CENTRO DE CIENCIAS BÁSICAS

TESIS

MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO

PRESENTA

Arturo Elías Ramírez

PARA OBTENER EL GRADO DE DOCTOR
EN CIENCIAS EXACTAS, SISTEMAS Y DE LA INFORMACIÓN

TUTOR

Dr. Alejandro Padilla Díaz

COMITÉ TUTORIAL

Dr. Carlos Alberto Ochoa Ortíz-Zezzatti
Dr. Julio César Ponce Gallegos
Dra. Aurora Torres Soto
Dr. Sergio Enríquez Aranda

Aguascalientes, Ags. 26 de Mayo del 2013

Autorizaciones

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

ANIVERSARIO UAA

Centro de Ciencias Básicas

M. en C. ARTURO ELÍAS RAMÍREZ
ALUMNO (A) DEL DOCTORADO EN CIENCIAS EXACTAS, SISTEMAS Y DE LA INFORMACIÓN,
P R E S E N T E .

Estimado (a) alumno (a) Elías:

Por medio de este conducto me permito comunicar a Usted que habiendo recibido los votos aprobatorios de los revisores de su trabajo de tesis y/o caso práctico titulado: **"MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO,** hago de su conoçimiento que puede imprimir dicho documento y continuar con los trámites para la presentación de su examen de grado.

Sin otro particular me permito saludarle muy afectuosamente.

A T E N T A M E N T E
Aguascalientes, Ags., 28 de mayo de 2013
"SE LUMEN PROFERRE"
EL DECANO SUSTITUTO

M. en C. JOSÉ DE JESÚS RUIZ GALLEGOS

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES CENTRO DE CIENCIAS BÁSICAS

c.c.p.- Archivo
JJRG.mjda

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

M. EN C. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como Tutor designado del estudiante **ARTURO ELÍAS RAMÍREZ**, registrado con el ID 9166 quien realizó el trabajo de tésis titulado **MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el VOTO APROBATORIO, para que el pueda proceder a imprimirlo, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 27 de Mayo de 2013.

Dr. Alejandro Padilla Díaz.
Tutor de Tesis.

c.c.p. Interesado
c.c.p. Secretario de Investigación y Posgrado.
c.c.p. Jefatura del Depto. de Ciencias de la Computación.
c.c.p. Consejero Académico
c.c.p. Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

M. EN C. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como Tutor designado del estudiante **ARTURO ELÍAS RAMÍREZ**, registrado con el ID 9166 quien realizó el trabajo de tésis titulado **MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el VOTO APROBATORIO, para que el pueda proceder a imprimirlo, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 24 de Mayo de 2013.

Dr. Carlos Alberto Ochoa Ortíz Zezzatti.
Tutor de Tesis.

c.c.p. Interesado
c.c.p. Secretario de Investigación y Posgrado.
c.c.p. Jefatura del Depto. de Ciencias de la Computación.
c.c.p. Consejero Académico
c.c.p. Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

M. EN C. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como Tutor designado del estudiante **ARTURO ELÍAS RAMÍREZ**, registrado con el ID 9166 quien realizó el trabajo de tésis titulado **MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el VOTO APROBATORIO, para que el pueda proceder a imprimirlo, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo ló anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 27 de Mayo de 2013.

Dr. Julio César Ponce Gallegos.
Asesor de Tesis.

c.c.p. Interesado
c.c.p. Secretario de Investigación y Posgrado.
c.c.p. Jefatura del Depto. de Ciencias de la Computación.
c.c.p. Consejero Académico
c.c.p. Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

M. EN C. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como Tutora designada del estudiante **ARTURO ELÍAS RAMÍREZ**, registrado con el ID 9166 quien realizó el trabajo de tésis titulado **MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el VOTO APROBATORIO, para que el pueda proceder a imprimirlo, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 27 de Mayo de 2013.

Dra. Aurora Torres Soto.
Asesora de Tesis.

c.c.p. Interesado
c.c.p. Secretario de Investigación y Posgrado.
c.c.p. Jefatura del Depto. de Ciencias de la Computación.
c.c.p. Consejero Académico
c.c.p. Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

M. EN C. JOSÉ DE JESÚS RUIZ GALLEGOS
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

PRESENTE

Por medio del presente como Tutor designado del estudiante **ARTURO ELÍAS RAMÍREZ**, registrado con el ID 9166 quien realizó el trabajo de tésis titulado **MODELOS DE PUNTUACIÓN CREDITICIA PARA INSTITUCIONES MICROFINANCIERAS MEXICANAS A TRAVÉS DE PROPUESTAS ENSAMBLADAS E HIBRIDADAS CON EL USO DE PSO**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el VOTO APROBATORIO, para que el pueda proceder a imprimirlo, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

ATENTAMENTE
"Se Lumen Proferre"
Aguascalientes, Ags., a 27 de Mayo de 2013.

Dr. Sergio Enríquez Aranda.
Asesor de Tesis.

c.c.p. Interesado
c.c.p. Secretario de Investigación y Posgrado.
c.c.p. Jefatura del Depto. de Ciencias de la Computación.
c.c.p. Consejero Académico
c.c.p. Minuta Secretario Técnico

Agradecimientos

El primer agradecimiento que quiero y debo externar es a Dios, por permitirme transitar por la vida en medio de bendiciones y oportunidades como lo es el poder finalizar el estudio de este doctorado, además de rodearme de gente que me ha brindado su apoyo y su afecto. Entre estas personas son destacables por su presencia continua mi gran y creciente familia, a quienes también tengo un sinfín de situaciones que agradecer. Especialmente agradezco a Clelia, quien ha tenido que soportar las ausencias, los desvelos y las desatenciones, quien no solo ha estado a mi lado, sino que se ha preocupado por la atención de Arturito, Ale y Montse, junto a quienes han permanecido estoicos brindándome su cariño incondicional tan valioso en los momentos de tribulación. Igualmente agradezco a mis papás Arturo y Silvia, fuente de sabiduría y amor paternal incondicional quienes junto a mis hermanos Silvia, Carlos Laura, Sofía y sus respectivos cónyuges e hijos siempre han tenido las palabras oportunas para no permitirme claudicar. Omito nombrar a todos los demás familiares con la idea de no excluir a nadie, mi agradecimiento va para todos ellos.

En el plano académico existe una gran cantidad de personas que han tenido una fuerte influencia en mí; pudiese mencionar gente desde mi etapa de jardín de niños pero la lista sería de enormes dimensiones por los que mencionaré únicamente a quienes estuvieron en esta última fase de mi formación. Comienzo por los doctores Padilla, primordialmente al Dr. Alejandro Padilla quien con sus conocimientos, experiencia y eterno optimismo me ha conducido en el camino del trabajo doctoral. Mención significativa tengo para el Dr. Alberto Ochoa, quien de manera relevante ha brindado su apoyo a muchos alumnos de este programa en nuestra formación y producción como investigadores. Al Dr. Francisco Luna quien pese a los enfados y quejas de los alumnos muestra las pautas necesarias para un trabajo de investigación con el rigor necesario para resultados de calidad. A las Dras. Eunice Ponce y Elba Díaz por todos sus comentarios y conocimientos en general y en lo particular los matemáticos, herramientas indispensables para la fundamentación de propuestas dentro del área de la inteligencia artificial.

Finalmente quisiera agradecer a los amigos, quienes aún por el distanciamiento y disminución en la frecuencia de convivencia, al paso del tiempo y en cada reencuentro es como si no hubiese pasado más de un día y en el abrazo se estrechan los lazos de amistad.

En general quiero agradecer a todas aquellas personas que de una manera directa o indirecta, consciente o inconsciente han contribuido para mi desenvolvimiento en la vida.

Dedicatoria

A mi familia, desde el más pequeño hasta el más grande, con quien convivo día a día y a quien muy ocasionalmente lo veo.

Clelia compañera en los vaivenes de la vida que a su lado se vuelven más intensos y placenteros en las crestas, pero más tolerantes y llevaderos en los valles, a través de un cariño y amor mutuo.

Ivette Alejandra, Arturo, Andrea Montserrat (I1, A3, M1 [IAM 131]), quienes de forma natural e inimaginable se han vuelto grandes faros que como a barco en la tormenta guían y dan rumbo certero a mi vida.

Silvia y Arturo labradores persistentes y entregados que hoy ven los frutos de lo sembrado y cultivado con paciencia y cariño.

Silvia, Carlos Alejandro, Laura y Sofía, hermanos y compañeros en el caminar por la vida desde la infancia y de quienes afortunadamente puedo sentir la presencia y fortaleza de una cofradía que ha crecido a la par de sus respectivas familias a quienes también dedico este trabajo.

A quienes su presencia física ya no nos acompaña, pero que fueron en vida un eslabón más de la cadena familiar, especialmente a quienes al comenzar el estudio doctoral aún estaban entre nosotros.

Contenido General Página

Índice de Tablas Página

Índice de Gráficos Página

Índice de Ecuaciones Página

Acrónimos

AI	Artificial Intelligence (Inteligencia Artificial)
ACS	Ant Colony Systems (Sistemas de Colonias de Hormigas).
ANN	Artificial Neural Networks (Redes Neuronales Artificiales).
AUC	Area Under Curve ROC (Área bajo la curva ROC).
CNBV	Comisión Nacional Bancaria y de Valores.
CONCAMEX	Confederación de Cooperativas de Ahorro y Préstamo de México
CONDUSEF	Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros.
Conf	Confiabilidad.
CPCPSOK	Clusterización PCA PSO K-medias
CRISP-DM	CRoss Industry Standard Process for Data Mining (Metodología para el Desarrollo de Proyectos en Minería de Datos).
DE	Differential Evolution (Evolución Diferenecial)
EA	Evolutionary Algorhitms (Algoritmos Evolutivos).
EP	Evolutionary Programming (Programación Evolutiva).
EPCPSO	Ensamble PCA PSO
ES	Evolutionary Strategies (Estrategias Evolutivas).
FIPS	Fully Informed Particle Swarm (Acumulación de Partículas Totalmente Informadas).
GA	Genetic Alghoritm (Algoritmo Genético).
GP	Genetic Programming (Programación Genética).
ICC	Índice de clasificación correcta.
IEC	Índice erróneo de clasificación.
IMF	Institución Micro-Financiera.
KBS	Knowledge Bases Systems. (Sistemas Basados en Conocimiento)
KMO	Coeficiente de Kaiser-Meyer-Olkin
LR	Logistic Regression (Regresión Logística).
MLR	Multiple Linear Regression (Regresión Lineal Múltiple).

NP	Nondeterministic polynomial time ("tiempo polinomial no determinista")
PCA	Principal Component Analysis (Análisis de Componentes Principales).
PCC	Probabilidad de Clasificación Correcta.
PSO	Particle Swarm Optimization (Optimización por Cúmulo de Partículas).
ROC	Receiver Operator Characteristics (Característica Operativa del Receptor).
SHCP	Secretaría de Hacienda y Crédito Público.
Se	Sensibilidad.
SPEF	Especificidad.
SOCAPS	Sociedades Cooperativas de Ahorro y Préstamos
SOFIPOS	Sociedades Financieras Populares.
SOFOM	Sociedad Financiera de Objeto Múltiple.
SVM	Support Vector Machines (Máquinas de Soporte Vectorial).
VPN	Valor predictivo negativo.
VPP	Valor predictivo positivo.

Resumen

El acceso a las oportunidades productivas de los países en desarrollo, implica en muchas ocasiones que las pequeñas empresas y las microempresas tengan que recurrir a los servicios financieros de crédito, favoreciendo desde mediados de los años 60 la aparición de las microfinanzas como herramienta para ofrecer estas oportunidades, a partir de un concepto de inclusión financiera que contempla la problemática de acceso al financiamiento con una perspectiva global. Sin embargo, las dificultades derivadas de la informalidad y la pobreza en ciertos sectores de la población afectan directamente a los métodos de gestión de las microempresas y sus riesgos asociados, propiciando el aumento de los riesgos de crédito de las instituciones de microfinanzas con respecto a otras instituciones financieras. Así mismo, la industria microfinanciera se ha visto afectada por la fuerte competencia proveniente de la banca tradicional que ha comenzado a dirigirse a los clientes tradicionales de las IMFs. Resulta fundamental por lo tanto, llevar a cabo un correcto análisis de la información para el desarrollo y expansión de la IMFs, adquiriendo metodologías para seleccionar un público adecuado y un sector económico deseado.

La estrategia más utilizada en la actualidad dentro de la banca comercial es la puntuación crediticia (credit scoring), que es un sistema de evaluación automático, más rápido, más seguro y consistente para determinar la concesión de créditos a partir de toda la información disponible, capaz de predecir la probabilidad de impago, asociada a una operación crediticia. Por lo tanto, la aplicación de la estadística, investigación de operaciones y técnicas inteligentes bajo la idea de minería de datos en la evaluación del riesgo de crédito y la predicción de quiebra ha sido un área de interés para los investigadores desde los años 70.

La idea de la propuesta consiste en presentar dos alternativas de sistemas de puntuación crediticia para IMFs mexicanas con una fácil interpretación para el oficial de crédito. El primer modelo utiliza un enfoque de hibridación y de ensamblado, a partir de PSO para asignar pesos en una forma hibridada a las técnicas estadísticas de MLR y RL en el enfoque ensamblado. El segundo modelo emplea el agrupamiento de clientes para implementar la puntuación crediticia mediante un algoritmo hibrido resultante de la combinación de PCA,

PSO y K-medias para realizar la fase de agrupamiento y con los grupos resultantes aplicar el primer modelo de puntuación crediticia propuesto para de manera ponderada obtener las valoraciones de cada cliente.

Los dos modelos parten de la reducción de la dimensionalidad de los datos empleando un modelo de selección de variables hibrido con la aplicación de PCA y PSO. De estos tres modelos propuestos se generan las aportaciones de este trabajo de tesis, el cual se sustenta en el uso hibridado de diferentes variantes del algoritmo PSO con diversas técnicas estadísticas y de investigación de operaciones.

Summary

Access to productive opportunities of developing countries, often implies that small and micro enterprises have to resort to credit financial services, promoting since the mid 60's the emergence of microfinance as a tool to provide these opportunities, from a financial inclusion concept, contemplates the problem of access to finance with a global perspective. However, the difficulties arising from the informality and poverty in certain sectors of the population directly affect the methods of micro-management and its associated risks, resulting in greater credit risk of microfinance institutions with respect to other financial institutions. Furthermore, the microfinance industry is now affected by strong competition: commercial banks have begun to target MFIs traditional customers, new MFIs have continued to be created in microfinance industry, and the microfinance clientele is becoming more sophisticated concerning the quality of service they require or expect. These factors may negatively affect the MFIs. It is essential thus to carry out a proper analysis of information for the development and expansion of MFIs, acquiring methods to select an appropriate audience and desired economic sector.

The strategy used at present in commercial banks is credit scoring, an automatic evaluation system, faster, safer and consistent lending determine from all available information, able to predict the probability of default, associated with a credit transaction. Therefore, the application of statistics, operations research and intelligent techniques under the idea of data mining in the credit risk assessment and bankruptcy prediction has been an area of interest for researchers since the 70s.

The idea of the proposal consists in present two credit scoring system alternatives for Mexican MFIs with an easy interpretation for the credit officer. The first model uses a hybridized and ensemble approach, based on PSO to assign weights in a hybridized form to the statistical techniques of MLR and logistic regression LR in the ensemble approach. The second model uses grouping of customers to implement credit score using an algorithm resulting of hybridizing PCA, PSO and K-means, to perform clustering stage and applying the first model of the proposed credit score to the resulting groups to obtain a weighted valuations of each client.

The two models are based on reducing the dimensionality of the data using a variable selection model by hybridizing PCA and PSO. Of these three models proposed are generated contributions of this thesis, which is based on the use of different variants hybridized PSO algorithm with various statistical techniques and operations research.

I

Capítulo I Introducción

I.1 Definición del problema

El acceso a oportunidades productivas en los países en vías de desarrollo significa acceso a la agricultura, comercio a pequeña escala, tecnología, sanidad, educación, etc. Sin embargo, el desempeño de toda actividad tendiente al aprovechamiento de estas oportunidades obliga a las personas a acudir a los servicios financieros de crédito para la pequeña y micro producción, motivando que desde mediados de los años 60, se haya considerado a las microfinanzas como un instrumento para brindar dichas oportunidades, partiendo de un concepto de inclusión financiera que contempla la problemática de acceso al financiamiento con una perspectiva global; desde los efectos del desarrollo económico de un país hasta las consecuencias devastadoras del estrecho acceso a productos financieros en términos de oportunidades y nivel de vida de los más necesitados (ProDesarrollo Finanzas y Microempresa, 2011). Estos servicios crediticios originan en diversas zonas de estos países, el surgimiento de sistemas basados en microactividades productivas e informales que garantizan la supervivencia de gran parte de la población, configurándose así un factor de prosperidad económica para el conjunto de la sociedad.

Con todo, las dificultades derivadas de la informalidad y de la pobreza en determinados sectores de la población afectan directamente a los métodos de gestión de negocio de las microempresas y sus riesgos asociados (Lara Rubio, 2010), incrementando los riesgos crediticios de las Instituciones Microfinancieras (IMFs) a los que cualquier organismo financiero se encuentra expuesto. Resulta fundamental por lo tanto, llevar a cabo un correcto análisis de la información para el desarrollo y expansión de la IMFs, adquiriendo metodologías para seleccionar un público adecuado y un sector económico deseado. La estrategia más utilizada en la actualidad dentro de la banca comercial es la puntuación crediticia (credit scoring), que es un sistema de evaluación automático, más rápido, más seguro y consistente para determinar la concesión de créditos a partir de toda la información disponible, capaz de predecir la probabilidad de impago, asociada a una

operación crediticia. La forma de funcionamiento de puntuación de crédito es muy sencilla en teoría. El procedimiento básico de trabajo se explica de la siguiente manera: la variable dependiente (Y) representa el riesgo de crédito (la probabilidad de pago). Las variables independientes (variables predictivas o Xi) se utilizan para explicar la variable dependiente (Jentzsch, 2007).

Adicionalmente, los historiales de crédito y las puntuaciones crediticias son esenciales en toda América Latina para ayudar a resolver tres problemas económicos específicos: a) Niveles de eficiencia inferiores a estándares internacionales en el sector financiero, b) el relativo estancamiento de los préstamos del sector privado y c) el riesgo de crisis financieras, que a menudo derivan, en parte, de los problemas de selección adversa en el sector bancario (Turner & Varghese, 2006).

La aplicación de técnicas estadísticas, de investigación de operaciones e inteligentes, bajo la idea de minería de datos, en la evaluación de riesgo crediticia y predicción de quiebra ha sido un área de interés para los investigadores desde la década de los 70 (Ahmad Ghodselahi & Amirmadhi, 2011). Usualmente, las propuestas generales de evaluación de riesgo crediticio consisten en aplicar algunas técnicas de clasificación sobre información de clientes previos con características similares, tanto de buenos como de malos pagadores, con la idea de encontrar alguna relación entre dichas características y los comportamientos de pago (Yu, Wang, & Lai, 2008).

La puntuación crediticia puede tomar diferentes formas. Se pueden distinguir tres tipos principales de propuestas (Thomas, 2000): (i) de juicio, (ii) estadística y (iii) no paramétricas. Las propuestas de juicio son aún las más empleadas en las microfinancieras y evalúan los riesgos de crédito basándose en la experiencia y opinión con la que cuenta el evaluador (Van Gool, Baesens, Sercu, & Verbeke, 2009). En contraste, las propuestas estadísticas se basan en los datos históricos e incluyen análisis discriminante, regresión lineal y regresión logística (C.-F. Tsai & Chen, 2010). Finalmente, las metodologías no paramétricas incluyen una variedad de métodos de investigación de operaciones y de inteligencia artificial (Yu, Yue, Wang, & Lai, 2010). Estos tres tipos de propuestas utilizan técnicas y metodologías de minería de datos, Keramati y Yousefi muestran un compendio

de una gran variedad de trabajos desarrollados al respecto (Keramati & Yousefi, 2011). Sin embargo, la mayoría de estos estudios se han realizado para la banca tradicional de los países industrializados y solo algunos pocos trabajos relacionados con IMFs en Latinoamérica y países de África del Sur (Rayo Canton, Lara Rubio, & Camino Blasco, 2010), siendo el primer trabajo documentado el realizado por Viganò en Burkina Faso (Viganò, 1993).

La explicación y predicción del riesgo de impago en microfinanzas deben abordarse de una manera distinta a la habitual en la banca comercial, más allá de de las limitantes de las bases de datos con las que se cuenta, que son mas deficientes que las manejadas por la banca comercial en los países desarrollados, por las diferencias existentes entre los tipos de clientes a los que va dirigido el préstamo, teniendo en cuenta que los prestatarios de las IMFs generalmente son desempleados o trabajan en ocupaciones informales o autoempleo, originando créditos pequeños, normalmente a corto plazo y, habitualmente, sin garantías que lo respalden, lo que restringe información fundamental empleada por los modelos de la banca comercial. Schreiner (Schreiner, 1999) manifiesta lo anterior de la siguiente forma ”la manipulación matemática es la parte fácil. La parte difícil es la recolección de información y el uso de las estimaciones de riesgo en la práctica”. Es decir, la principal complicación a la hora de elaborar un modelo de puntuación crediticia para microfinanzas radica en combinar una serie de variables de carácter subjetivo sobre las cuales existe un gran problema a la hora de encontrar datos e información.

Adicionalmente, las limitaciones que aplican para la puntuación crediticia de la banca comercial, que sugieren que un modelo que no mantenga una proporción similar de créditos que aporten elevados rendimientos, a créditos cuya contribución al margen de intermediación financiera no sea tan elevada, no se encuentra correctamente elaborado a causa de no recoger una amplia información de las características que puedan suponer un aumento del riesgo de impago (Mester, 1997). Sin embargo, esta limitación identificada para la banca comercial, no lo es tanto para las IMF’s dado que según la función económica y social que define al microcrédito y a las entidades que los conceden, éstas vienen otorgando créditos de diversa índole, en el sentido que el exceso de riesgo de uno de ellos

puede verse compensado por la elevada rentabilidad que puede aportar otro y por tanto, se pueden encontrar en ella factores de riesgo más distinguidos (Lara Rubio, 2010).

Los motivos anteriormente mencionados y otros no enunciados, son los causales de que en la actualidad exista una ausencia importante de mecanismos cuantitativos capaces de medir el riesgo de la actividad crediticia de las IMFs, en comparación con el resto del sector financiero.

I.1.1 Problema de Investigación.

El problema de investigación consiste en el desarrollo de mecanismos de puntuación crediticia para IMFs mexicanas, con un modelado no paramétrico a partir de propuestas ensambladas (ensemble), hibridadas y semi-supervisadas tomando como punto de partida el empleo de Optimización por Cúmulo de Partículas (PSO, por sus siglas en inglés Particle Swarm Optimization). La parte correspondiente a la propuesta agregada emplea un esquema con asignación de pesos a partir del empleo de PSO de las técnicas estadísticas de Regresión Lineal Múltiple (MLR, por sus siglas en inglés Multiple Linear Regression) y Regresión Logística (LR, por sus siglas en inglés Logistic Regression), lo correspondiente a enfoque semi-supervisado se efectúa a través de la creación de agrupamientos (clusters) de prestatarios con características similares que permitan la clasificación de los prestatarios de acuerdo a dichos agrupamientos, mediante la hibridación Análisis de Componentes Principales (PCA, por sus siglas en inglés Principal Component Analysis), PSO y K-medias para la asignación de los agrupamientos, para finalmente realizar una comparativa entre las metodologías propuestas y determinar si existe alguna mejora con respecto a lo revisado en la literatura.

Este problema de investigación se realiza de forma empírica con datos recabados por diversas IMFs mexicanas, y se pretende este concluido para mediados del 2013.

I.2 Justificación

La escasez de modelos y aplicaciones de puntuación en las microfinanzas es indicativo de un campo de investigación joven y poco explorado. Además, un alto porcentaje de los trabajos realizados evalúan la utilidad de la puntuación crediticia y sólo unos pocos se han

centrado en la aplicabilidad del concepto en las microfinanzas. De estos pocos la gran mayoría le ha dado un enfoque estadístico y no se han detectado estudios publicados con la utilización de un modelo de inteligencia artificial.

Existen opiniones divergentes sobre la utilidad de la puntuación crediticia para las microfinanzas. Algunos autores señalan como ventajas la reducción de pérdidas por impago, el potencial de comercialización de los diferentes segmentos y la disminución del tiempo que el analista de crédito pasa con los clientes individuales. Según Schreiner (Schreiner, 2001), "Aunque la calificación es menos poderosa en los países pobres que en los países ricos, y a pesar de que la calificación en las microfinanzas no reemplazará el conocimiento personal del carácter por los analistas de crédito y por los grupos de crédito, la calificación puede mejorar las estimaciones de riesgo y así disminuir los costos. Por tanto, la calificación complementa, pero no sustituye, los esquemas empleados actualmente en las microfinanzas".

El diseño de una aplicación de puntuación en microfinanzas ha de estar basada en el seguimiento de una estrategia innovadora en este sector, pues estriba en combinar una serie de variables de carácter subjetivo sobre las cuales existe un gran problema al momento de encontrar los datos e información.

Por lo anterior, la propuesta de este trabajo pretende entregar aplicaciones de puntuación crediticia en microfinanzas, a partir de metodologías y técnicas que son no solamente innovaciones dentro de área financiera, sino adicionalmente dentro del área de la minería de datos y la inteligencia artificial como son los métodos ensamblados, híbridos y semi-supervisados. El aprendizaje ensamblado es un paradigma de máquina de aprendizaje donde múltiples agentes se entrenan para resolver el mismo problema. En contraste a las propuestas de máquinas de aprendizaje que tratan de aprender una hipótesis de los datos de entrenamiento, los métodos agregados tratan de construir un conjunto de hipótesis y emplearlas de manera combinada (G. Wang, Hao, Ma, & Jiang, 2011). En esta propuesta los métodos ensamblados son MLR y LR con una asignación de pesos a partir de PSO. En máquinas de aprendizaje, las propuestas de hibridación han sido un área activa de investigación para mejorar el desempeño de la clasificación o predicción con respecto a una

propuesta simple. En general, la hibridación se basa en la combinación de dos técnicas diferentes de aprendizaje de máquina (Ahmad Ghodselahi & Amirmadhi, 2011). En el caso de esta propuesta se presenta un modelo de clasificación hibrida consistente de un agente como PSO para el pre-procesamiento de los datos de entrenamiento y formar los agrupamientos en combinación con un agente como K-medias para optimizar el establecimiento de los centroides de los agrupamientos y desde dichos agrupamientos determinar sus características para establecer la evaluación crediticia.

I.2.1 Preguntas de Investigación

Acorde al problema de investigación presentado y con la idea de dar un seguimiento a la justificación del desarrollo de este trabajo se plantean las siguientes preguntas de investigación.

i. ¿Pueden lograr propuestas de inteligencia artificial en la puntuación crediticia un impacto similar en las instituciones microfinancieras al obtenido en instituciones de banca comercial y de las hipotecarias? Los prestatarios en las microfinancieras tienen un perfil muy diferente, pues generalmente se trata de personas con pocos recursos, para los cuales los requisitos impuestos por la banca comercial y muchas otras instituciones que proporcionan créditos (Lara Rubio, 2010), los marginan de la posibilidad de acceder a ellos y los colocan como clientes de alto riesgo de incumplimiento en el pago del crédito solicitado.
ii. ¿Resulta relevante el uso de PSO para asignación de los pesos del método de voto mayoritario en una propuesta ensamblada dentro del entorno de puntuación crediticia en IMFs? Las propuestas ensambladas han demostrado tener un mejor desempeño en la solución de problemas que los obtenidos por propuestas simples (Nanni & Lumini, 2009), en donde un factor de consideración es la asignación de pesos que cada propuesta individual tendrá dentro de la solución agrupada. Los pesos se pueden calcular mediante el uso de diversas técnicas o herramientas, por lo que esta propuesta hace uso de PSO para llevar a cabo el cálculo y lograr una mejor adecuación de la solución.
iii. ¿Ante la dificultad de la recolección de información una propuesta semi-supervisada presenta mejores resultados de confiabilidad de pronóstico y reducción de error tipo

II en la puntuación crediticia que una propuesta supervisada? Una de las mayores problemáticas que se presenta en la implementación de puntuación crediticia es la irregularidad y el desequilibrio de los datos disponibles para desarrollar las fases de entrenamiento y prueba de los métodos supervisados (Van Gool, et al., 2009), con la idea de afrontar esta situación se presenta un modelo semi-supervisado, a través de agrupamientos optimizados con PSO.

iv. ¿Puede alcanzar la propuesta una confiabilidad como la que logran las máquinas de soporte vectorial (SVM por sus siglas en inglés Support Vector Machines) y las redes neuronales (ANN por sus siglas en inglés Artificial Neural Networks)? En la literatura se ha demostrado que dos de las mejores alternativas para mejorar la confiabilidad de los sistemas de puntuación crediticia son las máquinas de soporte vectorial y las redes neuronales con variantes e hibridación (Fu & Liu, 2011), sin embargo, estas propuestas son muy poco empleadas por las instituciones crediticias debido a lo poco explicativas que son sus procesos en la obtención de resultados. Con la idea de que se comprenda el proceso del cálculo de puntuación crediticia esta propuesta hace uso de elementos que más se han empleado por su facilidad de explicación como lo son MLR y LR de manera ensamblada para superar sus deficiencias, así como de un sistema de agrupamiento para establecer reglas que faciliten y expliquen el pronóstico de la puntuación calculada.

I.3 Hipótesis

De acuerdo a las preguntas de investigación planteadas se pueden derivar en tres hipótesis de investigación es este trabajo, las cuales se presentan a continuación.

1. La primera hipótesis está directamente relacionada con el impacto de una propuesta del uso de técnicas de inteligencia artificial en minería de datos para apoyar la fijación de una puntuación crediticia dentro de las IMFs con el mismo impacto que en las instituciones bancarias internacionales teniéndose así las siguientes primeras hipótesis.
    - Hi1: El uso de una propuesta de técnica inteligente en minería de datos aplicados a los conjuntos de datos de las IMFs mexicanas empleando modelos ensamblados con PSO, MLR y LR produce pronósticos con una confiabilidad

superior a los modelos aplicados tradicionalmente en la clasificación de puntuación crediticia de sus prestatarios.

- H01: El uso de una propuesta de técnica inteligente ensamblada no produce mejores pronósticos de puntuación crediticia de las IMF mexicanas debido a la pobreza de los conjuntos de datos como lo maneja la literatura.

2. La segunda hipótesis versa en la mejora a la propuesta a partir de un esquema semisupervisado de agrupamientos de los prestatarios para posteriormente aplicar el modelo de puntuación a los clientes agrupados, generándose las siguientes hipótesis:
    - Hi2: La implementación de agrupamientos hibridada para la clasificación de los prestatarios mediante PCA, PSO, y K-medias y el modelo ensamblado de puntuación crediticia produce resultados con mayor grado de confiabilidad al momento de recalcular su puntuación crediticia.
    - H02: El uso de agrupamientos no supera la propuesta inicial en cuanto a la confiabilidad de la puntuación crediticia.
3. Finalmente la última hipótesis plantea el comportamiento de las propuestas realizadas frente a otras propuestas expuestas en la literatura, siendo las hipótesis resultantes las que se enuncian a continuación.
    - Hi3: Los modelos propuestos de puntuación crediticia son equiparables a los modelos más exitosos presentados en la literatura como SVM y ANN en cuanto a confiabilidad de pronóstico de puntuación crediticia y de disminución de errores tipo II (falsos positivos).
    - Ha3: Los modelos propuestos aunque no son mejores se aproximan de una manera adecuada en cuanto a confiabilidad de puntuación crediticia y disminución de errores tipo II, por los modelos más exitosos presentados en la literatura.
    - H03: Los modelos propuestos son superados ampliamente en el cálculo de la confiabilidad de pronóstico de puntuación crediticia y disminución en errores tipo II, por otros modelos propuestos en la literatura.

I.4 Objetivos

La gran trascendencia demostrada en el crecimiento reportado de las IMFs a nivel mundial y específicamente en México, implica una gestión de rigurosa calidad de las microfinancieras, en especial en lo que al tratamiento de crédito se refiere, motivos por los cuales en este trabajo se pretenden alcanzar los siguientes objetivos:

I.4.1 Objetivo General:

Desarrollar modelos inteligentes que permitan el cálculo de puntuación crediticia a las Instituciones Microfinancieras, que ofrezcan una evaluación y medición del riesgo de impago, cuyos resultados sean válidos para la propuesta de una aplicación de negocio para estas entidades, a partir de las metodologías de minería de datos de agregación, hibridación y agrupamiento con un sustento de PSO.

I.4.2 Objetivos Específicos

- Proceder a una revisión teórica de las metodologías y modelos de puntuación crediticia en la banca comercial y las microfinanzas, con objeto de establecer diferencias entre ellos.
- Implementar una aplicación de puntuación crediticia a partir de una propuesta de agregación de los métodos estadísticos de Regresión Lineal Múltiple y Regresión Logística, con ponderación de pesos mediante Optimización por Acumulación de Partículas.
- Implementar una aplicación de puntuación crediticia semi-supervisada mediante la hibridación de las técnicas de PCA, Optimización por Acumulación de Partículas y K-medias.
- Comprobar la funcionalidad de las aplicaciones desarrolladas a partir pruebas empíricas realizadas con información real de Instituciones Microfinancieras.

I.5 Diseño del modelo propuesto

La investigación tiene un alcance explicativo, a partir del desarrollo de propuestas que implementan la puntuación crediticia mediante un enfoque ensamblado y otro hibrido semi-supervisado, con la idea de afrontar el principal desafío del microcrédito que es el gestionar el riesgo del incumplimiento de un cliente, ya sea por caer en moratoria o pagar

debidamente pero no volver a solicitar créditos adicionales, realizando las pruebas correspondientes de las propuestas de manera empírica con información proveniente de IMFs mexicanas, las cuales por motivos de confidencialidad no pueden ser nombradas dentro de este documento.

Los métodos agregados se emplean para mejorar el desempeño y confiabilidad de las tareas de clasificación. Los sistemas clasificadores múltiples se basan en la agregación de un grupo de clasificadores de tal manera que su fusión logra un mayor rendimiento que los clasificadores individuales (Ahmad Ghodselahi & Amirmadhi, 2011). La idea principal de la mayoría de los métodos para construir agregación de clasificadores es modificar el conjunto de datos de entrenamiento, construyendo clasificadores en estos n conjuntos nuevos de entrenamiento y entonces combinarlos en una nueva regla de decisión final (Nanni & Lumini, 2009). En el caso de la propuesta presentada los clasificadores a considerar son MLR y LR y la regla de decisión final se logra a partir de la aplicación de PSO, que es quien otorga los pesos a cada uno de los clasificadores para alcanzar el objetivo de la agregación. El modelo propuesto se presenta en la figura I.1.

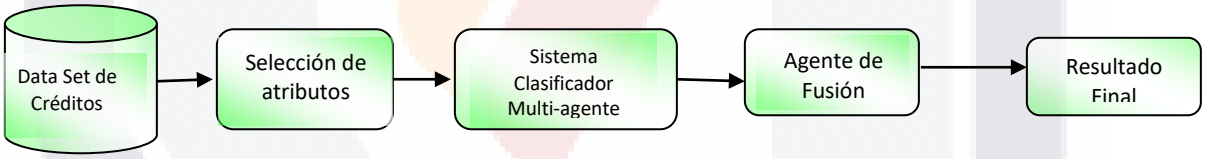


Figura I-1 Representación del modelo ensamblado propuesto para la asignación de puntuación crediticia.

Por su parte los métodos de agrupamiento son una técnica de minería de datos con un enfoque de tipología no supervisada, aunque es frecuentemente empleado en esquemas semisupervisados (Chandola, Banerjee, & Kumar, 2009). El método de K-medias es un algoritmo de agrupamiento geométrico bien conocido que se basa en el trabajo de Lloyd (Lloyd, 1982). El procedimiento K-medias pertenece al grupo de técnicas de agrupamiento conocidos como métodos de partición u optimización. Sin embargo, el algoritmo sufre de algunos inconvenientes de consideración. La función objetivo del K-medias no es convexa y por lo tanto puede contener varios mínimos (máximos) locales (Ahmadyfard & Modares, 2008). El resultado de K-medias, por lo tanto depende considerablemente de la elección inicial de los centroides de agrupamiento. La tesis implementa una alternativa mejorada de

K-medias mediante la hibridación de PCA, K-medias y PSO. PSO es un algoritmo poblacional basado en la técnica de optimización estocástica que se puede utilizar para encontrar un equilibrio óptimo, o cerca de una solución óptima para un problema numérico y cualitativo (Cui & Potok, 2005). El algoritmo PCA se emplea con k-medias para reducir el problema de dimensionalidad al que se enfrenta k-medias cuando la dimensión de los datos es grande y pierde su efectividad. Sin embargo k-medias con PCA no presenta mucha optimización, experimentalmente se puede observar que la optimización de k-medias ofrece resultados más precisos, así el uso de PSO optimiza k-medias con PCA para el agrupamiento de conjuntos de datos de alta dimensión. La figura I.2 muestra el modelo propuesto para esta fase.

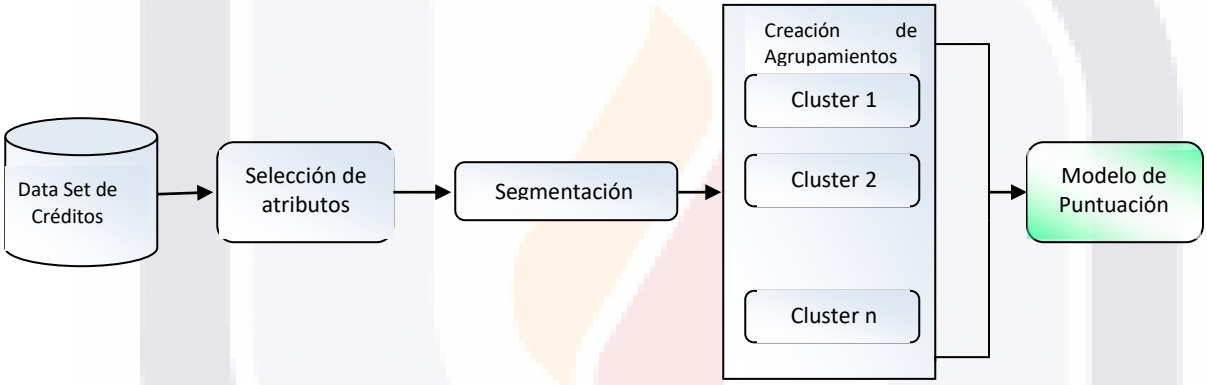


Figura I-2 Representación del modelo hibrido semisupervisado para la definición de agrupamientos de prestatarios y su clasificación.

Las propuestas de este trabajo se centran en 3 puntos principales, selección de variables a través de la disminución de la dimensión del problema original (secciones IV.3.3.1 y V.1). Los resultados de la fase de selección de variables se emplean tanto en la obtención del modelo de puntuación crediticia (secciones IV.3.3.2 y V.2), como en la de agrupamiento de prestatarios (secciones IV.3.3.3 y V.3). Sin embargo todas van dirigidas hacía el mejor establecimiento de un modelo de puntuación crediticia que ofrezca un pronóstico de puntuación confiable.

En la figura I.3 se muestra la idea general del modelo correspondiente a la aportación total de este trabajo de tesis, así partiendo del conjunto de datos de la IMF los cuales se depuran acorde a lo establecido por CRISP-DM (CRoss Industry Standard Process for Data Mining)

se lleva a cabo una selección de atributos empleando la propuesta hibridada de PCA-PSO, con la idea de reducir la alta dimensionalidad de los datos. Al definirse los atributos a utilizar se establece un modelo de puntuación crediticia ensamblada, según se muestra en la figura I.1 empleando bagging con los métodos MLR y LR y voto mayoritario con la asignación de pesos mediante la metaheurística de PSO, este modelo de puntuación crediticia se aplica a una población muestra de entrenamiento y a una población de prueba, ambas poblaciones son muestras parciales del datset, así como a los nuevos prestatarios que solicitan un crédito. De manera paralela se realizan agrupamientos de los clientes para poder definir puntuaciones crediticias acorde a la probabilidad de pertenencia de los prestatarios a los diferentes agrupamientos obtenidos mediante el modelo mostrado en la figura I.2, así la probabilidad de pertenencia ponderada se emplea en conjunto con el modelo de puntuación crediticia ensamblado. Los resultados de los modelos anteriores se comparan entre sí para determinar cuál es el que obtiene mejores resultados y posteriormente evaluarlos con otras propuestas de puntuación crediticia adaptadas a partir de los desarrollos para la banca comercial.

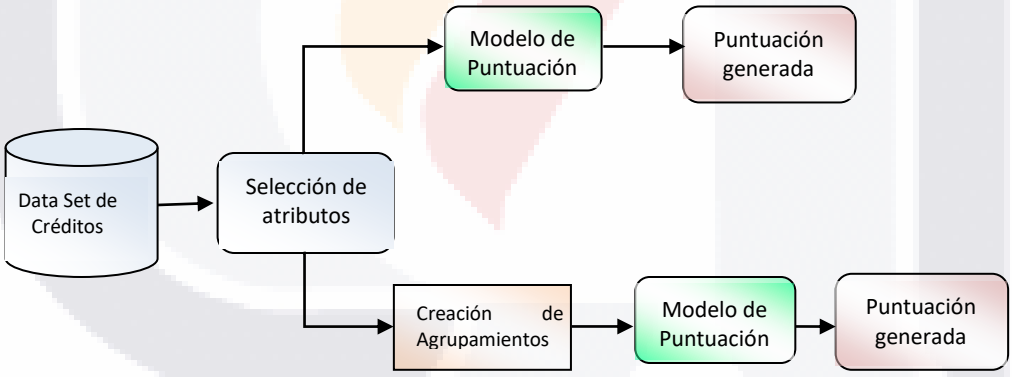


Figura I-3 Modelo general propuesto para el cálculo de puntuación crediticia.

Para lograr los objetivos trazados, el documento de Tesis Doctoral está dividido en tres bloques que agrupan las diferentes fases de desarrollo y cumplimiento de dichos objetivos como se expone a continuación.

I.6 Aportaciones del trabajo doctoral.

Este trabajo doctoral presenta dos contribuciones principales dentro del área de minería de datos, las cuales en esta propuesta se han desarrollado de manera independiente a través del modelado de diversos métodos, soportados por el uso de variantes del algoritmo PSO, pero que al final se integran para coadyuvar al proceso de puntuación crediticia dentro de las instituciones microfinancieras, como un ejemplo más de la gran variedad de actividades cotidianas que pueden verse favorecidas con el uso de minería de datos, como lo demuestran los capítulos Explaining Diverse Application Domains Analyzed from Data Mining Perspective (Ochoa, et al., 2012), y New Implementations of Data Mining in a Plethora of Human Activities (Ochoa, Ponce, Elias, Ornelas, et al., 2011), y el artículo Analysis of Cyber-bullying in a virtual social networking (Ochoa, Ponce, Elias, Jaramillo, et al., 2011).

La aportación inicial de la tesis se centra en la selección de variables en problemas de alta dimensionalidad, debido a que es la razón principal de la mayoría de los problemas en minería de datos se deriva de que un gran número de muestras a menudo tienen diferentes tipos de caracteres, y estas muestras suelen ser datos de alta dimensión, lo que significa que tienen varios atributos medibles. Estas dimensiones redundantes de datos a gran escala producen la maldición de la dimensionalidad en la minería de datos, que se genera por la alta dimensión de la geometría del espacio, y este tipo de espacios de datos son problemas representativos de la minería de datos. En este sentido se busca la reducción de la dimensionalidad mediante el uso de PCA que es un método recurrente en la solución de este tipo de problemas, pero que se ve mejorado con la hibridación con PSO binario para obtener la muestra óptima de variables representativa. Aunque PCA y PSO se han empleado en el uso de selección de características con anterioridad, la mayoría de las aplicaciones realizadas se abocan a problemas de reconocimiento de imágenes, en este sentido la aportación se refiere a la aplicación de minería de datos, específicamente como ya se comentó las instituciones microfinancieras.

Un segundo aspecto y el principal de contribución de este trabajo es el cálculo de puntuación crediticia en IMFs, el cual se lleva a cabo mediante dos metodologías. Al igual

que el problema de selección de variables, la literatura presenta muchas propuestas para la estimación de puntuación crediticia, sin embargo los trabajos reportados para microfinanzas son muy pocos y la mayoría de ellos emplean métodos meramente estadísticos, además en la práctica cotidiana solamente se basan en la decisión del oficial de crédito encargado de dar seguimiento al otorgamiento del crédito, perdiéndose así las ventajas de reducción de costos, aceleración en la toma de decisiones y registros de datos actualizados y con mayor confiabilidad que ofrecen los métodos de investigación de operaciones y de inteligencia artificial, estos últimos los considerados en este trabajo.

El primer método de puntuación crediticia se implementa a través de un modelo ensamblado de de MLR y LR hibridado con PSO. El ensamblado es una técnica que recientemente se ha empleado de manera recurrente y busca mejorar el rendimiento de los modelos participantes mediante la complementación de sus resultados, en este caso la combinación de los clasificadores MLR y LR, la forma de conjuntarlos es generalmente con la combinación directa de los resultados, sin embargo esta propuesta mejora este rendimiento mediante la hibridación con PSO para calcular los pesos de aporte de cada clasificador para a través de la técnica de voto mayoritario establecer un pronóstico más robustos, esta propuesta se presenta en el artículos Credit Scoring for Microfinance Institutions in México an Ensemble and Hybridized Approach (Elías, Padilla, & Padilla, 2012).

El segundo método de puntuación crediticia se desarrolla mediante un modelo de análisis de agrupamientos hibridado, es en este punto que se genera el artículo Outlier analysis for plastic card fraud detection a hybridized and multi-objective approach (Elías, Ochoa-Zezzatti, Padilla, & Ponce, 2011) pues se requiere del manejo de agrupamientos para el establecimiento de anomalías. La propuesta de tesis emplea la hibridación de PCA, PSO y K-medias para la creación de agrupamientos, de este trabajo se desprende la ponencia Credit Scoring Mechanisms for MFIs (Elias, Padilla, & Ochoa, 2012) dentro del Consorcio Doctoral del MICAI 2012, el cual será publicado en el Journal RCS del IPN, el cartel Algoritmo Inteligente para agrupamiento de Clientes en Microfinancieras (Elias & Padilla, 2012), dentro del III Congreso Internacional la investigación en el Posgrado. Este modelo

de agrupamiento es un aporte que no ha sido ampliamente explotado en la literatura, pues se ha empleado la hibridación pero solo entre 2 de los métodos y rara vez la combinación de los tres. Una vez definidos los agrupamientos se aplica el mecanismo de puntuación ensamblada para todos los prestatarios en las agrupaciones definiendo la distancia individual entre cada prestatario y cada agrupación y de esta forma establecer un valor ponderado en la asignación de pesos de manera ensamblada. Este modelo no se ha encontrado en ninguna opción de la literatura que se ha consultado.

I.7 Descripción de la tesis

En el primer bloque, se muestra un marco teórico que define la perspectiva desde la cual se visualiza este trabajo, iniciando con una descripción de la puntuación crediticia en general y particularmente para las microfinancieras desde el entorno mexicano, posteriormente se hace una revisión literaria de minería de datos y los desarrollos de puntuación crediticia en las diversas áreas financieras y específicamente en las microfinancieras. Este primer bloque comprende los capítulos II y III del trabajo de investigación, los cuales de forma correspondiente muestran cada uno de los aspectos de la literatura enunciados anteriormente.

En el segundo bloque de trabajo, integrado por los capítulos IV y V, se presentan los fundamentos teóricos de PSO y la presentación del la propuesta de Tesis de acuerdo a la metodología de desarrollo de proyectos de minería de datos CRISP-DM. En el capítulo IV se muestran el fundamento teórico de PSO y a partir de ello el modelado de las fases de preparación de los datos y modelado de la propuesta.

En el capítulo V, se exponen todos los aspectos de evaluación de resultados de los modelos de preparación de datos, cálculo de puntuación crediticia y agrupamiento de clientes acorde a lo presentado en el capítulo IV.

Finalmente, en el tercero de los bloques, se establecen una serie de análisis y discusiones en el capítulo VI con respecto a los resultados obtenidos en este trabajo y la comparativa frente a otros resultados derivados de trabajos relacionados. Por último dentro del capítulo

VII se definen las conclusiones y futuras líneas de investigación sobre el trabajo realizado, las cuales tratan de dar respuesta a los objetivos planteados en la presente Tesis Doctoral.

II

Capítulo II El fundamento de la puntuación crediticia

II.1 El Riesgo financiero

En el contexto de las instituciones financieras como bancos o compañías de seguros en general no existe la llamada "comida gratis" o en otras palabras, no hay beneficio sin riesgo, sin embargo el riesgo financiero suele corresponder a grandes pérdidas en una cartera de activos, las cuales deben afrontarse con la finalidad de obtener mejores dividendos debido a que el éxito de estas instituciones está directamente ligada a su capacidad de controlar y gestionar los riesgos relacionados.

De manera intuitiva el riesgo financiero suele relacionarse con carteras del mercado en las bolsas de valores tales como acciones y bonos, debido a los precios de mercado a la baja (un evento llamado riesgo de mercado), las pérdidas en una cartera de contratos de seguros debido a la ocurrencia de grandes siniestros (seguros o de riesgo de suscripción)(Eberlein, Frey, Kalkbrenner, & Overbeck, 2007). De igual forma las pérdidas en una cartera de obligaciones o préstamos, causada por el incumplimiento de algunos emisores o los prestatarios (riesgo de crédito) (Ya-qiong, 2007). Una categoría de riesgo adicional es el riesgo operacional, que incluye las pérdidas resultantes de procesos internos inadecuados o fallidos, fraudes o litigios.

Por lo tanto la gestión del riesgo puede ser vista como una competencia fundamental de una institución financiera: mediante el uso de su experiencia y su capital, una institución financiera puede tomar los riesgos y gestionarlos mediante diversas técnicas, tales como la diversificación, cobertura, o el recálculo y la transferencia de riesgos de vuelta a los mercados, etc.

Mientras que la gestión de riesgos, siempre ha sido una parte integral del negocio financiero y de seguros, la predicción del riesgo financiero se ha convertido en las últimas

décadas en una de las principales áreas de crecimiento en el modelado estadístico y probabilístico, además desde mediados de los 90 diversos modelos de inteligencia artificial como ANN, SVM y algoritmos genéticos (AGs) se han empleado con dicho propósito, en el capítulo III se realiza un análisis más exhaustivo de estas técnicas.

Las mayores acciones de predicción de riesgo se han emprendido hacia el manejo de carteras, opciones de precios y cotizaciones, además de otros instrumentos financieros o de fijación de precios en los bonos. Sin embargo, a pesar de la amplia variedad de servicios bancarios, los préstamos a las empresas y el público sigue constituyendo el núcleo de los ingresos de los bancos comerciales y otras instituciones de crédito (Kocenda & Vojtek, 2009), por ello y aunque no tan conocidas pero no menos importantes, las puntuaciones crediticias (credit scoring) y de comportamiento (behavioral scoring), son aplicaciones de riesgo financiero desarrolladas para pronosticar créditos de consumo (Thomas, 2000).

II.2 ¿Qué es la puntuación crediticia?

Desde una perspectiva técnica, el proceso de otorgamiento de crédito es una serie relativamente sencilla de las acciones sobre dos partes principales. Estas acciones van desde la aplicación inicial del préstamo a la devolución con éxito el préstamo o el incumplimiento en el pago.

La puntuación crediticia es un método para evaluar el riesgo al otorgar los créditos y con ello apoyar a las organizaciones a decidir si otorgan o no dichos créditos a las solicitudes de préstamos realizadas. El resultado del empleo de la puntuación crediticia produce un puntaje (score), en donde los prestatarios se clasifican como buenos pagadores o malos pagadores (Yu, Wang, & Lai, 2009). En general, los resultados de la puntuación están en función de numerosos factores relacionados con la vida financiera de un consumidor, incluyendo el historial de pagos de un individuo, la relación deuda-capital, la duración del historial de crédito, los tipos existentes de los créditos concedidos y otras numerosas variables relacionadas con a las transacciones recientes. A pesar de que las ponderaciones asignadas a cada una de estas clases generales de variables están disponibles para algunas puntuaciones de crédito, las variables específicas y el peso específico asignado a dichas

variables son propiedad de las instituciones desarrolladoras. Por lo general, las puntuaciones de crédito al consumo se escalan para facilitar su uso a la gama de entre 350 y 850.

Los solicitantes con buena puntuación tienen gran posibilidad de pagar sus obligaciones financieras, mientras que, los solicitantes con un puntaje malo tienen alta probabilidad de no realizar los pagos correspondientes. Las instituciones suelen emplear la puntuación crediticia para clasificar a sus solicitantes de crédito o prestatarios en términos del riesgo que conllevan, con la finalidad de aislar los efectos característicos por morosidad o impagos de varios solicitantes, lo anterior tomando como referencia los acuerdos del tratado de Basilea II (Hasan & Zazzara, 2006).

El puntaje crediticio es en consecuencia un problema de clasificación, donde las características de entrada son las respuestas a las preguntas de un formulario de solicitud de crédito en conjunto con los resultados de una comprobación con una agencia de referencia de crédito y la salida es la división en "buenos" y "malos" prestatarios. (Thomas, 2000; Yu, Wang, & Lai, 2008).

Así la necesidad de modelos confiables que permiten predecir con precisión la posibilidad de impago, así como detectar los potenciales buenos pagadores es imprescindible para que los interesados puedan tomar medidas tanto preventivas o correctivas.

II.2.1 El modelado de la puntuación crediticia.

Los modelos de puntuación crediticia se desarrollan por las instituciones financieras o por los investigadores con la intención de solucionar los problemas involucrados durante el proceso de evaluación de prestatarios. Para construir un modelo de puntuación los desarrolladores deben emplear datos históricos, técnicas estadísticas, matemáticas y/o de inteligencia artificial. Comúnmente, el enfoque genérico de evaluación de riesgo de crédito es la aplicación de algunas técnicas de clasificación de datos similares de los clientes anteriores, tanto a los clientes buenos como a los irregulares, con el fin de encontrar una relación entre las características y las fallas potenciales(Yu, Wang, & Lai, 2008), esta relación dada un grupo de clientes, puede ser descrita matemáticamente como sigue:

$$S = \{(x_1, y_1) \cdots (x_j, y_j) \cdots (x_n, y_n)\}$$ **Ecuación II-1**

con

$$f(x_1, x_2, \ldots, x_m) = y_n$$ **Ecuación II-2**

donde cada cliente x_j posee m atributos: x_{j1}, x_{j2},…, x_{jm}, y_j denota el tipo de prestatario, por ejemplo bueno o malo. *f* es la función o el modelo de puntuación crediticia que se asigna entre las características de los clientes (entradas) y su solvencia (salida), siendo la tarea del modelo de puntuación crediticia (función *f*) es predecir el valor de y_j.

Un modelo bien diseñado, por lo general asigna un porcentaje elevado de puntuaciones altas a los prestatarios cuya confianza sea alta por parte de la institución, mientras que otorgará puntajes bajos en los casos donde se estime un problema en el comportamiento del acreedor. Pero ningún modelo es perfecto, y algunas cuentas malas recibirán una puntuación más alta de lo que algunas cuentas buenas. La exactitud de la puntuación de crédito es fundamental para la rentabilidad de la institución financiera. Una mejora aún menor al 1% en la exactitud de la calificación crediticia de los solicitantes, puede repercutir en la reducción de una gran pérdida para las instituciones financieras, provocadas por el riesgo crediticio (Ahmad Ghodselahi & Amirmadhi, 2011). Adicionalmente los beneficios obtenidos por el desarrollo de un sistema de puntuación crediticia confiable son (C.-F. Tsai & Wu, 2008):

- La reducción del costo por el análisis de crédito.
- Habilitar una decisión más rápida.
- Asegurar las colecciones crediticias y disminuir los posibles riesgos.

En el modelado de puntuación crediticia, generalmente se tienen en cuenta dos aspectos en donde el concepto principal de predicción suele ayudar. En primer lugar es necesario identificar las técnicas de predicción de riesgo de consumo que incorporan las condiciones económicas y automáticamente ajustan los cambios económicos. En segundo lugar, antes de intentar minimizar el porcentaje de consumidores que defraudan, las empresas están esperando que se pueda identificar a los clientes que son más rentables. Parte del

catalizador de este desarrollo es el aumento masivo de la información sobre las transacciones de consumo que ha sucedido en las última décadas (Thomas, 2000).

La información sobre los prestatarios se obtiene a partir de su solicitud de préstamo y de las agencias de crédito. Sin embargo, hay también una gran cantidad de la información de solicitantes anteriores, sus datos en la solicitud de préstamo y su desenvolvimiento posterior. En muchas organizaciones se mantiene la información de millones de clientes anteriores. No obstante esta situación que puede ser ventajosa acarrea una problemática. La empresa tendrá los detalles de la solicitud de aquellos clientes a los que se rechazó el crédito, pero no el conocimiento de cómo ha sido su comportamiento crediticio posterior, lo cual produce un sesgo al ejemplo. Esto ocasiona un serio problema debido a que si la institución ha rechazado previamente un cliente como malo, esta decisión pudiera perpetuarse en cualquier sistema de puntuación que utilice dicha información y grupos de clientes potenciales pueden no tener nunca la oportunidad de demostrar su valor. Por otra parte generalmente existen razones de peso para rechazar la petición de este tipo de solicitantes, motivo por el cual los rechazados tienen un mayor índice de riesgo que aquellos que previamente han sido aceptados. Si se puede imputar que los clientes que han sido rechazados serán buenos o malos, ha sido cuestión de diversos debates.

II.3 Desarrollo de los modelos de puntuación crediticia.

II.3.1 Variantes en la puntuación crediticia.

El crédito al consumo se ha desarrollado por alrededor de 4,000 años. Existe un grabado en arcilla de tablas sumerias sobre dos campesinos que pidieron préstamos para compra de cereales con la promesa de pagar más en la temporada de cosecha. En la edad media la discusión sobre si era correcto cargar intereses a los préstamos no solo fue el punto central de una de las obras de Shakespeare, sino que lo plantearon tanto los teólogos católicos como musulmanes(Thomas, 2009). Sin embargo, es sólo hasta los últimos cincuenta años, con la llegada de las tarjetas de crédito (publicado por primera vez en los EE.UU. en 1958 y luego en el Reino Unido en 1966) y el crecimiento de la propiedad de la vivienda en conjunto con los préstamos hipotecarios, que el crédito al consumo se ha generalizado tanto. Este crecimiento en los préstamos de consumo no podría haber sido posible sin un

enfoque automatizado en la evaluación del riesgo de crédito, dando origen a la puntuación crediticia.

El puntaje de crédito se inició en la década de 1950 cuando los investigadores asociaron que los métodos estadísticos de clasificación, inicialmente el análisis discriminante (Fisher, 1936), podrían ser utilizados para clasificar los préstamos en buenos (no morosos) y malos (morosos) con las características del préstamo y de los prestatarios.

En un principio fue utilizado por las empresas de venta por correo y las casas financieras y fue hasta la llegada de las tarjetas de crédito que los bancos comenzaron a usarlo en primer lugar para las tarjetas de crédito, préstamos personales y finalmente, para las hipotecas (Thomas, 2009). Este uso inicial de la puntuación crediticia, a la cual se le conoce como puntuación de aplicación, apoya la decisión de conceder o negar un crédito a un nuevo solicitante. Su filosofía es pragmática, busca predecir no explicar empleando cualquier otra característica de mejora en la capacidad de discriminación del sistema. Por otra parte, se concentra en un riesgo muy específico - la posibilidad de que un prestatario tendrá 90 días de atraso en sus pagos de los próximos 12 meses. Si el préstamo resulta rentable para la entidad crediticia; si el prestatario continuará pagando más allá de este período, la cantidad empleada por el prestatario de la línea de crédito, ninguno de estos riesgos se consideran. El enfoque también supone que la relación entre las características del préstamo / prestatario y la solvencia se mantiene estable al menos durante un período de cuatro o cinco años (Mester, 1997). Con datos de los solicitantes de dos años anteriores y observándose su desempeño en el año siguiente. Este resultado se utiliza para determinar si el solicitante es malo (el riesgo específico ocurre) o bueno (no ocurre). Esta forma se emplea para construir un sistema de clasificación que separa los buenos de los malos prestatarios utilizando las características del préstamo y del prestatario. Los métodos de clasificación estándar resultan en una tarjeta de clasificación, que se construye sobre los datos de dos años anteriores y se emplea para cuales solicitantes son aceptados en los próximos años. Después de un tiempo el proceso se repite creándose una nueva tarjeta de puntos.

La segunda variante de la puntuación de crédito, puntuación de comportamiento, se introdujo en la década de 1980 al considerarse útil para evaluar el riesgo crediticio de los

clientes existentes, así como los nuevos solicitantes (Thomas, 2009). De igual forma la variable objetivo es saber sí el prestatario se retrasará en los próximos 12 meses, pero ahora es posible utilizar la información reciente (generalmente 12 meses) de los reembolsos y desempeño de compras del prestatario. Tales resultados son utilizados por casi todos los prestamistas y rutinariamente se actualizan cada mes. Las características más poderosas son si los prestatarios han sido morosos recientemente y la información actualizada de la oficina de crédito en su desempeño global de crédito (Mester, 1997). A pesar de que la puntuación de comportamiento es una extensión obvia de la puntuación de aplicación también se considera una oportunidad desaprovechada. En primer lugar, no se utiliza para apoyar una decisión específica sino que el prestamista la emplea como parte de una estrategia de relaciones con los clientes para determinar si hay que aumentar los límites de crédito, intentando elevar las ventas o la venta cruzada de otros productos. El objetivo de estas acciones es sin embargo para mejorar la rentabilidad del cliente, pero puede haber otras medidas en lugar de riesgo de impago en los próximos 12 meses, lo que da un mejor manejo de los beneficios. De igual forma la puntuación de comportamiento solamente emplea características estáticas del desempeño pasado de los clientes para estimar su estatus en un tiempo fijo futuro. Un alternativa puede ser construir un modelo dinámico de cómo ha actuado un cliente que permita una predicción de su comportamiento futuro (Lahsasna, Ainon, & Wah, 2010).

En los últimos años, la puntuación de crédito ha cambiado de acuerdo a los intereses de los prestamistas, ahora como soporte de sus objetivos de negocio en cuanto a la rentabilidad y las cuotas de mercado. Los prestamistas quieren optimizar todas las decisiones que toman con respecto al prestatario independientemente si se le ofrece o no un préstamo estándar (Kocenda & Vojtek, 2009). Incluso en la decisión inicial, los prestamistas tienen ahora una serie de variantes de un producto de crédito que pueden ofrecer, ya sea de platino, oro, plata o tarjetas de crédito estándar, o de crédito variable, tipo de interés fijo, y las hipotecas a tipo variable, y dentro de cada uno de ellos pueden decidir ¿Qué límite de crédito para ofrecer y qué tipo de interés y el pago (los componentes del precio) a cargo? Adicionalmente, el crecimiento de la Internet y el teléfono como formas de llevar a cabo el proceso de solicitud significa que las aplicaciones son esencialmente privadas, por lo que el producto se puede

"personalizar" dependiendo de las características de los solicitantes, teniendo en cuenta cargos variables. Del mismo modo los prestamistas tienen más probabilidades de ajustar el producto o la oferta de productos alternativos o adicionales durante su relación con el cliente y por lo tanto desean conocer el impacto que estos cambios tendrán sobre el riesgo de impago y la rentabilidad del cliente (Lahsasna, et al., 2010). Los prestamistas quieren usar la "puntuación crediticia" para ayudar a tomar decisiones de precios variables y determinar la rentabilidad a largo plazo de un cliente bajo diferentes acciones que emprendan los propios prestamistas. Por otra parte la rentabilidad se trata tanto de mercadotecnia como de la evaluación de riesgos y por lo tanto existe la necesidad de combinar el trabajo realizado por las organizaciones de mercadotecnia financiera y la evaluación de riesgos de los grupos de investigación de operaciones. En la actualidad estos grupos se ven como adversarios con un grupo que quiere tomar tantos aspirantes como le sea posible y el otro que trata de ser todo lo exigente posible sobre los que se seleccionan. Los modelos utilizados por los vendedores para segmentar a los clientes y para estimar sus tendencias de adquisición de créditos son muy similares a los utilizados por los equipos de evaluación de riesgo para determinar el número de diferentes tarjetas de puntuación desarrollar y entonces estimar la probabilidad de incumplimiento para cada cliente.

Otro factor que ha afectado la puntuación crediticia en los años recientes, es el cambio en las regulaciones bancarias introducidas por el documento Convergencia internacional de medidas y normas de capital, más comúnmente conocido como el Acuerdo de Basilea II (Basel, 2005, comprehensive version 2006). En la actualidad ya se encuentran en fase de implementación las resoluciones de los tratados del Acuerdo de Basilea III (Basel, 2010). Bajo las regulaciones de Basilea II, los bancos están autorizados a utilizar las estimaciones de sus propios sistemas internos de calificación de riesgo en la fórmula que determina el capital mínimo que han de dejar de lado para cubrir el riesgo de crédito en sus préstamos (Thomas, 2009). Es evidente que los préstamos a los consumidores a través de estos sistemas internos de calificación de riesgo son tanto de aplicación como de comportamiento. En efecto, los únicos casos en que a los bancos e instituciones crediticias les resulta conveniente mudarse a estos sistemas internos de puntuación, es sí los emplean para el crédito al consumo, debido a que principalmente el ahorro en capital se compara con

la alternativa impuesta externamente con los índices de capital en los prestamos de consumo. El Acuerdo requiere que las puntuaciones consideren muchas de las propiedades de los sistemas de puntuación de crédito existentes. Por ejemplo, define el valor predeterminado de 90 días de atraso en los próximos 12 meses (aunque algunos reguladores nacionales, como la Autoridad de Servicios Financieros en el Reino Unido habían modificado esta a 180 días de atraso). Sin embargo, requiere de igual forma un énfasis en validar la probabilidad de las estimaciones de impago, no solo en los próximos 12 meses sino a largo plazo con un hincapié en la necesidad de probar los modelos bajo presión y la implementación de cálculos completamente nuevos, como las pérdidas derivadas por el incumplimiento de pago, más allá de sólo asegurar la clasificación exacta de los deudores con respecto a la forma en que fueron juzgados previamente por los sistemas de puntuación de crédito. Resulta importante resaltar que en general en Latinoamérica y en México en particular, a excepción de las instituciones bancarias y financieras transnacionales los puntos del acuerdo recién comienzan a implementarse.

II.3.2 Cálculo de confiabilidad en los modelos de puntuación

Al desarrollarse modelos de puntuación crediticia el cálculo de su confiabilidad resulta imperativo. La forma más usual de realizar esta ponderación es mediante una matriz de clasificación como se muestra en la tabla II.1 (Boguslauskas & Mileri, 2009). En esta matriz la información sobre la confiabilidad de los clientes se codifica como sigue "0" cuando el cliente ha liquidado sus obligaciones financieras (el cliente es bueno) y "1" cuando el cliente ha tenido problemas de solvencia (el cliente es malo).

Tabla II-1 Matriz de Clasificación para determinar la confiabilidad de los clientes. 0=clientes buenos, 1=clientes malos.

Modelo	*Real*	
	1	*0*
1	Verdadero Positivo(VP)	Falso Positivo (FP)
0	Falso Negativo (FN)	Verdadero Negativo (VN)

Empleando los datos de la matriz de clasificación, el cálculo de la confiabilidad se puede estimar de acuerdo a lo que muestra la tabla II.2. El significado de las razones es:

- Índice de clasificación correcta (ICC): clientes clasificados correctamente, donde N es el número de clientes analizados.

- Índice erróneo de clasificación (IEC): clientes clasificados incorrectamente.
- Índice de falso negativo (Error tipo I) (α): clientes malos clasificados como buenos.
- Índice de falso positivo (Error tipo II) (β): clientes buenos clasificados como malos.
- Sensibilidad (Se): clasificación correcta de clientes malos.
- Especificidad (SPEF): clasificación correcta de clientes buenos.
- Valor predictivo positivo (VPP): porcentaje de clientes clasificados como malos que realmente son malos.
- Valor predictivo negativo (VPN): porcentaje de clientes clasificados como buenos que realmente son buenos.
- Confiabilidad (Conf): es una razón no comúnmente empleada de los aciertos diagnosticados por el modelo tanto de buenos clientes como de los malos.

Para validar de forma simple el desempeño de estas pruebas se puede emplear la medida-F. La medida-F es la media armónica del valor predictivo positivo y la sensibilidad.

Tabla II-2 Índices para el cálculo de confiabilidad de los modelos crediticios.

Índice	*Cálculo*
Índice de clasificación correcta	ICC = (VP + VN) / N
Índice erróneo de clasificación	IEC = (FP + FN) / N
Índice de falso negativo	α = FN / (FN + VP)
Índices de falso positivo	β = FP / (FP + VN)
Sensibilidad	Se = VP / (VP + FN)
Especificidad	SPEF = VN / (VN + FP)
Valor predictivo positivo	VPP = VP / (VP + FP)
Valor predictivo negativo	VPN = VN / (VN + FN)
Confiabilidad	Conf = (VP + VN) / (Pos + Neg)
Medida-F	$F = \frac{2 \cdot VPP \cdot Se}{VPP + Se}$

Para el análisis gráfico de la estimación de confiabilidad de los modelos de riesgo crediticio frecuentemente se emplea la curva Característica Operativa del Receptor (ROC por sus siglas en inglés receiver operator characteristics), la cual evalúa los resultados de la

clasificación de objetos en dos grupos, cuando se cuenta con un conjunto de entrenamiento (ver figura II.1).

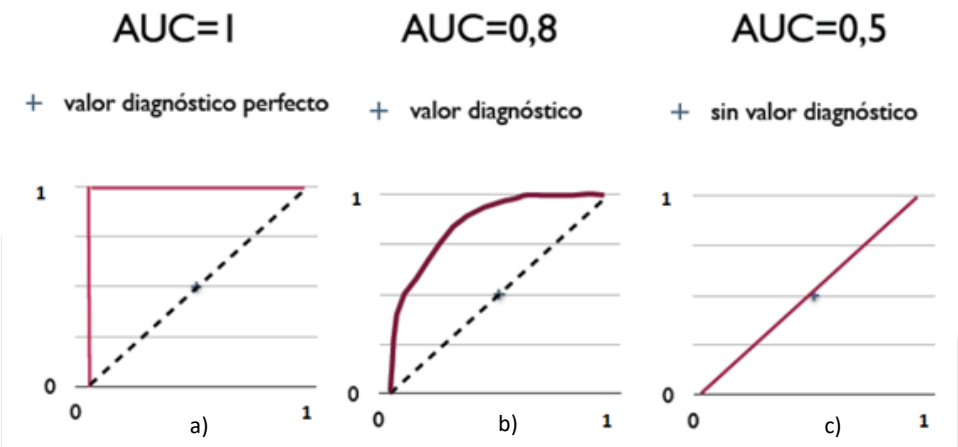


Figura II-1 Ejemplos de Curvas ROC. a) Modelo 100% confiable. b) Modelo con cierto grado de confiablidad. c) Modelo sin capacidad de clasificación.

En el caso de modelos perfectos de riesgo crediticio la curva ROC alcanza la esquina superior izquierda de la gráfica. Esto significa que todos (100%) los clientes no confiables se clasificaron correctamente, es decir la sensibilidad del modelo es máxima. En el caso de clasificación incorrecta parte de los clientes son iguales a 0. Así que la cercanía de la curva a la esquina superior izquierda del gráfico denota la precisión del modelo. Y a la inversa, una curva menos arqueada es más cercana a la diagonal denotando un modelo menos eficaz. La diagonal de la curva refleja un modelo sin ningún valor de diagnóstico, es decir, la imposibilidad absoluta de discriminar clases. Las curvas ROC permiten comparar la eficiencia de diferentes modelos. Una mayor cercanía a la esquina izquierda de la gráfica denota la mejor confiabilidad del modelo. Si las curvas ROC de diferentes modelos son muy cercanas y se cruzan, resulta muy complejo evaluar la eficiencia de los modelos visualmente. En estos casos la comparación se realiza en las áreas bajo las curvas ROC (AUC por sus siglas en inglés), las cuales tienen un valor entre 0 y 1. Sin embargo un modelo es aplicable solo si su curva ROC se encuentra por encima de la diagonal, por lo que la eficiencia de los modelos se evalúa de 0.5 (modelos sin valor) a 1 (modelos perfectos). Estos valores se pueden calcular estimando el área bajo la curva ROC.

II.4 Modelos de puntuación crediticia en las Instituciones Microfinancieras.

II.4.1 Las microfinanzas.

Las microfinanzas se refieren a los servicios financieros otorgados a personas de bajos ingresos de poblaciones que normalmente no tienen acceso al ahorro, crédito, información financiera y los seguros. Históricamente, a los pobres se les ha negado el acceso al crédito, y por lo tanto no pueden darse el lujo de gastar sus ingresos y consumo en la actividad empresarial.

Las microfinanzas tienen como objetivo fundamental impulsar la creación y el desarrollo de pequeñas actividades productivas. En términos generales se concibe a las microfinanzas como un tipo de financiamiento a pequeña escala, pero con la característica de que es para familias pobres. Así, las microfinanzas es el desarrollo de las finanzas al servicio de una población excluida del sistema tradicional capitalista (Hernández Romero & Almorín Albino, 2005).

Durante mucho tiempo se había pensado que las personas de escasos recursos no podían por su condición, obtener servicios financieros de calidad, porque ha persistido el mito de que la gente pobre presenta altos riesgos crediticios pues son personas que no pueden pagar o son incapaces de ahorrar, debido a que tienen demasiadas carencias para hacerlo (Conde Bonfil, 2000; Mansell Carstens, 1995).

Sin embargo, las familias de bajos ingresos quieren y pueden ahorrar, y lo hacen cuando tienen a su alcance instituciones e instrumentos apropiados a sus peculiaridades. Tales innovaciones, que se llevan a cabo en las microfinanzas, han contribuido a generar oportunidades de desarrollo y bienestar a las familias que participan en los proyectos de microfinanzas (Escalona Cortés, 2011).

Las actividades microfinancieras no son un fin en sí mismas, sino que son un instrumento que apoya otras actividades o metas básicas de las organizaciones; se centra en la mayoría de los casos en fortalecer la economía y oportunidades de desarrollo de las personas que viven en situación de pobreza, convirtiéndose en uno de los instrumentos más importantes de la política social de acceso a los servicios financieros, que trata de romper el círculo

vicioso en el que un individuo no puede comenzar a ser productivo por no conseguir un financiamiento para impulsar un pequeño esfuerzo empresarial porque no tiene pertenencias ni aval para garantizar el préstamo.

Han transcurrido aproximadamente 46 años desde que las microfinanzas se originaron y la evolución cada vez mayor de la industria ha provocado que las instituciones micro financieras (IMFs) se vean obligadas a ser reguladas y supervisadas por los organismos competentes de los sistemas financieros de las naciones o regiones donde tiene lugar su actividad, además de que su actividad ya no dista demasiado en relación a las entidades financieras bancarias. A este respecto la normativa impuesta por los gobiernos abre camino a la normalización de instituciones informales hacia el desempeño de labores y actividades que favorezcan la consecución de los objetivos marcados por las microfinanzas (Lara Rubio, 2010). El microcrédito por esta razón es considerado el principal producto que permite el acceso a los servicios financieros a un número elevado de microempresarios. La figura II.2 generada a través de la información recabada por Mixmarket.org sobre los activos, préstamos y cantidad de deudores de las IMFs por países en el 2005 y 2010, muestra la participación a nivel mundial de las instituciones microfinancieras.

Es claro que la actividad microfinanciera es más intensa en América Latina y Asia donde la situación económica privilegia la creación de instituciones financieras de esta índole. África donde la extrema pobreza y la inestabilidad social son muy graves, no garantiza las condiciones para que las IMFs puedan prosperar a excepción de Sudáfrica que si cuenta con una gran cantidad de actividad microempresarial.

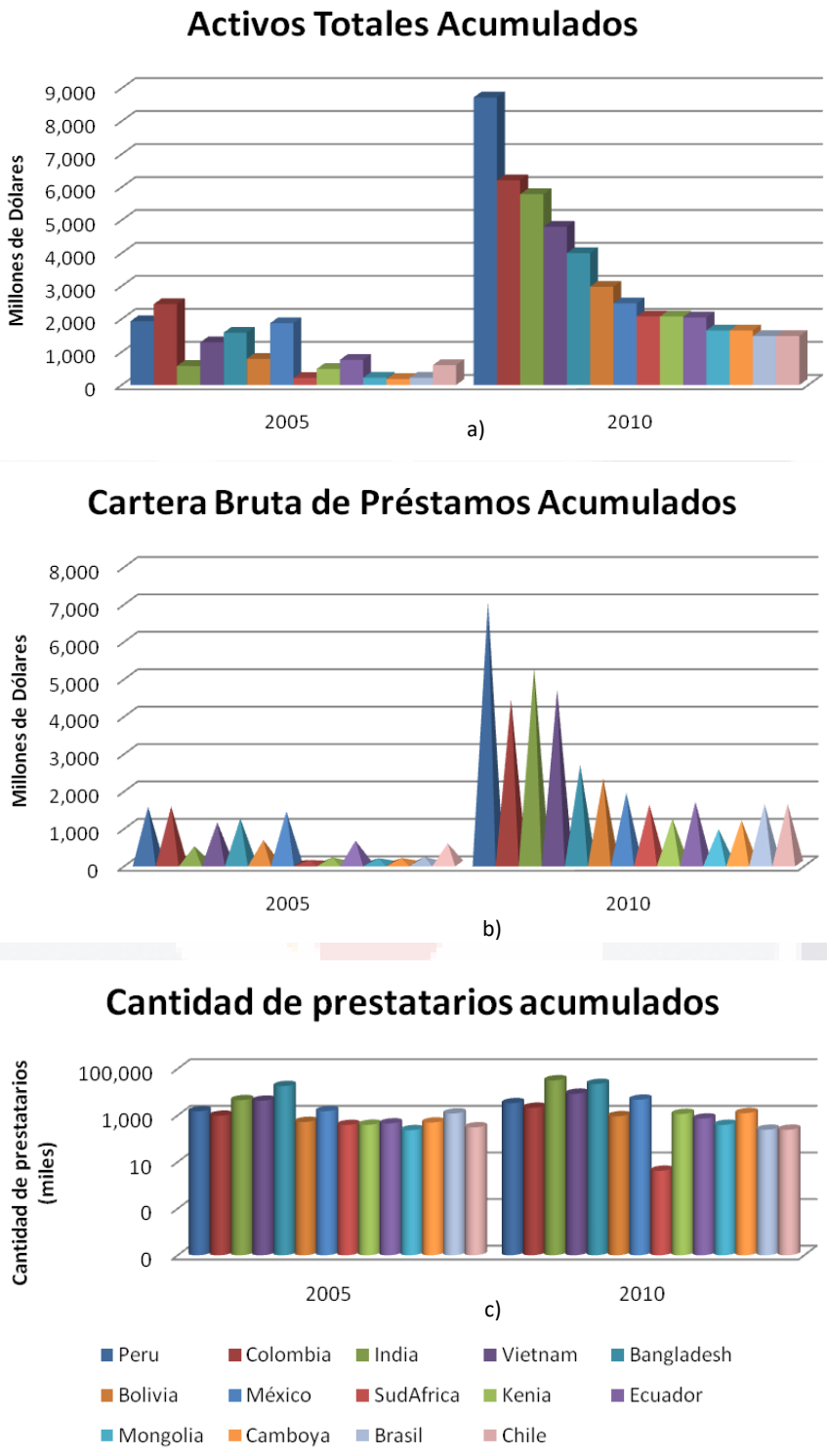


Figura II-2 Muestra informativa de las IMFs en los catorce países con mayor cantidad de activos acumulados al 2010. a) Cantidad de Activos en millones de dólares. b) Cartera bruta de préstamos en millones de dólares. c) Cantidad de deudores en miles.
(Datos tomados de www.mixMarket.org, noviembre del 2011).

II.4.2 Las microfinanzas en México

Al igual que otros países del mundo, en México, las microfinanzas cada vez adquieren un lugar importante en el financiamiento hacia las clases más desprotegidas, ello se nota en la gran cantidad de actores involucrados en diversos proyectos, programas y eventos dirigidos hacia el microfinanciamiento.

Las instituciones de microfinanzas en México, a diciembre de 2010, atendieron a 5’576,433 personas con productos de crédito, ahorro y microseguros. De estos, 5’401,921 son clientes activos de crédito. En total se cuenta con 238,992 ahorradores. En promedio, las instituciones de microfinanzas atienden a 96,000 clientes activos (sin tomar en cuenta 27,000 casos extremos) donde 80% son mujeres y 53% vive en zonas rurales. (ProDesarrollo Finanzas y Microempresa, 2011).

De las personas atendidas, 35% son por crédito individual y 65% mediante alguna metodología de crédito solidario (banca comunal o crédito grupal). El total de la cartera es de 24,544 millones de pesos. El crédito promedio de las instituciones microfinancieras es de $6,317 pesos. En total, las instituciones de microfinanzas emplean a 34,640 personas.

En lo referente a las instituciones de microfinanzas, su antigüedad promedio es de 7 años; esto significa que el sector es aún joven y, a pesar de ser visto por algunas instituciones líderes como de alto retorno sólo 61% de las instituciones son sostenibles financieramente (muchas no han logrado la sostenibilidad financiera por el poco tiempo que llevan operando). El promedio de retorno sobre activos de una institución microfinanciera es de 2.6%.

En cinco años, el número de clientes de crédito casi se ha quintuplicado, mientras el número de ahorradores apenas se ha duplicado, lo que da cuenta del acelerado ritmo del crecimiento del sector. La continuidad en la creación de nuevas instituciones así como la concentración por clientes atendidos sigue siendo alta: 82% de los clientes son atendidos por las 5 entidades más grandes; 11% está atendido por 10 instituciones y el 7% restante está atendido por 38 instituciones.

La cartera bruta de préstamos presenta un comportamiento similar en su distribución. Habría que destacar que todavía no se cuentan con indicadores que permitan conocer el grado de clientes "compartidos" entre dos o más instituciones. En la medida en que se utilicen sociedades de información crediticia podrá conocerse mejor este tema. La percepción es que en las zonas donde hay una mayor cantidad de instituciones trabajando, hay un mayor número de clientes que mantienen créditos abiertos de manera simultánea sobre todo en metodologías grupales.

Hay diversidad en los productos en cuanto a plazos, montos y precio otorgados por las instituciones. La más común es semanal: 50%; 25% quincenal; 8%, mensual y, 9% catorcenal. Los plazos van desde 3 hasta 18 meses. Las tendencias indican que existe un mayor desplazamiento hacia la especialización; 11 instituciones (18%) tienen al menos un producto especializado de crédito a la vivienda o mejora de ésta que en conjunto suman una cartera de $1,611,172,779.°° (7% de la cartera bruta total). 20% de las instituciones ofrecen algún tipo de seguro (en su mayoría es de vida) independiente al del saldo deudor (ProDesarrollo Finanzas y Microempresa, 2011).

En lo referente al marco regulatorio comúnmente se habla de microfinancieras o instituciones de microfinanzas, aún cuando en sí no se trate de algún tipo de institución en particular, sino una diversidad de instituciones que tienen en común el ofertar servicios financieros para las microempresas. Desde hace unos años las instituciones de microfinanzas, tanto en la figura jurídica como en los lineamientos de toma de decisiones, han tendido a ser más comerciales. La protección de los clientes, el cuidado del no sobreendeudamiento y la educación financiera son los principales retos que se plantean. Hoy, sólo 13% de las instituciones de microfinanzas son sin fines de lucro, dicha tendencia está marcando la incorporación de las entidades microfinancieras al sistema financiero formal, lo cual puede resultar factible e incluso conveniente, siempre que las que así lo hagan, cuenten con una posición razonablemente sólida desde el punto de vista organizativo y técnico y así poder cumplir con los requisitos de regulación exigidos, además de que los organismos de supervisión y regulación tengan en cuenta la naturaleza y

características peculiares de este tipo de entidades y establezcan un régimen de regulación y supervisión ad hoc.

Los intermediarios financieros que operan en México están regulados y supervisados por la Secretaría de Hacienda y Crédito Público (SHCP), el Banco de México y la Comisión Nacional Bancaria y de Valores (CNBV) que en conjunto con la Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros (CONDUSEF), registran desde su figura jurídica este tipo de instituciones pese a no existir un padrón único de éstas (Lara Rubio, 2010).

La mayoría de las instituciones de microfinanzas corresponden a la figura de Sociedad Financiera de Objeto Múltiple (SOFOM, figura que no está regulada excepto por la ley de transparencia), sin embargo un número de clientes considerable son atendidos por un banco y algunas Sociedades Financieras Populares (SOFIPOS). Si bien todavía hay asociaciones civiles que realizan las actividades de microfinancieras, la tendencia es hacia su desaparición.

De acuerdo con la lista de la CONDUSEF, existen 2,779 SOFOMES, de las cuales operan 2,105 (75%). Alrededor de unas 240 son de créditos personales, de nómina o microcréditos (11%), aunque es difícil saber, ya que del listado de la CONDUSEF, y debido a que son entidades de objeto múltiple, no se tienen clasificadas por especialidad. Las SOFIPOS, figura regulada que puede captar ahorro, son 37 y las Sociedades Cooperativas de Ahorro y Préstamos (SOCAPS) supervisadas por la CNBV son 57; por su parte, la Confederación de Cooperativas de Ahorro y Préstamo de México (CONCAMEX) agrupa 210 cooperativas (incluyendo las autorizadas) en 18 federaciones. En tanto que las instituciones de microfinanzas, el total es de 208, si se consideran las instituciones fondeadas por algunas instituciones de gobierno, la banca de desarrollo y las asociadas al organismo Prodesarrollo (ProDesarrollo Finanzas y Microempresa, 2011).

La figura II.3 muestra las diez IMFs de mayor alcance en México, en cuanto a la cartera bruta de préstamos y cantidad de prestatarios se refiere en datos recabados hasta el 2010.

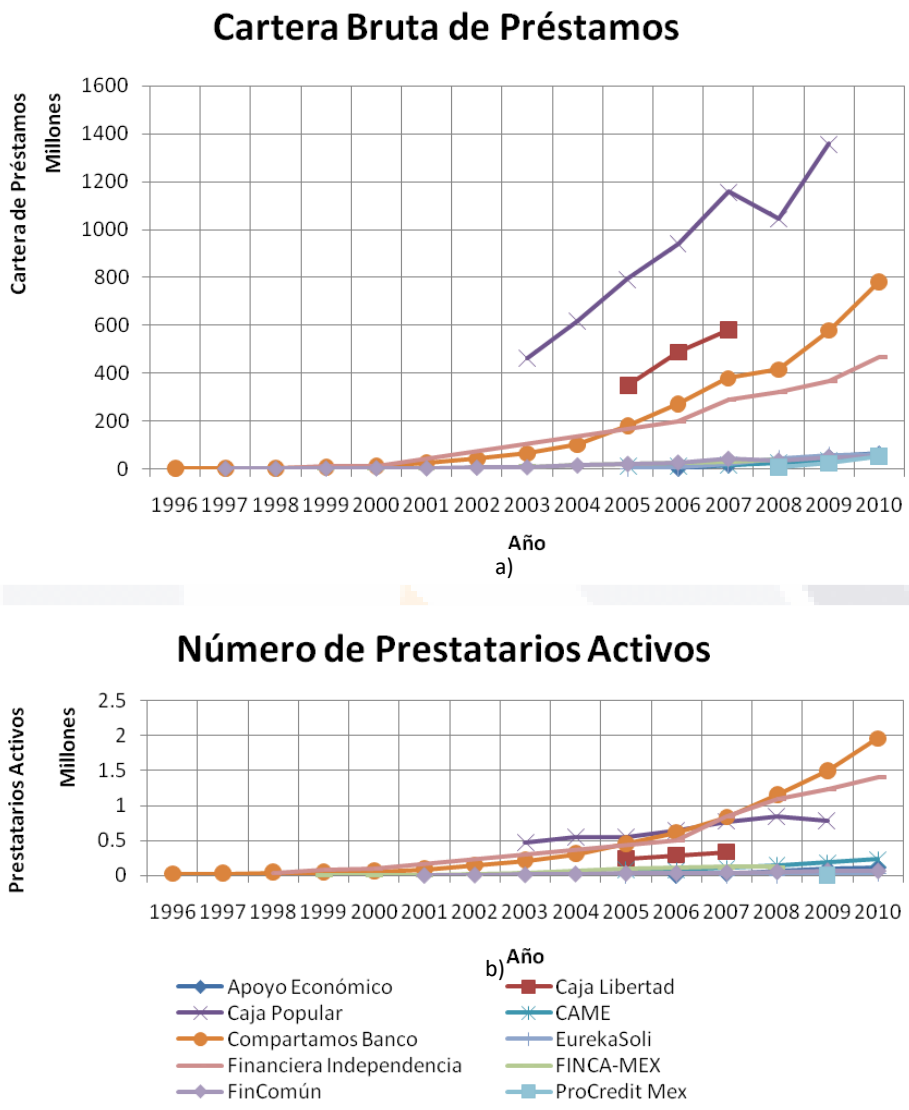


Figura II-3 Muestra de las 10 IMFs mexicanas con mayor actividad económica al 2010. a) Cartera bruta de préstamos. b) Número de prestatarios activos.
(Datos tomados de www.mixMarket.org, noviembre del 2011).

II.4.3 Limitaciones de la puntuación crediticia en microfinanzas

La cuestión fundamental que actualmente se plantea en los trabajos de investigación sobre puntuación crediticia es si los modelos estadísticos de riesgo de crédito son aplicables o no a las instituciones de microfinanzas. Schreiner (Schreiner, 2001) indica que aunque la calificación es menos enérgica en los países pobres que en los países ricos, y aunque la calificación en las microfinanzas no reemplazará el conocimiento personal del carácter por parte de los analistas de crédito y de los grupos de crédito, la calificación puede mejorar las estimaciones del riesgo y disminuir los costos. Por tanto, indica el autor, la calificación

complementa, pero no sustituye, las tecnologías que actualmente se aplican en las microfinanzas.

La construcción de un modelo de puntuación crediticia para las instituciones de microfinanzas implica una serie de limitaciones y desventajas que incrementan la dificultad para llegar a unos resultados razonables, en general un modelo de puntuación crediticia emplea medidas cuantitativas de los resultados y características de los préstamos pasados para predecir el rendimiento futuro de los préstamos con características similares. Para las organizaciones de crédito en los países ricos la puntuación crediticia ha sido una de las herramientas más importantes en el logro de una mayor eficiencia. La información usada en estos modelos, en países ricos, se basa en la proporcionada por el historial crediticio (buró de crédito) y en la experiencia laboral (salario) del cliente. En microfinanzas, sin embargo la mayoría de los acreditados son pobres y trabajan por su cuenta propia (Schreiner, 2001).

Si lo que se pretende es analizar el comportamiento de pago de un cliente de microcrédito, se requiere una base de datos amplia que recoja la historia de los préstamos que han resultado impagados en algunas de sus cuotas desde su concesión hasta su fecha de vencimiento. Por otro lado, las IMFs normalmente no incluyen en sus bases de datos la información referente a los clientes a los que se les denegó el crédito porque, en su momento, no pasaron la evaluación estándar del analista, además de que generalmente se otorga el crédito. En consecuencia, solo se podrá contar con la información de aquellas solicitudes que están en la fase de aprobación.

Las microfinanzas requieren la intervención de un asesor de crédito en la recopilación y captación de información para el historial de crédito. Sin embargo, este proceso podría estar muy influenciado por la opinión subjetiva de dicho funcionario.

Los datos recogidos por el analista u oficial de crédito han de ser integrados correctamente en el sistema de información de gestión de la entidad de microfinanzas. Esta labor es delicada y requiere un programador u operario destinado exclusivamente a tal función. La razón es simple: si se cometen errores en esta fase, o se demora la introducción de los datos en un tiempo razonable, el modelo diseñado perdería su efectividad.

El incumplimiento de pago de igual forma tiene variantes con respecto a la forma en que se le trabaja en la banca tradicional donde los atrasos se consideran entre 90 y 180 días según se definió anteriormente, en el caso de las IMFs el incumplimiento de pago debe definirse con cautela, por lo que es necesario identificar todo atraso que conlleve un costo para la organización. Para ello se han de verificar las siguientes condiciones (Rayo Canton, et al., 2010):

a) El atraso percibido ha de ser real y no estimado, según fechas concretas marcadas en la contratación del crédito, en función del método estipulado para su amortización por las partes contratantes.
b) El atraso ha de producirse en, al menos, una cuota de amortización del microcrédito.
c) El atraso considerado ha de suponer un incremento en el coste para la entidad más que proporcional al habitual en caso de no sucederse esta contingencia. Generalmente, estos incrementos suelen darse en términos de costes administrativos debido al incremento monetario que supone realizar un seguimiento y gestionar el pago de un crédito cuyo reembolso mantiene un retraso considerable.

En el modelo de calificación riesgo de morosidad para los créditos de una organización de microfinanzas en Bolivia, (Schreiner, 1999) establece el atraso costoso como un “atraso de 15 días o más”, sin argumentar los motivos en los que se basa para determinarlo.

En general, se puede observar que la problemática de implementación de un sistema de puntuación crediticia de en las IMFs es de un grado de complejidad mucho mayor al que se presenta en instituciones bancarias y de crédito en el primer mundo pese a manejar montos de crédito muy inferiores.

III

Capítulo III La minería de datos y la puntuación crediticia.

III.1 Uso de minería de datos y aprendizaje automático en puntuación crediticia.

El gran auge que se ha dado en las entidades crediticias y el número de prestatarios que aplican para obtener créditos de casi cualquier índole, tornaron imposible en términos económicos como operativos realizar otra acción diferente a la de automatizar los procesos de otorgamiento de créditos (Thomas, 2000). En años recientes, varios métodos cuantitativos se han propuestos para la evaluación de la puntuación crediticia. Los modelos de puntuación crediticia se pueden dividir en dos categorías: los modelos tradicionales, que incluyen los modelos de juicio y estadísticos, así como los modelos novedosos, que generalmente son modelos no-paramétricos que están soportados por técnicas de inteligencia artificial, tanto los modelos estadísticos como los novedosos se ubican dentro del marco de la minería de datos y los algoritmos de aprendizaje automático (machine learning).

El campo de la minería de datos aborda la cuestión de como emplear los datos históricos de la mejor manera para descubrir las regularidades generales y mejorar el proceso de toma de decisiones, motivos por los cuales a la minería de datos se le ha definido también como "descubrimiento de conocimiento" o "análisis de datos avanzado", siendo la combinación de la inteligencia artificial, el aprendizaje automático, las estadísticas y los sistemas de bases de datos (Figura III.1) las disciplinas encargadas de descubrir nuevos patrones (Han & Kamber, 2006), a partir de los grandes conjuntos de datos generados en este caso particular, por los clientes aspirantes a créditos.

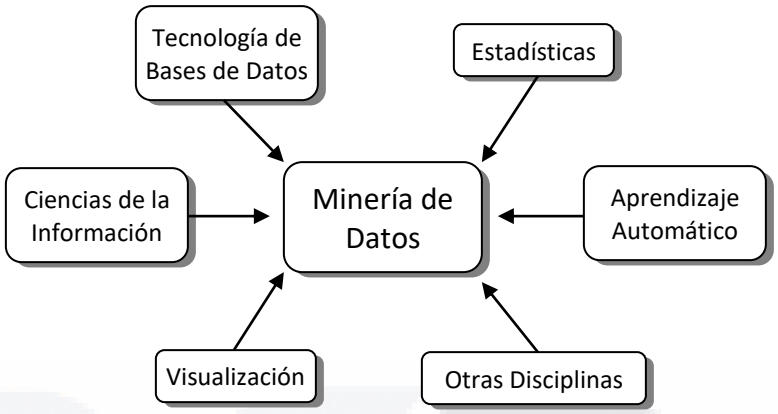


Figura III-1 Confluencia de varias disciplinas en minerías de datos.
(adaptado de (Han & Kamber, 2006)).

Aunque los algoritmos de aprendizaje automático son parte medular en los procesos de la minería de datos, es importante recalcar que dichos procesos de igual forma involucran otros pasos fundamentales, incluyendo la construcción y el mantenimiento de la base de datos, formatos de datos y la limpieza, la visualización y resumen de datos, el uso del conocimiento y experiencia humana para formular las entradas para el algoritmo de aprendizaje y evaluación de las regularidades empíricas que descubre, y determinar cómo distribuir los resultados (Mitchell, 1999).

III.2 Metodología de minería de datos

Las funciones de la minería de datos se emplean para definir la clase de patrones que se encontrarán en las tareas correspondientes. En general, las tareas de minería de datos se pueden clasificar en dos categorías: La minería predictiva cuyo objetivo es predecir el valor particular de un atributo basado en otros atributos, donde el atributo a predecir es comúnmente llamado “clase” o variable dependiente, mientras que los atributos usados para hacer la predicción se llaman variables independientes. La minería descriptiva, que tiene por objetivo encontrar patrones (correlaciones, tendencias, grupos, trayectorias y anomalías) que resuman relaciones en los datos (Fayyad & Uthurusamy, 2002).

La clasificación en dos categorías es perfectamente entendible si se parte de la idea de que al final el descubrimiento de conocimiento, como todos los métodos y metodologías de la inteligencia artificial parten de la premisa de intentar reproducir el comportamiento del ser

humano en este caso para aprender, aunque en realidad son más los tipos de inferencia más empleados que desarrolla el ser humano y son las siguientes (Handl, 2006):

- Deducción: el razonamiento deductivo aplica una regla de conocimiento preposicional a un caso individual y sus conclusiones son necesariamente ciertas.
- Inducción: un argumento inductivo toma un número de observaciones, consistentes de casos individuales y sus resultados asociados e intenta predecir una regla general que relacione los casos con sus resultados.
- Abducción: el razonamiento abductivo emplea una regla y un resultado observado y desarrolla hipótesis sobre este resultado particular que resulta en una instancia de la aplicación de la regla, además de que el antecedente de la regla debe por lo tanto ser verdad.
- Transducción: es una cuarta forma más reciente de realizar inferencias, la cual a diferencia de la inducción, generaliza de lo observado, casos específicos a casos específicos y no a principios generales. Así, estos métodos de transducción no intentan desarrollar un modelo general que pueda ser usado subsecuentemente para la deducción, sin embargo se emplea la inferencia para caso a caso.

Para alcanzar estas inferencias la minería de datos tiene técnicas funcionales específicas que le permiten llevar a cabo el establecimiento de patrones para el descubrimiento del conocimiento, los cuales se muestran a continuación (Han & Kamber, 2006):

- Caracterización y Discriminación. Los datos se pueden asociar con clases o conceptos, que pueden resultar útiles para describir clases y conceptos individuales en términos resumidos concisos y precisos. Estas descripciones se pueden derivar a partir de (1) la caracterización de los datos, resumiendo los datos de la clase bajo estudio (a menudo nombrada clase objetivo) en términos generales, o (2) la discriminación de los datos, mediante la comparación de la clase objetivo con uno o varios conjuntos comparativos de clase, o (3) la combinación de caracterización y discriminación.
- Asociaciones y Correlaciones. Existen muchas clases de patrones frecuentes, que incluyen conjuntos de elementos, subsecuencias y subestructuras. Un conjunto de elementos frecuente se refiere típicamente a los elementos que usualmente aparecen

juntos en operaciones transaccionales, como la leche y el pan. Un patrón secuencial frecuente o subsecuencia se presenta como el resultado de acciones derivadas de sus antecesoras, como el caso de los clientes que primero compran una PC, seguida de una impresora, una cámara digital y así subsecuentemente. Una subestructura o patrón estructurado se puede referir a diferentes formas estructurales, como gráficas, árboles o mallas que ocurren frecuentemente y se pueden combinar con conjuntos de datos o subsecuencias. La minería de patrones frecuentes puede guiar a descubrir asociaciones y correlaciones interesantes dentro de los datos.

- Clasificación y Predicción. La clasificación es el proceso de encontrar un modelo (o función) que describe y distingue clases o conceptos de datos, con la finalidad de de poder utilizar el modelo para predecir la clase de objetos cuya etiqueta de clase se desconoce (ver Figura III.2). El modelo derivado se basa en el análisis de un conjunto de datos de entrenamiento y puede representarse de varias formas, tales como reglas de clasificación, árboles de decisión, formulas matemáticas o redes neuronales.

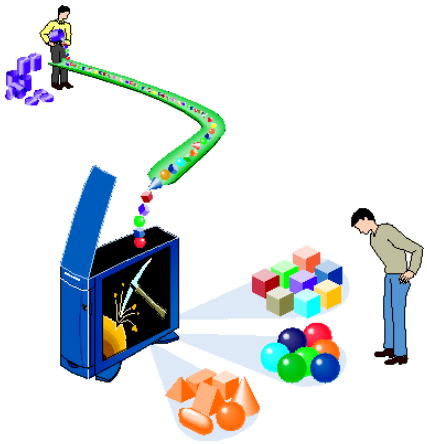


Figura III-2 Idea general de la técnica de clasificación para el establecimiento de patrones en minería de datos.

Mientras que la clasificación predice las etiquetas categóricas (discretas, sin orden), los modelos de predicción lo hacen con funciones de valores continuos, es decir, se emplean para predecir los valores de datos numéricos que faltan o no están disponibles en lugar de etiquetas de clase. Aunque el término predicción puede referirse tanto a la predicción numérica como a la de etiquetas de clase. La predicción también incluye la identificación de las tendencias de distribución basadas en los datos disponibles.

La clasificación y la predicción pueden requerir un análisis de relevancia, para identificar atributos que no contribuyan y que puedan excluirse de los procesos de clasificación o predicción.

- Análisis de Agrupamientos (Cluster). A diferencias de la clasificación y predicción, los cuales analizan los datos de los objetos con clases etiquetadas, el análisis de agrupamiento lo hace sin el empleo de clases etiquetadas conocidas. El agrupamiento puede emplearse para generar dichas etiquetas. Los objetos se agrupan basándose en el principio de maximizar la similitud intraclases y minimizarla en extraclases (Figura III.3). Cada agrupamiento formado se puede ver como una clase de objetos a partir de las cuales se pueden derivar reglas. El agrupamiento puede también facilitar la formación taxonómica, es decir, la organización de observaciones en una jerarquía de clases que agrupan eventos similares juntas.

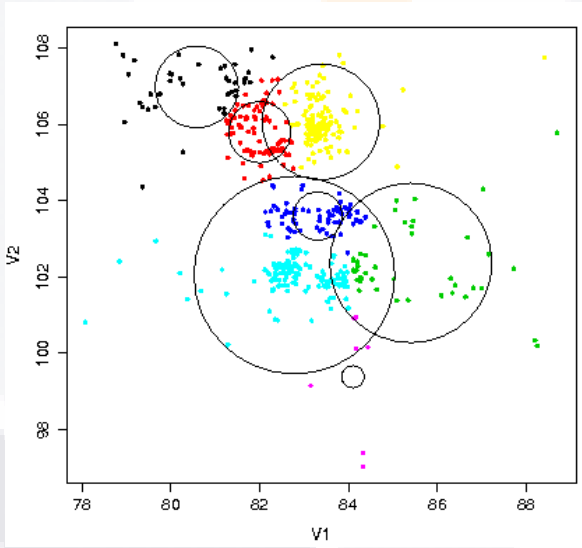


Figura III-3 Representación de agrupamiento de datos de acuerdo a la similitud guardada entre ellos.

- Análisis de anomalías (outlier): una base de datos puede contener objetos de datos que no cumplan el comportamiento general o modelado de los datos, a los cuales se les conoce como anomalías. La mayoría de los métodos de minería de datos descartan las anomalías como ruido o excepciones. Sin embargo, en algunas aplicaciones como en la detección de fraude, los eventos raros pueden resultar más interesantes.

- Análisis de evolución: El análisis de evolución de datos describe y modela regularidades o tendencias para objetos cuyo comportamiento cambia con el tiempo, aunque puede incluir todas las funcionalidades anteriores.

Con la idea de conducir los análisis de descubrimiento de conocimiento en minería de datos los procesos comúnmente se definen en las siguientes etapas (1) selección, (2) preprocesamiento, (3) Transformación, (4) Minería y (5) Interpretación/Evaluación (ver Figura III.4).

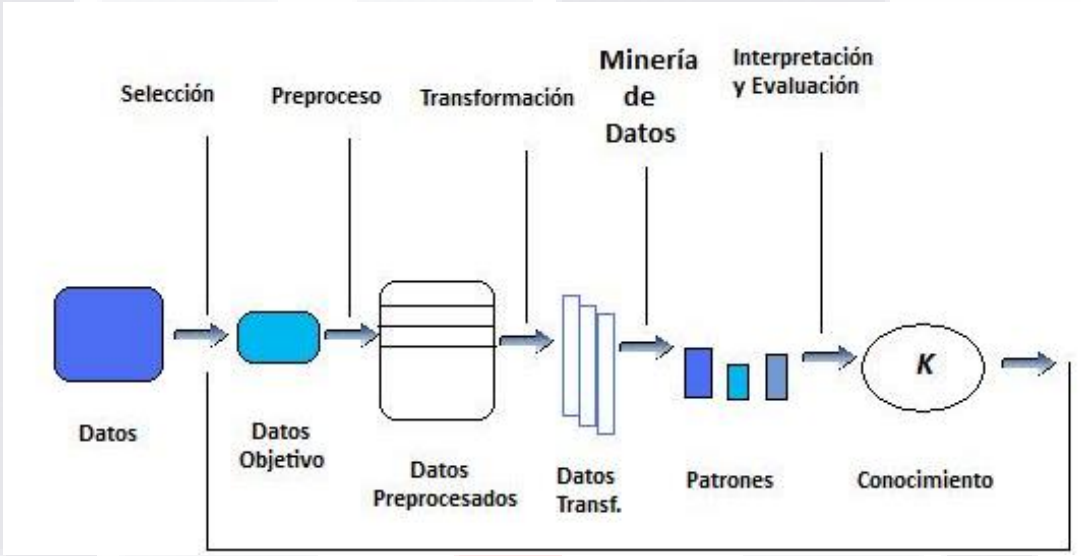


Figura III-4 Proceso del Descubrimiento de Conocimiento.

En la actualidad existen diversas variantes de este tema, siendo una de las más comunes en la actualidad el modelo de referencia CRISP-DM que proporciona una descripción del ciclo de vida del proyecto de minería de datos (Olson, 2009), el cual se muestra en la Figura III.5.

Este contiene las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. El ciclo de vida del proyecto de minería de datos consiste en seis fases. La secuencia de las fases no es rígida. El movimiento hacia adelante y hacia atrás entre fases diferentes es siempre requerido. El resultado de cada fase determina que la fase, o la tarea particular de una fase, tienen que ser realizados después. Las flechas indican las más importantes y frecuentes dependencias entre fases, las cuales se describen a continuación:

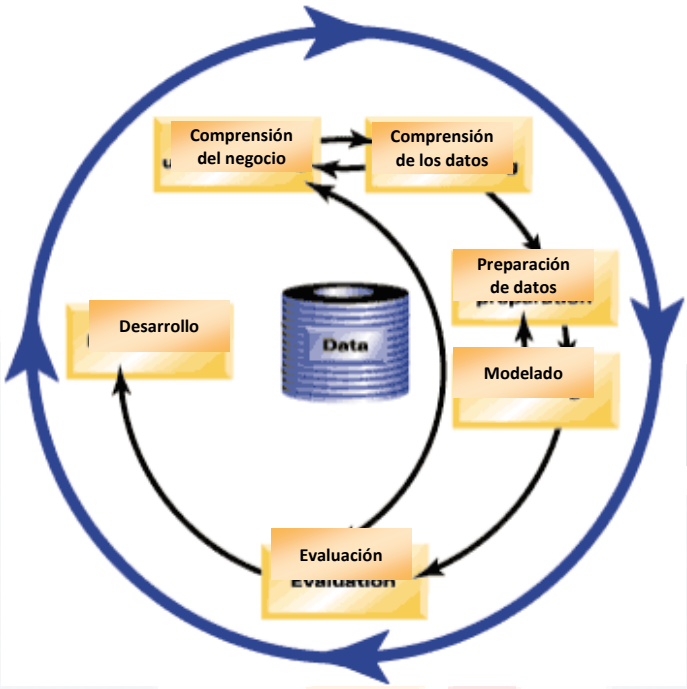


Figura III-5 Fases de la Metodología CRISP-DM 2.0.

- Comprensión del negocio. Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.
- Comprensión de los datos. La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.
- Preparación de datos. La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos en las herramientas de modelado) de los datos en brutos iníciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

- Modelado. En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.
- Evaluación. En esta etapa en el proyecto, se ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos. Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida.
- Desarrollo. La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

El modelado es una de las partes más directamente ligadas con la manipulación de la información dentro de la minería de datos y los procesos de descubrimiento de conocimiento y aprendizaje, existiendo una gran cantidad de algoritmos para desarrollar dichas actividades, los cuales se pueden clasificar en una taxonomía de acuerdo a los resultados esperados de ellos como se describe a continuación:

- Supervisados: los métodos funcionales de clasificación supervisada usualmente son entrenados sobre un conjunto finito de datos de entrenamiento los cuales pueden conducir al fenómeno estadístico de sobre entrenamiento. En la actualidad un vasto rango de métodos supervisados de clasificación están disponibles, los cuales pueden diferir significativamente en su grado de poder explicativo. En particular los sistemas de aprendizaje estadístico tales como Naive Bayes, maquinas de soporte vectorial o redes neuronales, entregan una clasificación sin un intento de explicarla. En contraste los métodos de clasificación basada en reglas como los árboles de decisión, las reglas de asociación, la programación genética o la programación lógica inductiva le brinda al usuario la oportunidad de reconstruir el proceso de clasificación y de interpretar el resultado final (Handl, 2006).
- No supervisados: a menudo se emplea como sinónimo del modelo de funcionalidad de análisis de grupos o agrupamiento, como la identificación de grupos homogéneos de los elementos de datos es una de las tareas principales definidas en este tipo de técnicas clasificación. La clasificación tradicional de los algoritmos de agrupamiento primeramente distinguen entre métodos jerárquicos, de particionamiento y basados en densidad. Donde alguna de las categorías se emplean basándose en los criterios de agrupamiento (implícito o explicito) (Handl, 2006).
- Semi-supervisados: en ciertos escenarios de clasificación puede ser deseable el combinar las fuerzas de las técnicas de clasificación supervisada y no supervisada, o sea el explotar tanto el conocimiento previo de las etiquetas de clases como la distribución desconocida de los datos, las propuestas semi-supervisadas apuntan a esto. A través del uso combinado de datos etiquetados y no etiquetados es posible alcanzar un grado dado de guía externa al algoritmo de clasificación, el cual permita que la estructura intrínseca en los datos sea tomada en cuenta (Handl, 2006).
- Reforzados: inspirado en la psicología conductista, los algoritmos de aprendizaje por refuerzo se refieren a como un agente debiera comportarse en un entorno determinado con el fin de maximizar alguna noción de acumulación de recompensas, generalmente dicho entorno se formula como un proceso de decisión de Markov (MDP, por sus siglas en inglés). Los algoritmos por refuerzo difieren de los supervisados en que nunca se

presentas pares correctos de entrada/salida, ni acciones de corrección sub-óptimas explícitas. Además de que hay un enfoque de desempeño en línea, lo que implica encontrar un equilibrio entre la exploración y explotación de los conocimientos actuales.(Kaelbling, Littman, & Moore, 1996).

En general, la minería de datos se ha convertido en una herramienta analítica útil en todos los aspectos de estudio en seres humanos, por ejemplo la medicina, la ingeniería y la ciencia. Es un medio necesario para hacer frente a las masas de datos que se producen en la sociedad contemporánea. Dentro de los negocios, la minería de datos ha sido especialmente útil en aplicaciones como la detección de fraudes, análisis de crédito, y segmentación de clientes. Este tipo de aplicaciones impactan en gran medida la industria de servicios. La minería de datos proporciona una manera de obtener rápidamente una nueva comprensión basada en gran escala de análisis de datos.

III.3 Modelos de puntuación crediticia más comúnmente empleados.

A continuación se presenta una perspectiva de la forma de operación de los modelos de puntuación crediticia más comúnmente empleados, de acuerdo a la metodología que emplean para su implementación.

Regresión lineal múltiple (Análisis de discriminante). Es una técnica multivariante que permite estudiar simultáneamente el comportamiento de un grupo de variables independientes con la intención de clasificar una serie de casos en grupos previamente definidos y excluyentes entre sí (Fisher, 1936). Para lograrlo se tiene que maximizar la varianza entre los grupos respecto a la varianza dentro del grupo. La siguiente ecuación explica el análisis discriminante:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \qquad \textbf{Ecuación III-1}$$

donde Z es la puntuación discriminante, β son los coeficientes y X son las variables independientes. El análisis de discriminante se puede emplear si la variable independiente es categórica y las variables independientes son métricas (cuantitativas). Con el fin de utilizar análisis discriminante, los datos tienen que ser independientes y con una

distribución normal, además de requerirse una matriz de covarianza para cumplir con el supuesto de variación homogénea (Rencher, 2002). El análisis de discriminante se ha empleado para resolver problemas de clasificación financiera, de negocios, de investigación de mercados (Jo, Han, & Lee, 1997; Kim, Kim, Kim, Ye, & Lee, 2000; Trevino & Daniels, 1995). Para los problemas de puntuación crediticia, muchos investigadores han propuesto y empleado el análisis de discriminante y sus variantes (T.-S. Lee, Chiu, Chou, & Lu, 2006; T.-S. Lee, Chiu, Lu, & Chen, 2002).

La principal ventaja de esta técnica está en la diferenciación de las características que definen cada grupo, así como las interacciones que existen entre ellas. Se trata de un modelo apropiado para clasificar buenos y malos pagadores a la hora de recuperar un crédito.

Regresión logística (logit). Los modelos de regresión logística permiten calcular la probabilidad que tiene un cliente para pertenecer a uno de los grupos previamente definidos (buen cliente o mal cliente). Es una técnica de modelado estadístico en el que la probabilidad de un resultado dicotómico se relaciona con un conjunto de posibles variables explicativas de la forma:

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \cdots \qquad \textbf{Ecuación III-2}$$

donde p es la probabilidad del resultado de interés, β_0 es el término de cruce y β_i representa el coeficiente β asociado con la variable independiente correspondiente X_i ($i = 1, \cdots, n$). La clasificación se realiza de acuerdo con el comportamiento de una serie de variables independientes de cada observación o individuo. La principal ventaja del modelo de regresión logística radica en que no es necesario plantear la hipótesis de partida, como por ejemplo la distribución de las variables, mejorando el tratamiento de las variables cualitativas o categóricas. Además este modelo presenta la ventaja de medir la probabilidad de incumplimiento al mantener la variable explicada siempre dentro de un rango de variación entre cero y uno. Wiginton (1980) en (Wiginton, 1980) fue uno de los primeros autores en publicar un modelo de puntuación crediticia aplicando esta metodología. Este autor realizó un estudio comparado entre el análisis discriminante y el modelo Logit en el que determinó que dicho modelo ofrecía un porcentaje de clasificación mejor que el

análisis discriminante. Posteriormente otros investigadores han empleado la regresión logística para diseñar modelos de puntuación crediticia para préstamos personales, empresariales y aplicaciones de tarjetas de crédito.

Modelos de Probabilidad Lineal. Los modelos de probabilidad lineal utilizan un enfoque de regresión por cuadrados mínimos, donde la variable dependiente (variable dummy) toma el valor de uno (1) si un cliente es fallido, o el valor de cero (0) si el cliente cumple con su obligación de pago. La ecuación de regresión es una función lineal de las variables explicativas. Orgler (1970) en (Yair E. Orgler, 1970) fue el precursor de esta técnica usando el análisis de regresión en un modelo para préstamos comerciales. Este mismo autor recurrió a dicha técnica para construir un modelo de puntuación crediticia para préstamos al consumo (Y.E. Orgler & Sciences, 1971), destacando el alto poder predictivo de las variables sobre el comportamiento del cliente, clasificadas fundamentalmente en cuatro grandes grupos: liquidez, rentabilidad, apalancamiento y actividad.

Modelos de Programación Lineal. Método encuadrado dentro de los modelos no paramétricos de puntuación crediticia. En general, este tipo de modelos presentan mayor validez cuando se desconoce la forma que pueda mantener la relación funcional entre las variables. Los modelos de programación lineal permiten programar plantillas o sistemas de asignación de rating sin perder de vista el criterio de optimización de clientes correctamente clasificados. En la década de los 80's del siglo pasado algunos investigadores (Hand, 1981; Kolesar & Showers, 1985; Showers & Chakrin, 1981), sentaron las bases de aplicabilidad de esta técnica en la actividad bancaria; a partir de ellos, otros autores han desarrollado esta metodología para predecir la omisión de pago de créditos.

Árboles de Decisión. La principal ventaja de esta metodología es que no está sujeta a supuestos estadísticos referentes a distribuciones o formas funcionales. Aunque conllevan una comprensión interna difícil sobre su funcionamiento, presentan relaciones visuales entre las variables, los grupos de la variable respuesta y el riesgo; por ello, este método es muy usado en la puntuación crediticia. Los algoritmos más comunes para construir los árboles de decisión son el ID3, C4.5 y C5. En cada uno de ellos se persigue la separación

óptima en la muestra, de tal modo que los grupos de la variable respuesta ofrecen distintos perfiles de riesgo.

La aportación de (Leo. Breiman, Friedman, Stone, & Olshen, 1984) fue determinante para el desarrollo de otros trabajos utilizando esta técnica. Algunos de ellos, (Carter & Catlett, 1987; Coffman, 1986; Makowski, 1985) aplicaron modelos de árboles de decisión para la clasificación de clientes en términos de puntuación crediticia. Adicionalmente se realizó un estudio comparado de esta metodología con el análisis discriminante, confrontando así una técnica paramétrica ante otra no paramétrica (Boyle, Crook, Hamilton, & Thomas, 1992).

Redes Neuronales. Es una metodología catalogada dentro de las técnicas no paramétricas de puntuación crediticia. Las redes neuronales artificiales tratan de imitar al sistema nervioso, de modo que construyen sistemas con cierto grado de inteligencia. La red está formada por una serie de procesadores simples, denominados nodos, que se encuentran interconectados entre sí. Como nodos de entrada se consideran las características o variables de la operación de crédito. El nodo de salida sería la variable respuesta definida como la probabilidad de no pago. La finalidad de cada nodo consiste en dar respuesta a una determinada señal de entrada. El proceso de puntuación crediticia mediante el uso de esta técnica resulta complicado, pues el proceso interno de aprendizaje funciona como una "caja negra" (capa oculta), donde la comprensión de lo que ocurre dentro requiere de conocimientos especializados.

Existen trabajos comparativos de esta técnica con otras técnicas alternativas de clasificación de clientes (Davis, Edelman, & Gammerman, 1992). Con posterioridad, se realizaron trabajos describiendo algunas de las aplicaciones de las redes neuronales empleadas en las decisiones gerenciales sobre el crédito y sobre la detección del fraude (Ripley, 1994; Rosenberg & Gleit, 1994). Desde entonces, gracias al avance en nuevas tecnologías, se han diseñado sistemas avanzados para el objetivo de la clasificación de 'buenos' y 'malos' clientes potenciales.

Redes Neuronales de Propagación hacia atrás. Las redes neuronales de propagación hacia atrás han sido extremadamente populares por su capacidad única de aprendizaje (Widrow, Rumelhart, & Lehr, 1994), y han demostrado un buen desempeño en diferentes

aplicaciones en diversas áreas de investigación como las aplicaciones médicas (Tolle, Chen, & Chow, 2000) y juegos (H. Chen, et al., 1994). Una red neuronal típica de esta forma consiste de una estructura de tres capas: los nodos de la capa de entrada, los nodos de la capa de salida y los nodos de la capa oculta.

Las redes de propagación hacia atrás están totalmente conectadas, en capas, con los modelos de seguimiento hacia adelante (feed –forward). El flujo de activación va de la capa de entrada a través de la capa oculta, hacia la capa de salida. La salidas típicamente tienen salidas iníciales de un conjunto de pesos aleatorios y se van ajustando con cada par de entradas y salidas. Cada par se procesa en dos etapas, una pasada hacia adelante y otra hacia atrás. El paso hacia adelante requiere de un ejemplo de entrada a la red y permitir el flujo de activaciones hasta que alcancen la capa de salida. Durante el paso hacia atrás, la salida actual de la red se compara con el objetivo de salida y se calcula un error de salida para las unidades de salida. Los pesos asignados a las unidades de salida se ajustan para reducir el error (un método de gradiente descendente). El error estimado de las unidades de salida se emplea para derivar una estimación de error para las unidades en la capa oculta. Finalmente, los errores se propagan hacia atrás a las conexiones base de las unidades de entrada. Las redes de propagación hacia atrás actualizan sus pesos incrementalmente hasta que la red se estabiliza.

Máquinas de soporte Vectorial. Las máquinas de soporte vectorial (SVM) son un método de aprendizaje automático introducido por Vapnik (Vapnik, 1995). Se basan en el principio de la minimización estructural del riesgo de la teoría de aprendizaje computacional. Se ha posicionado a los algoritmos SVM en la intersección del aprendizaje automático y la práctica (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998), lo que los ha convertido en candidatos viables para combinar las fortalezas de las teorías más impulsadas con la facilidad de analizar los métodos estadísticos convencionales y más manejo de datos, con una distribución gratuita y métodos robustos de aprendizaje automático.

Las propuestas de SVM se han aplicado en varias aplicaciones financieras, principalmente en el área de la predicción de series de tiempo y la clasificación (Tay & Cao, 2002; T. Van Gestel, et al., 2001).

Modelos Híbridos. Al igual que en muchos aspectos de la inteligencia artificial en la puntuación crediticia una tendencia muy vigente es el de hibridación. La razón para ello es que un gran número de algoritmos han sido reportados que no siguen de forma integral los conceptos de una metaheurística clásica simple (Lozano & García-Martínez, 2010), para resolver este inconveniente se busca lo mejor de una combinación de metaheurísticas (y cualquier otra clase de métodos de optimización) que se desempeñen juntas para complementarse y producir una sinergia provechosa, a lo cual se le denomina hibridación (Raidl, 2006; Talbi, 2002).

Algunas de las posibles razones que justifican la hibridación son (Grosan & Abraham, 2007; Sinha & Goldberg, 2003): 1.- mejorar el desempeño de los algoritmos evolutivos, 2.- mejorar la calidad de soluciones obtenidas por los algoritmos evolutivos y 3.- incorporar los algoritmos evolutivos como parte de un sistema mayor.

Las diferentes instancias de hibridación de metaheurísticas con algoritmos evolutivos se pueden agrupar en diferentes categorías. Dos grupos se derivan de la taxonomía ya bien desarrollada para hibridación de metaheurísticas (Raidl, 2006), la cual se desarrolló sobre las bases de sus estrategias de control:

- Metaheurísticas híbridas colaborativas: Se basan en el intercambio de información entre diferentes metaheurísticas (y posiblemente otras técnicas de optimización) corriendo secuencialmente o en paralelo.
- Metaheurísticas híbridas integradoras: En este caso, un algoritmo se considera un componente subordinado, acoplado de otro algoritmo.

Existen una gran variedad de modelos para poder implementar la minería de datos, los mencionados anteriormente son lo que la literatura muestra que más frecuentemente se emplean en el pronóstico de puntuación crediticia.

III.4 Análisis de la literatura en modelos de puntuación crediticia.

La importancia en la valoración del riesgo crediticio ha provocado un creciente incremento en la investigación correspondiente al tema. Partiendo de una gran cantidad de modelos estadísticos y técnicas de optimización. Tales como el análisis de discriminante lineal ((Fisher, 1936), el análisis de regresión logística (logit) (Wiginton, 1980), el análisis probit

(Grablowsky & Talley, 1981), programación lineal (F. Glover, 1990), programación entera (Mangasarian, 1965), k-vecino más cercano (Henley & Hand, 1996) y árboles de clasificación (Makowski, 1985) son muy empleados en la valoración de riesgo crediticio y tarea de modelado. Aunque estos métodos se pueden utilizar para evaluar el riesgo crediticio, la capacidad de discriminar a los clientes buenos de los malos es un aspecto que aún puede mejorarse. Estudios recientes han revelado que técnicas de inteligencia artificial (artificial inteligence (AI)), tales como las redes neuronales artificiales (artificial neural networks (ANN))(Lai, Yu, Wang, & Zhou, 2006; Malhotra & Malhotra, 2003; Smalz & Conrad, 1994), los algoritmos genéticos (genetic algorhitms (GA)) (M.-C. Chen & Huang, 2003; Varetto, 1998) y las máquinas de soporte vectorial (support vector machine (SVM)) (Huang, Chen, Hsu, Chen, & Wu, 2004; Tony Van Gestel, Baesens, Garcia, & Van Dijcke, 2003) tienen ventaja sobre los modelos estadísticos y las técnicas de optimización para la evaluación de riesgo crediticio, principalmente en las situaciones donde las variables dependientes e independientes empleadas en los modelos exhiben relaciones no lineales complejas.

Pese a que la mayoría de los métodos de clasificación se pueden emplear para evaluar el riesgo crediticio, para los modelos individuales mencionados anteriormente (M.-C. Chen & Huang, 2003; Fisher, 1936; F. Glover, 1990; Grablowsky & Talley, 1981; Henley & Hand, 1996; Huang, et al., 2004; Makowski, 1985; Malhotra & Malhotra, 2003; Mangasarian, 1965; Smalz & Conrad, 1994; Tony Van Gestel, et al., 2003; Varetto, 1998; Wiginton, 1980), resulta complejo indicar que el comportamiento de uno de ellos es consistentemente mejor que los otros en todas las circunstancias. En la mayoría de los casos, el rendimiento de estos modelos individuales es dependiente del problema. El trabajo realizado por algunos investigadores concluyen que generalmente los clasificadores combinados, aquellos que integran dos o más métodos simples de clasificación, han demostrado una exactitud mayor de predicción que los métodos individuales, dicha combinación se conoce comúnmente como “hibridación” (Ahmad Ghodselahi, 2011). La investigación en esta área de clasificadores combinados desde hace algunos años se ha incrementado en la evaluación de riesgo crediticio. Ejemplos recientes son las técnicas de discriminante neuronales (Adnan, 2010; T.-S. Lee, et al., 2002; C.-F. Tsai & Chen, 2010; Yu, Wang, & Lai, 2008),

neuro-difusos (Malhotra & Malhotra, 2002; Piramuthu, 1999) y SVM en diversas instancias (W. Chen, Ma, & Ma, 2009; Y. Wang, Wang, & Lai, 2005; D. Zhang, Hifi, Chen, & Ye, 2008; L.-l. Zhang, Hui, & Wang, 2009). Sin embargo las últimas tendencias motivadas por los esquemas híbridos es la integración de múltiples clasificadores en una salida ensamblada, el aprendizaje conjunto, ha resultado ser un método más eficaz para lograr una clasificación de alto rendimiento. Existe un creciente interés en demostrar que la aplicación de un solo clasificador se puede mejorar con los métodos de ensamblado (ensemble methods)(Ahmad Ghodselahi, 2011; Seni & Elder, 2010). Algunos de los trabajos recientemente han demostrado la efectividad de los métodos de ensamblado (C.-F. Tsai & Wu, 2008; Yu, Wang, & Lai, 2008; Yu, et al., 2010), en donde las redes neuronales y las máquinas de soporte vectorial son nuevamente muy empleadas. Nanni y Lumini (Nanni & Lumini, 2009) compararon el desempeño de clasificadores ensamblados contra sencillos. El resultado demostró que la aplicación de métodos ensamblados conduce a un mejor desempeño de clasificación. De igual forma se empieza a trabajar en modelos ensamblados híbridos (Hsieh & Hung, 2010).

En los años recientes las máquinas de soporte vectorial se han aplicado ampliamente, sin embargo las propuestas híbridas y ensambladas se están empleando con mayor frecuencia debido a que disfrutan de las ventajas de dos o más modelos para realizar los cálculos correspondientes de puntuación crediticia, pero en general son recomendables los modelos que tienen la capacidad de estimar la probabilidad de impago y que además son simples de interpretar y entender (Keramati & Yousefi, 2011).

La tabla III.1 muestra diferentes propuestas desarrolladas para puntuación crediticia empleando diversos modelos de minería de datos.

Tabla III-1 Propuestas de puntuación crediticia en minería de datos.

		Referencias
TECNICAS	Redes Neuronales	(Abdou, Pointon, & El-Masry, 2008), (Angelini, di Tollo, & Roli, 2008), (Yang, Wu, Fu, & Luo, 2008), (Yu, Wang, & Lai, 2008), (Derelioğlu, Gürgen, & Okay, 2009), (M.-C. Tsai, Lin, Cheng, & Lin, 2009), (Šušteršič, Mramor, & Zupan, 2009), (Zhou & Lai, 2009), (Adnan, 2010), (Adnan, 2011), (Nwulu, Oroja, & Ilkan, 2011), (Tarsauliya, Kala, Tiwari, & Shukla, 2011), (Wong & Versace, 2012)
	Clasificadores Bayesianos	(Jiang, 2009), (Baesens, Egmont-Petersen, Castelo, & Vanthienen, 2002), (X.-s. Li & Guo, 2006), (Antonakis & Sfakianakis, 2009), (Panigrahi, Kundu, Sural, & Majumdar, 2009), (Yi & Li Hua, 2009), (J. L. Zhang & Härdle, 2010), (Luo, Xiong, & Zhou, 2011)
	Análisis Discriminante	(Ionescu, Murgoci, Gheorghe, & Ionescu, 2009), (Stefan & Svetlozar T, 2009)
	Regresión Logística	(Schwarz & Arminger, 2005), (Sohn & Kim, 2007), (Dong, Lai, & Yen, 2010), (Yap, Ong, & Husain, 2011)
	K-Vecinos Cercanos	(F.-C. Li, 2009), (H.-L. Chen, et al., 2011)
	Árboles de Decisión	(Pang & Gong, 2009), (Sudjianto, et al., 2010), (D. Zhang, Zhou, Leung, & Zheng, 2010), (Zurada, 2010), (Siami, Gholamian, Basiri, & Fathian, 2011), (Wei-Li, 2011), (Olson, Delen, & Meng, 2012)
	Sistemas basados en reglas Difusas	(Yu, Wang, Lai, & Zhou, 2008b), (Xikun & Zhengzheng, 2010), (Hájek), (Capotorti & Barbanera)
	Máquinas de Soporte Vectorial	(Stecking & Schebesch, 2007), (Martens, Huysmans, Setiono, Vanthienen, & Baesens, 2008), (Schebesch & Stecking, 2008), (Yu, Wang, Lai, & Zhou, 2008a), (Bellotti & Crook, 2009), (L. Zhang & Hui, 2009), (Xu, Zhou, & Wang, 2009), (Zhou, Lai, & Yu, 2009), (Maldonado & Paredes, 2010), (Hájek & Olej, 2011)
	Modelos Híbridos	(Yu, Lai, Wang, & Zhou, 2007), (C.-F. Tsai & Wu, 2008), (Yao, Wu, & Yao, 2009), (C.-F. Tsai & Chen, 2010) (Verikas, Kalsyte, Bacauskiene, & Gelzinis, 2010), (Bahrammirzaee, Ghatari, Ahmadi, & Madani, 2011), (Farquad, Ravi, Sriramjee, & Praveen, 2011), (Fu & Liu, 2011), (Ahmad Ghodselahi, 2011), (Sreekantha & Kulkarni, 2010), (G. Wang & Ma, 2012)

III.4.1 Revisión literaria de puntuación crediticia en micro financieras.

El riesgo de crédito en las entidades de microfinanzas se manifiesta de la misma forma que en el ámbito bancario. Desde sus orígenes, la actividad microfinanciera ha requerido sistemas de gestión adecuados para minimizar los costes. Las limitaciones e inconvenientes en la elaboración de sistemas de calificación estadística del cliente potencial plantean dificultades a la hora de construirlos, hecho que se refleja en la escasa literatura existente

hasta la fecha sobre modelos de puntuación crediticia para las Instituciones Micro Financieras. Hay autores que discuten sobre la conveniencia o no y sobre la posibilidad de éxito de los modelos de puntuación crediticia para las microfinanzas (Dennis, 1995), (Kulkosky, 1996) y (Schreiner, 2003), quienes aportan las limitaciones, ventajas e inconvenientes de los modelos de evaluación del riesgo de crédito en las microfinanzas. Por su parte (Schreiner, 2004), afirma que los modelos planteados por (Sharma & Zeller, 1996), (Reinke, 1998) y (Zeller, 1998) no son estadísticamente válidos, al tiempo que indica que los modelos de (Sharma & Zeller, 1996) y de (Zeller, 1998) no son viables por estar construidos sobre grupos mancomunados, argumentando que la puntuación crediticia no tiene validez para préstamos en grupo.

Los modelos de puntuación crediticia en microfinanzas publicados hasta la actualidad generalmente están diseñados en las regiones de América Latina y del Sur de África, como lo evidencian diversas investigaciones (Dinh & Kleimeier, 2007; Vogelgesang, 2003). El primer modelo de puntuación financiera para microfinanzas fue desarrollado para una institución de microfinanzas de Burkina Faso (Viganò, 1993). Sobre una muestra de 100 microcréditos, y contando con 53 variables iníciales, Viganò utilizó el análisis discriminante para la elaboración del modelo. Como consecuencia del reducido tamaño muestral, el autor tuvo que reagrupar las 53 variables en 13 factores, aunque ello complica la identificación de las características explicativas del no pago del microcrédito.

Un estudio similar se realizó para una IMF de Bangladesh, donde contaron con 868 créditos para el análisis (Sharma & Zeller, 1996). Tras aplicar una metodología Tobit, basada en estimación por máximo-verosimilitud, los autores obtuvieron 5 variables significativas de las 18 que inicialmente contaban con información. De igual forma se utilizó un modelo Probit para la elaboración y construcción de una puntuación crediticia para una entidad de microcrédito de Sudáfrica, en el que aceptó las 8 variables explicativas disponibles para una muestra de 1,641 microcréditos (Reinke, 1998). Otro modelo estadístico de clasificación del cliente para una institución de microfinanzas de Madagascar, también implementado con metodología Tobit (Zeller, 1998). El autor disponía de una muestra de 168 observaciones, incorporando 7 de las 18 variables que tenía por crédito.

Con una muestra de 39,956 microcréditos, se desarrolló un modelo en el que se empleó la regresión logística binaria en clientes de Bancosol (Bolivia), y en el que se incluían las nueve variables independientes disponibles (Schreiner, 1999). Dichas variables fueron resumidas en 1) experiencia como prestatario; 2) historial de morosidad; 3) género; 4) sector de actividad; 5) cantidad desembolsada; 6) garantías; 7) sucursales; 8) oficiales de crédito; y, 9) la fecha del desembolso. También para Bolivia, (Vogelgesang, 2003) formuló dos aplicaciones estadísticas para dos entidades de 8.002 y 5.956 casos respectivamente, mediante un modelo de utilidad aleatoria bajo los supuestos de Greene (Greene, 1992). En la región de Latinoamérica, se trabajó con una puntuación crediticia de Pymes de México, Colombia y Brasil respectivamente (Miller & Rojas, 2004), mientras que en Nicaragua se realizó lo mismo para microfinancieras (Milena, Miller, & Simbaqueba, 2005) . En Mali, se volvió a emplear la regresión logística para una muestra de 269 créditos de una entidad microbancaria del país (Diallo, 2006). Diallo solo obtuvo 5 variables significativas en su modelo.

Una aplicación de puntuación crediticia para la banca minorista de Vietnam mediante el uso de la regresión logística binaria (Dinh & Kleimeier, 2007). La muestra para el modelo estaba conformada por 56.037 créditos de todo tipo (microempresas, consumo, hipotecarios, personales). Obtuvieron 17 variables significativas de 22 variables explicativas, aplicando una combinación adecuada de los conceptos de sensibilidad y especificidad sobre el porcentaje de aciertos.

La tabla III.2 presenta una muestra de los modelos de puntuación crediticia publicados, la mayoría de los cuales se han elaborados para países en vías de desarrollo, en donde el tamaño de muestra es el total de observaciones empleadas, combinando los conjuntos de entrenamiento y prueba. Donde número de entradas es el número de variables disponibles en la información y entre paréntesis se indican las variables empleadas en el modelo final. Las siglas PCC, Se y SPEF corresponden a los términos Probabilidad de Clasificación Correcta, Sensibilidad y Especificidad respectivamente, los dos últimos por lo general se aplican mediante ROC, o bien, el esquema de AUC. En la tabla se observa que son pocas las aplicaciones desarrolladas para puntuación crediticia en las Instituciones

MicroFinancieras, además de que la mayoría de ellos presentan el uso de modelos estadísticos y solo uno de ellos emplea técnicas de inteligencia artificial, aunado a que en la vida cotidiana de las financieras los métodos de evaluación crediticia más comúnmente empleados son los de juicio (Lara Rubio, 2010).

Tabla III-2 Modelos de puntuación crediticia estadísticos o de inteligencia artificial implementados en IMFs.

Modelos de Puntuación Crediticia en Microfinancieras Recientemente Publicados					
(Autor, Fecha) País	*Tipo de Institución*	*Tamaño de la muestra*	*Número de Entradas*	*Técnica(s)*	*Métricas de Desempeño*
(Diallo, 2006), Mali	Microfinanciera	269	17(5)	Regresión Logística Análisis de Discriminante	PCC, R^2
(Dinh & Kleimeier, 2007)	Banca Minorista	56037	22(17)	Regresión Logística	PCC, Se, SPEF.
(W. Jia, Vadera, Dayson, Burridge, & Clough, 2010), India	Microfinanciera	3657	23(8)	Modelo Basado en Ejemplo, Agrupamiento, Redes Bayesinas, Arboles de decisión.	Se, SPEF
(Rayo Canton, et al., 2010) Perú	Microfinanciera	5678	35(9)	Regresión Logística	ROC, Se, SPEF.
(Karlan & Zinman, 2011) Filipinas	Microfinanciera	1601	12(12)	Modelo de utilidad aleatoria	PCC, Pseudo R^2
(Kinda & Achonu, 2012) Senegal	Microfinanciera	759	18(8)	Modelo de Regresión Logística	PCC

Los modelos de puntuación crediticia en las instituciones microfinancieras tienen el mismo fin último que los modelos en otros dominios financieros y crediticios: una discriminación adecuada entre los buenos y los malos prestatarios. Sin embargo, no todas las buenas prácticas de los otros dominios se pueden adaptar a los esquemas de microfinanzas quedando un área de desarrollo e investigación muy amplio por explorar.

En los siguientes capítulos se presenta el desarrollo de la propuesta de puntuación crediticia de este trabajo empleando una combinación de diversos métodos a partir de su hibridación con PSO.

IV

Capítulo IV Optimización por Acumulación de Partículas y la Solución Propuesta

IV. 1 Computación Natural y Optimización por Acumulación de Partículas

Los paradigmas de computación convencional a menudo presentan dificultad al trabajar con problemas cotidianos, como son los caracterizados por ruido, por datos incompletos o multidimensionales, debido a su construcción inflexible. Los sistemas naturales han evolucionado por millones de años para resolver tales problemas, los cuales frecuentemente resuelven con la interacción de varios elementos simples produciendo nuevos comportamientos complejos, que han servido como punto de partida para varios paradigmas de cómputo bioinspirado.

En términos generales las propuestas de computación natural se pueden catalogar en tres clasificaciones distintas (F. E. Glover & Kochenberger, 2003): i) los modelos de inteligencia humana, como las redes neuronales artificiales (Hebb, 2002; McCulloch & Pitts, 1943) (ANN por sus siglas en inglés, Artificial Neural Networks) y los sistemas basados en conocimiento (Feigenbaum, Buchanan, & Lederberg, 1970) (KBS por sus siglas en inglés Knowledge Bases Systems), ii) las poblaciones competitivas, por ejemplo los algoritmos evolutivos (Friedberg, 1958) (EAs por sus siglas en inglés, Evolutionary Algorithms) que a su vez se pueden subdividir en algoritmos genéticos (Goldberg, 1989; Holland, 1962, 1992) (GAs por sus siglas en inglés, Genetic Algorithms), estrategias evolutivas (Rechenberg, 1965) (ESs por sus siglas en inglés Evolution Strategies), programación evolutiva (Fogel, Owens, & Walsh, 1966) (EP por sus siglas en inglés Evolutionary Programming), más recientemente programación genética (Koza, 1992) (GP por sus siglas en inglés Genetic Programming) y evolución diferencial (Storn & Price, 1997) (DE por sus siglas en inglés Differential Evolution); y finalmente iii) las poblaciones

cooperativas, como los sistemas de colonias de hormigas (Colorni, Dorigo, & Maniezzo, 1991; Dorigo & Stützle, 2004) (ACS por sus siglas en inglés Ant Colony Systems) y la optimización por acumulación de partículas (Russell C Eberhart, Kennedy, & Shi, 2001; J. Kennedy & Eberhart, 1995).

La mayoría de esos paradigmas de cómputo natural comparten atributos de alto nivel tales como la habilidad de trabajar datos con ruido o incompletos, la habilidad para resolver problemas combinatorios, detectar propiedades emergentes donde el comportamiento complejo contrasta con toda la simplicidad de los agentes, las normas o interacciones implicadas, el uso de algún tipo de competencia y cooperación, la necesidad de algún tipo de aprendizaje y a menudo son computacionalmente eficientes con la posibilidad de una implementación distribuida (Dréo, Pétrowski, Siarry, & Taillard, 2006). El inconveniente de tales paradigmas es su inherente no-determinismo, y al mismo tiempo que ofrecen soluciones aproximadas a problemas para los cuales una solución analítica o método numérico tradicional no es posible o imposible, presentan desafíos en términos de diseño y validación (Talbi, 2002). La Figura IV.1 muestra una clasificación general de métodos de optimización de un solo objetivo y ubica las metaheurísticas y técnicas de hibridación en este contexto.

IV.2 Metodología de Optimización por Cúmulo de Partículas

PSO está inspirada en el comportamiento social de los enjambres naturales, como sistemas auto organizados y descentralizados; y sus correspondientes conexiones con la computación evolutiva, ha ganado gran aceptación entre los investigadores y ha demostrado un buen desempeño en varios dominios de aplicación, con gran potencial para la hibridación y especialización, además de demostrar algunos comportamientos emergentes de interés (Banks, Vincent, & Anyakoha, 2007). La idea original de la propuesta de cúmulo de partículas es emplear varios agentes autónomos (partículas) que actúan juntos en una forma simple para producir un comportamiento emergente (Reeves, 1983), donde el sistema de partículas de una manera estocástica genera series de puntos en movimiento, los cuales típicamente se inicializan en lugares predefinidos.

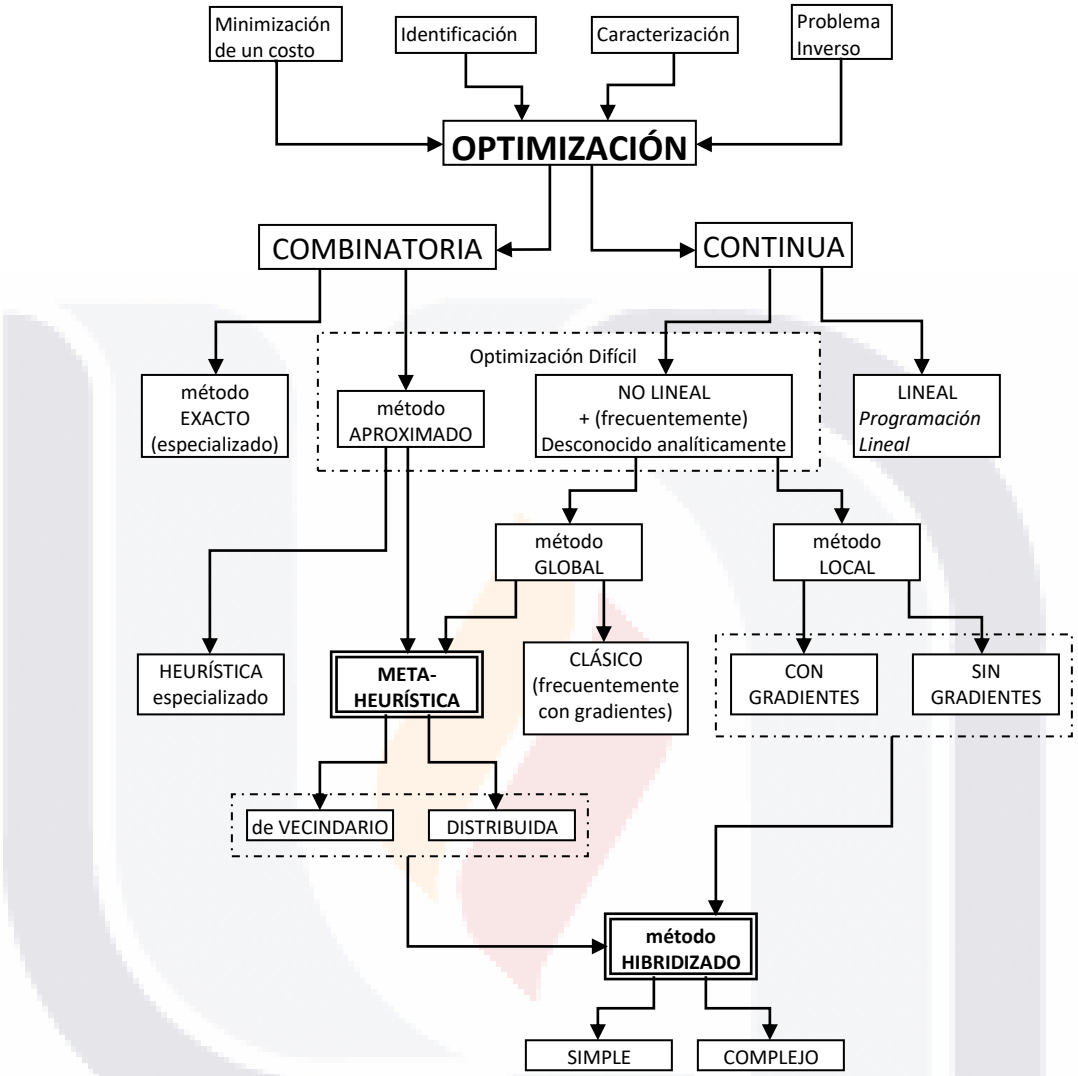


Figura IV-1 Clasificación general de métodos de optimización mono-objetivo. (adaptado (Dréo, et al., 2006)).

A cada partícula se le asigna un vector velocidad inicial el cual de manera iterativa se ajusta por algún factor aleatorio. Entonces cada partícula se mueve de acuerdo a su vector de velocidad y su posición actual que se ajusta a un ángulo restringido para que el movimiento parezca realista. Con la idea de lograr una dinámica superior en el movimiento al de las partículas simples, se representa el comportamiento de grupo (como el de una bandada de pájaros). Empleando el sistema de Reeves como base para un algoritmo de grupo de orden superior (en términos de los objetos que se están modelando), se toma la

partícula en movimiento y se agregan la orientación y la comunicación entre objetos (Reynolds, 1987). Estos comportamientos adicionales permitieron el uso de Boids (acrónimo de Bird-oid objects) individuales para seguir algunas reglas simples de las parvadas: los Boids deben evitar chocar con los Boids cercanos, deben intentar igualar cada vector de velocidad con el de los demás, finalmente deben de tratar de permanecer cercanos. El desarrollo de este modelo subyacente, incrementa la inteligencia de los individuos, elimina el requisito de establecer trayectorias individuales. El mayor nivel de autonomía plantea varias preguntas adicionales, tales como la resolución de conflictos. Reynolds utiliza un enfoque determinista, prioridad de pedido, para resolver estas cuestiones, pero señala que las decisiones podrían ser no deterministas.

Teniendo en mente extender el modelo de Reynolds y reflejar los comportamientos sociales se fija un objetivo más realista: la búsqueda de comida (J. Kennedy & Eberhart, 1995). Este objetivo propicia que los investigadores empleen problemas matemáticos no triviales como función de aptitud para los miembros de la parvada. Así en PSO un número de partículas se localizan en el espacio de búsqueda y se evalúa la función objetivo en su posición actual. Cada partícula determina su movimiento a través del espacio de búsqueda combinando algunos aspectos de la historia de su posición actual y su mejor posición (mejor aptitud), con los de uno o más miembros de la parvada, aunados a algunas perturbaciones aleatorias. La siguiente iteración se desarrolla después de que todas las partículas se han movido. Eventualmente el conjunto total de partículas, como una parvada de pájaros buscando alimento colectivamente, es probable que se mueva cerca de un óptimo de la función de aptitud.

Cada individuo del enjambre de partículas se compone de tres vectores D-dimensionales, donde D es la dimensión del espacio de búsqueda. Estos son la posición actual $\vec{x}_i$, la mejor posición anterior $\vec{p}_{best}$ y la velocidad $\vec{v}_i$.

La posición actual $\vec{x}_i$ se puede considerar un conjunto de coordenadas que describen un punto en el espacio. Donde en cada iteración del algoritmo, la posición actual se evalúa como una solución al problema. Si esa posición es mejor que cualquiera de las encontradas anteriormente, entonces las coordenadas se almacenan en un vector $\vec{p}_i$. El mejor valor

general resultante de la función se almacena en una variable que se puede denominar $\vec{p}_{besti}$ (por la "mejor anterior") para ser comparada con los resultados de las iteraciones posteriores. La finalidad es mantener una búsqueda de las mejores posiciones actualizando $\vec{p}_i$ y $\vec{p}_{besti}$, eligiendo nuevos puntos mediante la adición de coordenadas $\vec{x}_i$ obtenidas a través del ajuste de las velocidades $\vec{v}_i$, lo que efectivamente puede ser visto como un tamaño de paso. PSO es más que una colección de partículas. Una partícula por sí misma no tiene el poder de resolver algún problema, el progreso ocurre solo cuando las partículas interactúan.

Una vez que se ha definido el objetivo a lograr, las variables que resultaron ser redundantes se eliminan del problema, proporcionando un modelo sencillo y más eficiente. El algoritmo resultante para calcular la siguiente posición de la partícula (x) es:

$$v_{t+1} = v_i + \varphi_1\beta_1(p_i - x_i) + \varphi_2\beta_2(p_g - x_i) \qquad \textbf{Ecuación IV-1}$$

$$x_{t+1} = x_t + v_{t+1} \qquad \textbf{Ecuación IV-2}$$

Donde las constantes φ_1 y φ_2 determinan el balance entre la influencia del conocimiento individual (φ_1) y el grupal (φ_2) (ambos se inicializan en 2), β_1 y β_2 son valores aleatorios distribuidos uniformemente por algún límite superior, β_{max}, qué es un parámetro del algoritmo, en conjunto estos parámetros se conocen como coeficientes de aceleración. El comportamiento de PSO cambia radicalmente de acuerdo a los valores que le sean asignados y que se pueden interpretar como componentes de una fuerza de atracción producida por fuentes de rigidez aleatorias, que permiten expresar el movimiento de una partícula como la integración de la segunda ley de Newton. Por lo anterior el valor de 2 para φ_1 y φ_1 permite que la velocidad de las partículas se mantenga bajo control.

Los parámetros p_{besti} y p_g son las mejores posiciones individual y grupal previas; mientras que x_i es la posición actual en la dimensión considerada. Nótese que el signo de los paréntesis resulta en una aceleración de las partículas hacía los mejores puntos previos conocidos en el espacio. Con ello se balancea la aceleración hacia las mejores posiciones

globales y locales. Para un espacio n-dimensional la velocidad de la partícula se calcula para cada dimensión, $i=1, 2,..., n$ y posteriormente se resuelven en un vector final para actualizar la posición de la partícula. Para cuestiones de minimización, p_{besti} al momento $t+1$ se calcula mediante la ecuación IV.3.

$$p_{besti}(t+1) = \begin{cases} p_{besti}(t), si\ f(x_i(t+1)) \geq f(p_{besti}(t)) \\ x_i(t+1), si\ f(x_i(t+1)) \geq f(p_{besti}(t)) \end{cases} \qquad \textbf{Ecuación IV-3}$$

Donde $f : R^D \rightarrow R$ es la función de aptitud, midiendo la distancia entre la solución candidata y la solución óptima.

En el momento t, p_g se calcula empleando la ecuación IV.4.

$$p_g \in \left\{ \{p_{g0}(t), \cdots, p_{gn_s}(t)\} | f(p_g) = min\{p_{g0}(t), \cdots, p_{gn_s}(t)\} \right\} \qquad \textbf{Ecuación IV-4}$$

Donde n_s es el número de partículas en el enjambre. De la ecuación IV.4, la mejor posición global es la posición óptima encontrada hasta el momento en todas las partículas. La tabla IV.1 muestra el algoritmo general de PSO.

Tabla IV-1 Algoritmo PSO Clásico

- Entradas
  - $t = 0$
  - *Nube* ← Inicializar la nube de partículas.

1) Mientras no se alcance la condición de parada hacer
   a. $t = t + 1$
   b. Para i = 1 hasta tamaño (*Nube)* hacer
      i. Evaluar cada partícula x_i de la *Nube*
      ii. Sí *aptitud_* x_i es mejor que *aptitud_mejorpos*$_i$ entonces
         1. *mejorpos*$_i \leftarrow x_i$
         2. *aptitud_mejorpos*$_i \leftarrow$ *aptitud*x_i
      iii. Fin Si
      iv. Sí *aptitud_mejorpos*$_i$ es mejor que *aptitud_mejorpos* entonces
         1. *mejorpos* ← *mejorpos*$_i$
         2. *aptitud_mejorpos* ← *aptitud_mejorpos*$_i$

v. Fin Si

c. Fin Para

d. Para *i* = 1 hasta tamaño (*Nube)* hacer

i. Calcular la velocidad v_i de x_i en base a los valores x_i, $mejorpos_i$ y *mejorpos*

ii. Calcular la nueva posición de x_i, de su valor actual y de v_i

e. Fin Para

2) Fin Mientras

Salida: Devuelve la mejor solución encontrada.

Adicionalmente se debe considerar el tamaño de la población, que a menudo se fija empíricamente basándose en la dimensionalidad y la dificultad percibida del problema, donde valores en el rango de 20 a 50 son comunes (Poli, Kennedy, & Blackwell, 2007), así como su ubicación y velocidad de cada una de las partículas de la población, lo cual se hace de manera aleatoria y se ajustan de manera iterativa, de tal forma que la partícula oscila estocásticamente alrededor de las posiciones p_{besti} y p_g.

IV.2.1 Variaciones en el algoritmo.

A pesar de que la formulación básica de las ecuaciones de PSO ha permanecido sin cambios (Banks, et al., 2007), la idea de reducir la posibilidad de que las partículas salgan del espacio del problema y oscilen sin control, situación que puede influir notablemente en el equilibrio entre la exploración y la explotación y provocar fallos en la convergencia del modelo a conducido proponer mejoras y variantes en el algoritmo canónico de PSO .

Una de las primeras modificaciones consiste en controlar las velocidades dentro de un rango de velocidades límite $[-V_{max}, +V_{max}]$, con V_{max} usualmente ubicado entre 0.1 y 1.0 veces la máxima posición de la partícula (Russell C. Eberhart, Simpson, & Dobbins, 1996), sin embargo el establecimiento de estos límites fijos siguen presentando problemas de estabilidad y convergencia. Intentando lograr un mejor control del alcance de la búsqueda, así como reducir la importancia del parámetro V_{max} después de experimentar con el algoritmo estándar, se tiene una nueva modificación de PSO (Y. Shi & Eberhart, 1998):

$$v_{t+1} = wv_t + \varphi_1\beta_1(p_i - x_i) + \varphi_2\beta_2(p_g - x_i) \qquad \textbf{Ecuación IV-5}$$

En donde se identifica que sin la memoria de velocidad el enjambre podría contraerse a la mejor solución global encontrada dentro de sus límites iníciales (búsqueda local). De manera inversa, con la memoria de velocidad, el enjambre se comporta en el sentido opuesto provocando una búsqueda global. Para ayudar en el balance entre la exploración y la explotación, es que se integra este nuevo parámetro denominado inercia, *w*. Los experimentos iníciales sugirieron que un valor entre 0.8 y 1.2 proveían buenos resultados, aunque en un trabajo posterior (Russel C. Eberhart & Shi, 2000), se estableció que el valor fijado típicamente en 0.9 (reduciendo el movimiento gradual de cada partícula, permitiendo mayor exploración inicial) y reducido linealmente a 0.4 (acelerando la convergencia al óptimo global) produce un mejor desempeño durante la corrida de optimización.

Un método alternativo de implementar PSO, es el de emplear un coeficiente limitante χ (Clerc & Kennedy, 2002), para controlar el comportamiento de las partículas en el enjambre, en lugar de aplicar inercia a la memoria de la velocidad, obteniendo las ecuaciones IV.6 y IV.7.

$$V_{t+1} = \chi\{v_t + \varphi_1\beta_1(p_{besti} - x_i) + \varphi_2\beta_2(p_g - x_i)\} \qquad \textbf{Ecuación IV-6}$$

$$\chi = \frac{2}{\left|2-\varphi-\sqrt{\varphi^2-4\varphi}\right|} \text{ donde } \varphi\text{- } \varphi_1\text{+ } \varphi_2\text{, } \varphi\text{>4} \qquad \textbf{Ecuación IV-7}$$

Sin embargo, se demostró que el modelo es algebraicamente equivalente al de inercia y las mejoras en el desempeño que se podrían alcanzar a través de un amplio rango de problemas (Russel C. Eberhart & Shi, 2000), volviéndose así este el algoritmo canónico actual de PSO.

La versión estándar de PSO considera como factores de influencia dos fuentes genéricas, la propia y la del mejor vecino. La información del resto de los vecinos no se emplea, una variación interesante del algoritmo consiste en considerar como interactúan las partículas con el resto de sus vecinos a la cual se le ha denominado Acumulación de Partículas Totalmente Informadas (FIPS por sus siglas en inglés, Fully Informed Particle Swarm), en donde la partícula es afectada por todos sus vecinos, sin considerar en algunas ocasiones la influencia de los éxitos anteriores (Mendes, Kennedy, & Neves, 2004). FIPS se puede describir a través de las ecuaciones IV.8 y IV.9.

$$V_{t+1} = \chi\left\{v_t + \frac{1}{K_i}\sum_{n=1}^{K_i}\varphi_n\beta_n\,(p_{nbr_n} - x_i)\right\}$$ **Ecuación IV-8**

$$x_{t+1} = x_t + v_{t+1}$$ **Ecuación IV-9**

Donde K_i es el número de vecinos de la partícula i y nbr_n son el i-ésimo vecino. Se puede observar que la formulación es la misma que en la acumulación de partículas tradicional si solo se considera la partícula en estudio y la mejor del vecindario para $K_i=2$. Con mejores parámetros, FIPS aparenta encontrar mejores soluciones en menos iteraciones que el algoritmo canónico, pero es mucho más dependiente de la topología de la población (Poli, et al., 2007).

Un caso particular de PSO es el binario que fue desarrollado por Kennedy (James Kennedy & Eberhart, 1997). En esta versión, la posición de la partícula se codifica como una cadena binaria que imita los cromosomas de un algoritmo genético. La función de velocidad de la partícula se usa como la distribución de probabilidad para la ecuación de posición. Es decir, la posición de la partícula en una dimensión es generada al azar usando dicha distribución. La ecuación IV.10 define como se actualiza la posición de la partícula.

$$aleat < \frac{1}{1+e^{-v_i^{t+1}}}\begin{cases} si\ X_i^{t+1} = 1 \\ no\ X_i^{t+1} = 0 \end{cases}$$ **Ecuación IV-10**

Un bit de valor [1] en cualquier dimensión en la posición del vector indica que el parámetro referenciado se emplea en la siguiente generación, mientras que un valor de [0] indica que este parámetro no se considera en la siguiente generación.

IV.2.2 Especializaciones en el algoritmo.

En general, se han realizado muchos ajustes y personalizaciones del algoritmo básico desde su planteamiento original. Algunos se han traducido en la mejora del rendimiento general, y algunos tienen un mejor rendimiento en determinados problemas. Algunos de los de mayor impacto y que parecen tener un aporte más prometedor para el futuro del paradigma se muestran en la tabla IV.2 a continuación.

Tabla IV-2 Ejemplos de casos de Especialización de PSO.

Especialización	*Descripción*	*Referencia*
Cúmulo de Partículas Binaria (Binary particle swarms)	Consiste en operar con cadenas de bits en lugar de números reales.	(James Kennedy & Eberhart, 1997), (Sadri & Suen, 2006), (Khanesar, Teshnehlab, & Shoorehdeli, 2007), (S. Lee, Soak, Oh, Pedrycz, & Jeon, 2008), (Jun & Chang, 2009), (Yuan, Nie, Su, Wang, & Yuan, 2009), (Muhammad, Selvan, Masra, Ibrahim, & Abidin, 2011)
Problemas dinámicos	Son un reto para PSO. Típicamente se modelan mediante una función que cambia en el tiempo, haciendo que la memoria de las partículas se vuelva obsoleta.	(Carlisle & Dozier, 2000), (Schoeman & Engelbrecht, 2006), (Blackwell, 2007), (Blackwell, Branke, & Li, 2008), (Du & Li, 2008), (L. Liu, Wang, & Yang, 2008), (Cui, Charles, & Potok, 2009), (Hashemi & Meybodi, 2009), (Kamosi, Hashemi, & Meybodi, 2010), (Novoa-Hernández, Corona, & Pelta, 2011)
Funciones con ruido	Las funciones de aptitud con ruido son importantes, debido a que frecuentemente se encuentran en problemas del mundo real. Al evaluar la función el ruido puede provocar que la exploración de PSO más de una vez en la misma posición, entregue valores de aptitud diferentes.	(Parsopoulos & Vrahatis, 2004)
Hibridación	Varios investigadores han intentados adaptar los parámetros de PSO como respuesta a la información del ambiente. Técnicas de computación evolutiva y otros métodos se han empleado también por los investigadores de cúmulo de partículas.	(Angeline, 1998), (Marinakis, Marinaki, Doumpos, & Zopounidis, 2009), (Ahn, An, & Yoo, 2010), (Bengoetxea & Larrañaga, 2010), (Fındık, Babaoğlu, & Ülker, 2010), (Niknam & Amiri, 2010), (Sarkar & Das, 2010), (Gao & Xu, 2011), (D. Jia, Zheng, Qu, & Khan, 2011), (Kumar, Sharma, & Sadu, 2011), (Thangaraj, Pant, Abraham, & Bouvry, 2011), (Yang Shi, Liu, Gao, & Zhang, 2011), (Valdez, Melin, & Castillo, 2011), (Robati, Barani, Nezam Abadi Pour, Fadaee, & Rahimi Pour Anaraki, 2012), (Voglis, Parsopoulos, Papageorgiou,

Especialización	*Descripción*	*Referencia*
		Lagaris, & Vrahatis, 2012)
Problemas combinatorios	Los problemas combinatorios se presentan frecuentemente en entornos de planificación en escenarios como el problema del agente viajero y se han desarrollado varias técnicas de PSO para afrontarlos.	(Rezazadeh, Ghazanfari, Saidi-Mehrabad, & Jafar Sadjadi, 2009), (Consoli, Moreno-Pérez, Darby-Dowman, & Mladenović, 2010), (Fakhfakh, Cooren, Sallem, Loulou, & Siarry, 2010), (Lin, et al., 2010), (Marinakis & Marinaki, 2010), (Moslehi & Mahnam, 2011), (Nimtawat & Nanakorn, 2011), (Qi, 2011), (J. Wang, Cai, Zhou, Wang, & Li, 2011),

IV.2.3 Aplicaciones de PSO

La optimización por acumulación de partículas se puede emplear y se ha usado en una gran diversidad de aplicaciones. Áreas en donde PSO ha demostrado particular desempeño incluyen problemas multimodales y problemas para los cuales no existe un método especializado disponible o en donde todos los métodos especializados entregan resultados no satisfactorios. Utilizando PSO como paradigma se puede subdividir las diversas aplicaciones en dos propuestas principales: la primera explota su capacidad para optimizar de manera eficiente, que a menudo requiere la adaptación de PSO para satisfacer las necesidades específicas del problema y en la segunda el problema se adapta para permitir el uso de PSO. Una tercera propuesta, menos común, se emplea la metáfora social original de aprendizaje individual y grupal para proporcionar una mayor comprensión de cómo los individuos se comportan dentro de los grupos. Las aplicaciones son tan numerosas y diversas que se pudiera escribir un capítulo completo para revisar solamente aquellas que son más paradigmáticas. Poli (Poli, et al., 2007) realiza una clasificación de 26 diferentes categorías en un análisis de más de 1100 publicaciones en la base de datos de IEEE Xplore obteniendo que las principales categorías de aplicación son las siguientes: análisis de imagen y video (7.6% de los artículos consultados); diseño y restructuración de redes de electricidad y despacho de carga (7.1%); aplicaciones de control (7.0%); aplicaciones en electrónica y electrodomésticos (5.8%); diseño de antenas (5.8%); generación de potencia y sistemas de potencia (5.8%); planificación (5.6%); aplicaciones de diseño (4.4%) diseño y optimización de redes de comunicación (4.4%); aplicaciones biológicas, medicas y

farmacéuticas (4.3%); agrupamiento, clasificación y minería de datos (4.3%); procesamiento de señal (3.8%); redes neuronales (3.8%); problemas de optimización combinatoria (3.5%); robótica (3.4%); predicción y pronóstico (2.9%); modelado (2.8%); detección o diagnóstico de fallas y la recuperación de éstas (2.3%); sensores y redes de sensores (1.9%), aplicaciones en gráficas computacionales y visualización (1.7%); diseño u optimización de máquinas y motores eléctricos (1.4%), aplicaciones en metalurgia (1.3%); seguridad y aplicaciones militares (1.3%); finanzas y economía (1.0%).

IV.3 Planteamiento de la propuesta a través de PSO

El problema de puntuación crediticia es como ya se mencionó un problema de clasificación que se está afrontando mediante un enfoque de minería de datos apegándose a la metodología CRISP, la cual se analizó en el capítulo III.

Las diversas técnicas de minería de datos empleadas en este trabajo se ven robustecidas con la aplicación conjunta de PSO, así a continuación se hace una descripción de cómo se implementaron los métodos y materiales en la solución propuesta de la tesis.

IV.3.1 Entendiendo el problema.

Esta propuesta, tiene como idea principal facilitar el proceso de asignación de créditos en instituciones microfinancieras, teniendo presente la obtención de una ganancia a partir del pago del monto acreditado así como de los intereses devengados por el préstamo.

El entendimiento del proceso de asignación de créditos, así como de los factores (datos) que se toman en cuenta para dicho proceso fueron las primeras actividades a realizar. Teniendo en cuenta la dificultad que se presenta en la recolección de los datos inicial en este tipo de problemas de minería de datos, debido a que se trata de una actividad financiera, las instituciones por política no suele facilitar la información relacionada con el giro del negocio, sin embargo una vez superado este inconveniente los pasos a seguidos por la metodología CRISP son verificar la calidad de los datos, descubrir los primeros conocimientos a partir de los datos y/o descubrir subconjuntos interesantes para forma hipótesis en cuanto a la información oculta.

IV.3.2 Preparación y selección de los datos.

La preparación de los datos es un aspecto importante para la minería de datos, debido a que en el mundo real los datos tienden a estar incompletos, con ruido e inconsistentes. La preparación de los datos incluye la limpieza, la integración, la transformación y la reducción de los datos. Por ello es importante la fase de entendimiento y lograr una visión global de los datos. Un resumen descriptivo de los datos suele ser una técnica común para identificar las propiedades típicas de los datos e identificar cuales datos se pueden tratar como anómalos.

Para muchas tareas de preparación es deseable tener el conocimiento del comportamiento de los datos en cuanto a su tendencia central y su dispersión. Medidas de tendencia central incluyen la media, la mediana, la moda y el rango medio, mientras que medidas de dispersión incluyen cuartiles, rango de intercuartiles y la varianza. Estas estadísticas descriptivas son de gran ayuda para entender la distribución de los datos.

La limpieza de datos consiste de rutinas que depuran los datos en cuanto a valores perdidos, aislando datos imprecisos, identificación de anomalías, que generalmente se eliminan, aunque en algunas situaciones resultan de utilidad en el descubrimiento de conocimiento, así como la resolución de inconsistencias. En este trabajo la limpieza de datos se llevó a cabo mediante un análisis detallado de los datos disponibles, adecuando los valores perdidos a la política de promediar los valores de los datos existentes, en una primera instancia y posteriormente con el manejo de inconsistencia y datos imprecisos.

En conjunto con la limpieza de datos la integración es un proceso que se realizó, sobre todo para campos cuyos valores no están bien validados, como suelen ser los nombres de personas y domicilios que son datos que frecuentemente se abrevian o sufren de errores de captura, provocando inconsistencias y redundancias, al momento de integrar las diversas tablas de interés para el proceso de minería de datos.

La transformación de los datos resulta de mucha utilidad debido a que propician su normalización. Así, con el fin de acelerar el proceso de convergencia y reducir la influencia del desequilibrio de los datos en la precisión del modelo, se realizó la transformación de los

datos de la muestra para normalizar el valor de las variables en el rango [0, 1]. El primer paso de este proceso fue clasificar las variables en secuenciales y discretas.

En el grupo de variables discretas se empleó la ecuación IV.11

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \textbf{Ecuación IV-11}$$

Donde, X'$\in [0,1]$ representa la variable pretratada, X_{min} y X_{max} representan el valor máximo y mínimo de la variable X respectivamente.

Las variables en el grupo secuencial, se observa que obedecen a una distribución normal, aproximadamente, esto es $X_i \tilde{} N(\mu,\sigma^2)$, para su normalización se utilizó la ecuación IV.12.

$$X' = \emptyset\left(\frac{X - \mu}{\sigma}\right) \qquad \textbf{Ecuación IV-12}$$

Donde la función $\phi(X)$ representa la probabilidad acumulada de la distribución normal y se expresa mediante la ecuación IV.13.

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \qquad \textbf{Ecuación IV-13}$$

Finalmente la preparación de datos tiene el proceso de reducción de los datos, la selección de un conjunto apropiado de características a menudo explota los criterios de diseño, tales como la minimización de la redundancia y la no correlación. Este proceso consiste en encontrar el conjunto óptimo d del total de variables m del problema en estudio (Tu & Chuang, 2007). Una posibilidad es hacer una búsqueda exhaustiva entre todos los subconjuntos de variables posibles $\binom{m}{d}$ y elegir el mejor de acuerdo con el criterio de optimización considerado. Sin embargo, tal propuesta tiene un alto costo computacional. Se han empleado gran cantidad de propuestas para llevar a cabo este procedimiento, sin embargo en este trabajo se realizó con una adecuación de la propuesta realizada por Voss (Voss, 2005) en el manejo de PCA y PSO.

El análisis de componentes principales es un método de extracción de características eficaz (Hui, et al., 2005), lo que implica un procedimiento matemático que transforma un número de variables correlacionadas posiblemente en un menor número de variables no correlacionadas llamadas componentes principales. Esto se logra mediante la selección de un conjunto de vectores ortonormales u_i (Jackson, 1991) que se emplean para definir las nuevas variables como una combinación lineal de las variables originales. Una matriz de covarianza ponderada se utiliza para calcular los vectores ortonormales base u_i y se define como lo enuncia la ecuación IV.14.

$$s_{jk} = \frac{\sum_{i=i}^{N} W_i(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{n-1} \qquad \textbf{Ecuación IV-14}$$

La Figura IV.2 ilustra la forma que toma la matriz de covarianza a partir de la ecuación IV.14.

$$S = \begin{bmatrix} s_1^2 & s_{12} & \dots s_{1n} \\ s_{12} & s_2^2 & \dots s_{2n} \\ \vdots & \vdots & \vdots \\ s_{1n} & s_{2n} & \dots s_n^2 \end{bmatrix}$$

Figura IV-2 Matriz de covarianzas ponderada necesaria para el cálculo de los vectores ortonormales en el proceso de PCA.

Donde n es el número de dimensiones, N es el número de puntos definidos en la matriz de covarianza y W_i es el peso dado al punto i. W_i se define como una función de la k *iteración/generación*, como lo muestra la ecuación IV.15.

$$W_i = \left(\frac{k}{k_{max}}\right)^{\gamma} \qquad \textbf{Ecuación IV-15}$$

γ permite un control no lineal sobre las ponderaciones aplicadas.

La matriz de covarianza se emplea para calcular los componentes principales con una matriz de vectores propios (eigenvectores) ortonormales U, donde U se define implícitamente como:

$$U'SU = \Lambda \qquad \textbf{Ecuación IV-16}$$

La matriz *U* contiene los vectores propios como columnas:

$$U = [u_1|u_2|\cdots|u_n] \qquad \textbf{Ecuación IV-17}$$

La diagonal principal Λ de la matriz derivada de la ecuación IV.16 representa un arreglo compuesto por los valores propios (eigenvalores). Así, suponiendo que los valores propios se representan con $\lambda_i(i=1,2,\cdots,N)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ mientras que los vectores propios son $U_i(i=1,2,\cdots,N)$, entonces:

$$\Lambda = \begin{bmatrix} \lambda_1 d & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}, U = [U_1 \cdots U_N] \qquad \textbf{Ecuación IV-18}$$

La ecuación IV.16 sirve como referencia para calcular los componentes principales de la muestra, seleccionando los vectores propios que están compuestos por los primeros valores propios $m(m<N)$ más grandes U^m. De esta forma las primeras *m* eigenvariables son:

$$W_{eig} = U^m X \qquad \textbf{Ecuación IV-19}$$

Donde *X* corresponde a la matriz original entrenamiento, así se logra la proyección del conjunto de variables *X*, en el subespacio individual de cada variable. De esta forma, el subespacio decidido por los primeros *m* ejes principales puede reducir los datos iníciales al máximo. La definición del valor más apropiado de *m* para tener un mejor desempeño se logra a través de la tasa de contribución o la tasa de contribución acumulativa mediante la ecuación (IV.20).

$$R = \lambda_j \Big/ \sum_{i=1}^{N} \lambda_i \Rightarrow R = \sum_{j=1}^{m} \lambda_j \Big/ \sum_{i=1}^{N} \lambda_i \qquad \textbf{Ecuación IV-20}$$

Un valor apropiado de *m* genera un valor de $R \in [85\%\text{-}95\%]$.

El modelado de la selección de los datos, así como el resto de modelos propuestos se analizan en la siguiente sección. La tabla IV.3 muestra el algoritmo básico de PCA.

Tabla IV-3 Algoritmo básico de PCA

- Entradas
  - $X \leftarrow$ Matriz de datos muestrales (centrada)

1) $S \leftarrow$ Obtener la matriz de covarianzas (simétrica y mayor igual a 0) a partir de X empleando la ecuación IV.14
2) $U \leftarrow$ Obtener la matriz de vectores propios ortonormales mediante su descomposición espectral de S, representados por columnas según la ecuación IV.17
3) Λ representa la diagonal principal de U un arreglo compuesto por los valores propios (eigenvalores).
4) Suponiendo que los valores propios se representan con $\lambda_i(i=1,2,\cdots,N)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ mientras que los vectores propios son $U_i(i=1,2,\cdots,N)$, se define la ecuación IV.18
5) $W_{eig} \leftarrow$ Nuevas variables (eigenvariables) a partir de los componentes principales empleando ecuación IV.19
6) Se seleccionan las primeras m eigenvariables a partir de la ecuación IV.20 para reducir la dimensión de los datos iniciales, teniendo en cuenta que un valor genera un $R \in [85\%\text{-}95\%]$.

IV.3.3 Modelado.

La fase de modelado se desarrolló en varias instancias, por tal motivo se emplearon diversos modelos implementados con diversas técnicas de minería de datos en forma conjunta e hibrida variantes del algoritmo PSO.

IV.3.3.1 Modelado de selección de variables.

La tarea del modelo de selección de variables es buscar el subconjunto de variables más representativas, en esta propuesta se empleó el algoritmo PSO binario a partir del espacio de variables extraído con PCA. Cada partícula en algoritmo representa un posible solución candidata (subconjunto de variables). La evolución es controlada por una función de aptitud (fitness) que se define en términos de la separación de clases (índice de dispersión) que proporciona una indicación de la aptitud esperada en ensayos futuros.

Las partículas se crearon con una representación cromosómica con un código inicial generado aleatoriamente imitando un cromosoma de un algoritmo genético; cada una de ellas se codificó con una cadena de alfabeto binario $P=F_1\ F_2 \ldots\ F_1,\ n=1,2,\ldots,m$; en donde m es la longitud del vector propio extraído por PCA. Cada gen del cromosoma de longitud m representa la variable seleccionada, *"1"* denota que la variable correspondiente se ha

seleccionado, otro caso denota rechazo. El algoritmo PSO binario se utiliza para búsquedas del gen espacio 2^m para el subconjunto óptimo de variables, donde el óptimo se define con respecto a la separación de clases. Por ejemplo, para un conjunto de datos con dimensión *(n=10)* $P=F_1F_2\ F_3F_4\ F_5F_6\ F_7F_8\ F_9F_{10}$ se analiza utilizando el PSO binario para seleccionar las variables, se puede seleccionar un subconjunto de variables menores que n. Es decir, PSO puede elegir de forma aleatoria 6 variables F_1, F_2, F_4, F_6, F_8, F_9 mediante el establecimiento de los bits 1, 2, 4, 6, 8 y 9 en la partícula cromosoma. Para cada partícula, la eficacia del subconjunto de variables seleccionadas para retener la máxima precisión en la representación del conjunto original se evalua con base al valor de aptitud.

La función de aptitud considera los m-genes de la partícula para representar los parámetros de PSO para evolucionar iterativamente. En cada generación, cada partícula (o individuo) se evalúa, y el valor de *bondad* o *aptitud* se regresa mediante una función de aptitud. Esta evolución es conducida por la función de aptitud *F* que evalúa la calidad de las partículas evolucionadas en términos de su habilidad para maximizar el término de separación de clases indicado por el índice de dispersión entre las diferentes clases (C. Liu & Wechsler, 2000).

Los valores $w_1, w_2, \ldots, w_L$ y $N_1, N_2, \ldots, N_L$ se emplearon para denotar las clases y el número de casos para cada clase respectivamente. Sé supuso que $M_1, M_2, \ldots, M_L$ y M_0 son las medias de las clases correspondientes y la media general en el espacio de variables, M_i se puede calcular mediante la ecuación IV.21.

$$M_i = \frac{1}{N_i}\sum_{J=1}^{Ni} W_{j^{(i)}}, \quad i = 1,2,\ldots.,L \qquad \textbf{Ecuación IV-21}$$

En donde $W_j^{(i)}, \quad j=1,2,\ldots,N_i$, representa los casos muestra para w_i y la media general se expresa mediante la ecuación IV.22.

$$M_0 = \frac{1}{N}\sum_{i=1}^{L} N_i M_i \qquad \textbf{Ecuación IV-22}$$

Donde *n* es el número total de casos para todas las clases. Así, el índice de la función de aptitud entre clases F esta dado por la ecuación (IV.23).

$$F = \sqrt{\sum_{i=1}^{L} (M_i - M_0)^t (M_i - M_0)} \qquad \textbf{Ecuación IV-23}$$

Una muestra de la idea del algoritmo de selección de variables empleando PSO a partir de PCA se muestra en la figura IV.3.

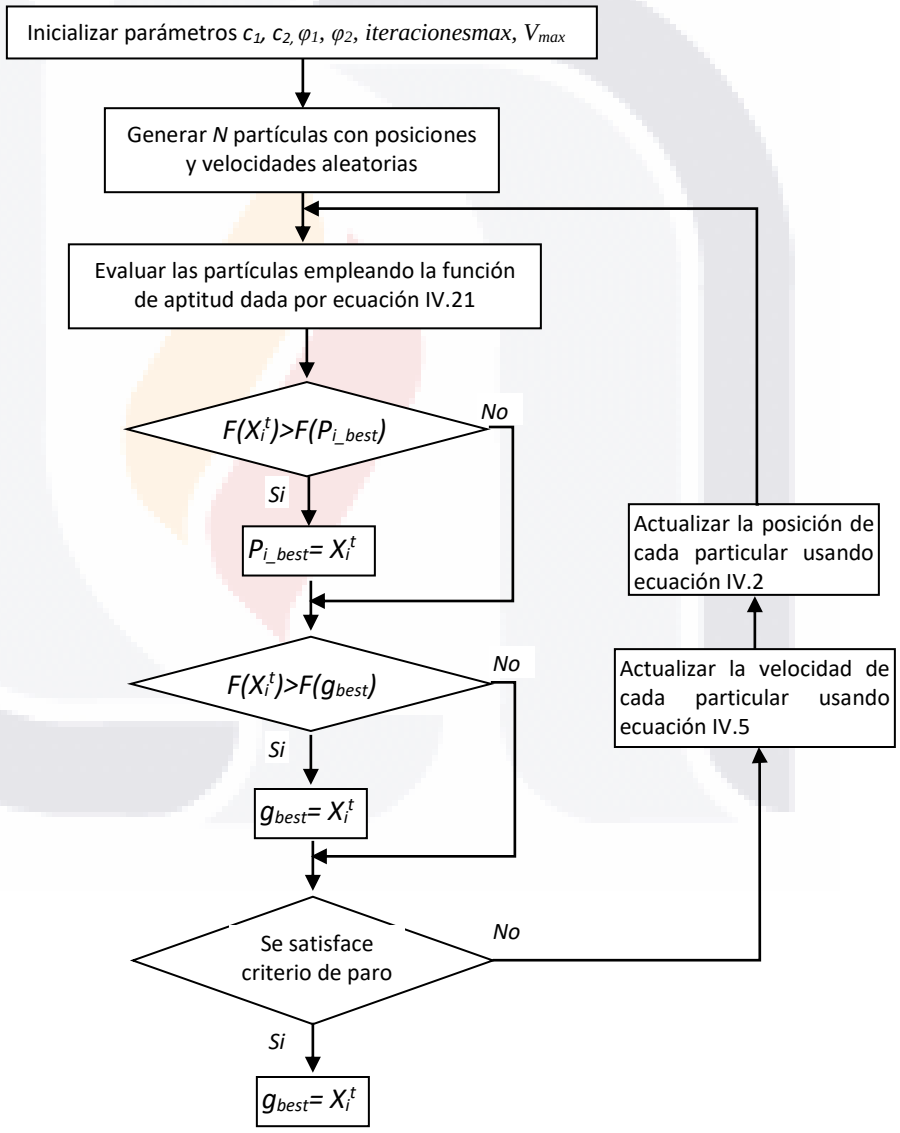


Figura IV-3 Algoritmo de selección de variables basado en PSO a partir de los resultados de PCA.

IV.3.3.2 Modelado de puntuación crediticia.

En el modelo de puntuación crediticia se recurrió al paradigma de aprendizaje ensamblado (ensemble learning) el cual es un paradigma de aprendizaje de máquina (machine learning) donde múltiples aprendices son entrenados para resolver el problema. En contraste con las propuestas típicas de aprendizaje de máquina que tratan de aprender una hipótesis a partir de datos de entrenamiento, lo métodos de ensamblaje tratan de construir un conjunto de hipótesis y combinarlos para su uso (Nanni & Lumini, 2009). Esta metodología se utiliza para mejorar el desempeño y la confiabilidad de la tarea de clasificación. Los sistemas de clasificadores múltiples se basan en la agrupación de un conjunto de dichos clasificadores cuya fusión logra un mejor rendimiento que un solo clasificador. La idea fundamental de la mayoría de las propuestas para construir clasificadores por ensamble consiste en ajustar los datos de entrenamiento en *n* conjuntos, uno por cada clasificador y combinar los resultados individuales mediante una regla de decisión final (G. Wang, et al., 2011). El racionamiento parte de la idea de que puede resultar más difícil optimizar el diseño de un clasificador complejo único que optimizar la combinación de varios clasificadores relativamente simples. Adicionalmente en los modelos de ensamblaje el error y la desviación de un clasificador se compensan por los otros miembros del ensamble en la tarea de clasificación (Ahmad Ghodselahi & Amirmadhi, 2011).

La capacidad de generalización de un ensamblaje es usualmente mucho más fuerte que el de un aprendiz simple, lo cual hace los métodos de ensamble muy atractivos. En la práctica, para lograr un buen ensamble, se deben satisfacer dos condiciones: precisión y diversidad. Para la primera condición, se considera simplemente que la base de aprendizaje debe ser más exacta que adivinar al azar. En esta propuesta, se aplicó PSO como base de aprendizaje para cumplir la condición anterior. Para la segunda condición, se quiere decir que cada aprendiz tiene su base de conocimientos sobre el problema y tiene un patrón diferente de errores en comparación con otros aprendices. Existen diferentes métodos para la construcción de diversas bases de aprendizaje, el método que se implantó en esta propuesta es la versión de empaquetamiento denominada Bagging (Bootstrap Aggregating) propuesto por Breiman (Leo Breiman, 2001). La tabla IV.4 muestra el algoritmo de funcionamiento para Bagging.

Tabla IV-4 Algoritmo simple de Bagging.

1) Sea *M* el número de predictores requeridos.
2) $d = \{(x_1,y_1), \ldots, (x_1,y_1)\}$
3) Para i = 1 hasta *M repetir*
    a. Generar una nueva muestra d_{bag}, eligiendo *N* muestras desde *d* con reemplazo.
    b. Entrenar un estimador f_i con la muestra d_{bag} y agregarlo al ensamble.
4) Fin Para
5) El resultado final del ensamble esta dado por la ecuación IV.24

La base del modelo ensamblado, se construyó a partir de los modelos estadísticos simples de regresión logística y regresión lineal múltiple. La construcción del modelo de análisis de discriminante se obtuvo recurriendo a la ecuación III.1, mientras que el modelo de regresión logística se consiguió utilizando la ecuación III.2. Así el principio del modelo de pronóstico ensamblado se puede describir con la ecuación IV.24.

$$f_t = \sum_{i=1}^{m} w_i f_{it}$$
$$s.t. \quad \sum_{i=1}^{m} w_i = 1 \quad t = 1,2,\cdots,n$$

Ecuación IV-24

Donde f_{it} $(i=1,\ldots, m, t=1,\ldots,n)$ representa el valor pronosticado del *i-ésimo* modelo simple en el *t-ésimo* ejemplo; w_i representa el peso del *i-esimo* modelo simple; f_t representa el valor pronosticado del modelo ensamblado. El modelo ensamblado se creó a partir de la ecuación IV.25.

$$f = w_1\hat{y}_1 + w_2\hat{y}_2$$
$$s.t. \quad w_1 + w_2 = 1$$

Ecuación IV-25

En este punto se aplicó el algoritmo PSO mejorado a través de las ecuaciones IV.5 para establecer la velocidad de cada partícula y IV.2 para su posición y con ello se calcularon los pesos w_1 y w_2, primeramente se seleccionaron los parámetros del algoritmo los cuales incluyen el tamaño de población *m*, las constantes de aceleración φ_1 y φ_2, la velocidad máxima v_{max} y el número máximo de generaciones t_{max}. La tabla IV.5 muestra el algoritmo PSO mejorado.

Tabla IV-5. Algoritmo PSO mejorado.

- Entradas:
  - Función de aptitud a optimizarse.

1) Iniciar el contador de generación es $t = 0$.
2) Iniciar aleatoriamente las posiciones x_i y v_i de las n partículas del cúmulo S.
3) Evaluar la función objetivo con las posiciones x_t.
4) Encontrar $p_{best.}$
5) Encontrar p_{gbest}.
6) Mientras $t < maxG$ Hacer
   - Para $i \leftarrow 1$ hasta size(S) Hacer
     i. Generar una nueva velocidad v_{t+1}. Empleando ecuación IV.5.
     ii. Calcular las nuevas posiciones x_{t+1}. Empleando ecuación IV.6.
     iii. Evaluar la función objetivo con las nuevas posiciones x_{t+1}.
   - Fin Para
   - Encontrar p_{best}.
   - Encontrar p_{gbest}.
   - $t = t + 1$.
7) Fin mientras.
8) Regresar la mejor solución encontrada.

En algunos casos, se manejan los valores de frontera v_{max} y v_{min} para controlar las velocidades de las partículas. Un valor muy pequeño de la velocidad puede motivar que las partículas queden atrapadas en un óptimo local, por otra parte, un valor muy grande puede propiciar que las partículas oscilen alrededor de la posición. La idea de trabajar con el algoritmo modificado de PSO en esta propuesta fue mejorar el desempeño mediante el balanceo de la capacidad de búsqueda global y local. Así un peso inercial grande facilita una búsqueda global, mientras que uno pequeño agiliza la búsqueda local. El decremento lineal del peso inercial de un valor relativamente grande a uno pequeño durante la ejecución de PSO, propicia una tendencia de búsqueda más global al inicio de la corrida que va decayendo hacía una búsqueda mayormente local al final de la ejecución (Y. Shi & Eberhart, 1998). El cálculo del peso inercial se logró mediante la ecuación IV.26:

$$w_t = w_{max} - \frac{w_{max} - w_{min}}{t_{max}} * t \qquad \textbf{Ecuación IV-26}$$

Para poder evaluar la eficacia del clasificador se recurrió a la función de confiabilidad que se describe en la ecuación IV.27.

$$CP = \frac{VP + VN}{VP + FP + FN + VN} \qquad \textbf{Ecuación IV-27}$$

Teniendo en cuenta que se considera disminuir el valor del error Tipo II, o índice de falso negativo según se analizó en el capítulo II, la función de aptitud del algoritmo se desarrolló para minimizar el comportamiento del algoritmo como lo define la ecuación IV.28.

$$F = M * \left(\frac{1}{n_1} \sum_{i=1}^{n_1} (y_i - f_i)^2 + k * \frac{1}{n_2} \sum_{j=1}^{n_2} (y_j - f_j)^2 \right) \qquad \textbf{Ecuación IV-28}$$

Donde, n_1 y n_2 representan el número de ejemplos acreditados y rechazados respectivamente; *f*, *y* representan la salida actual del modelo ensamblado y la salida esperada; *M* es un número positivo grande empleado para realizar el cambio de aptitud observable de las partículas; *k* es una variable empleada para controlar la ocurrencia de el error tipo II. El error tipo II se obtuvó mediante la aplicación de la ecuación IV.29.

$$ErrT2 = \frac{FP}{VP + FP} \qquad \textbf{Ecuación IV-29}$$

El esquema del modelo empleado con las consideraciones descritas se muestra en la Figura IV.4 a continuación:

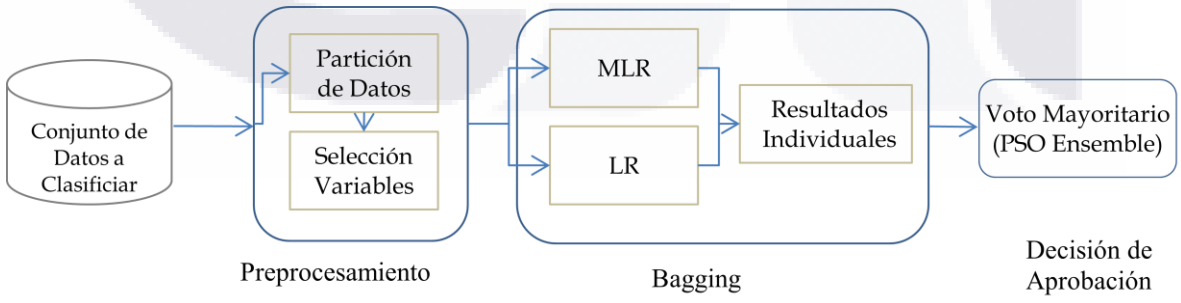


Figura IV-4 Modelo de puntuación crediticia mediante ensamble.

La tabla IV.6 muestra el algoritmo general propuesto para la obtención de puntuación crediticia de manera ensamblada.

Tabla IV-6. Algoritmo propuesto de cálculo de puntuación crediticia mediante ensamblado.

- Entradas:
    a. Conjunto de Datos a partir de PCAPSO $X = [d_1, d_2, \ldots, d_n]$, done d_i= prestatario indvidual, n = # de prestatarios. Acorde al algoritmo de la Figura IV.3.
    b. $M \leftarrow$ Número de predictores (2; LRA y LR).
    c. Función a optimizar con PSO, descrita en la ecuación IV.28.
- Salida:
    a. Puntuación crediticia ensamblada.

1) Calcular los valores predictores empleando Bagging.
    a. Utilizar LRA y LR como predictores, empleando las ecuaciones III.1 y III.2 respectivamente.
    b. según el algoritmo de la tabla IV.4.
2) Obtener los pesos de voto mayoritario mediante PSO repitiendo los pasos del algoritmo PSO mejorado de la tabla IV.5.
    a. Emplear como función objetivo la ecuación IV.28.
    b. Adecuar el valor inercial empleando la ecuación IV.26.
    c. Entregar los valores de los pesos de voto mayoritario w_1 y w_2 de la ecuación IV.25.
3) Calcular la puntuación crediticia ensamblada por medio de la ecuación IV.25
4) Regresar el valor de la puntuación crediticia ensamblada.

Este modelo de puntuación crediticia se aplicó de igual forma a los prestatarios que se conjuntaron en agrupamientos como se explica en la sección siguiente.

IV.3.3.3 Modelado con agrupamiento de prestatarios.

Tradicionalmente el análisis de agrupamientos se ha empleado como una herramienta descriptiva, en la cual el algoritmo se emplea para crear grupos de observaciones a partir de sus características. En este trabajo con la idea de mejorar el desarrollo de tarjetas de puntuación crediticia se manejó el análisis de agrupamiento no solo como una metodología para clasificar individuos con algunas características específicas (variables), sino también como una parte de un proceso de predicción en conjunto con el modelo de clasificación ensamblado; para obtener buenos resultados al momento de clasificar pero también para conocer los perfiles de los clientes nuevos que se incorporan a la IMF.

El modelo de agrupamiento se sustentó en una propuesta mejorada de K-medias que en términos generales intenta asignar un conjunto de n prestatarios (observaciones) en un

número *k* de agrupamientos donde cada observación se ubica en el agrupamiento en base a la distancia mínima hacía el centroide del agrupamiento, adoptando la distancia como el índice de evaluación de similitud. El centroide de un agrupamiento es el valor medio de todas las observaciones en el agrupamiento, además cada prestatario solo puede pertenecer a un agrupamiento.

El proceso completo de K-medias se puede dividir en dos fases. La primera consiste en asignar cada observación al agrupamiento con el centroide más cercano. En la segunda se calculan nuevos centroides de los agrupamientos de acuerdo a los prestatarios que lo forman al finalizar el paso uno. Este proceso se repite hasta que los agrupamientos permanecen sin cambios o se satisfacen ciertas condiciones de paro. Para la inicialización de los prestatarios a los agrupamientos, usualmente se emplea un método de partición aleatorio. Los procedimientos más comunes son: i) asignar aleatoriamente cada observación a un agrupamiento y proceder la fase 2 o ii) elegir aleatoriamente *k* prestatarios y fijar la posición de estas observaciones como los centroides de los *k* agrupamientos. En general, el algoritmo de k-medias funciona como se describe a continuación en la tabla IV.7.

Tabla IV-7 Algoritmo general del método K-medias

1) Introducir el número *k* de los agrupamientos a considerar, y los datos a agrupar.
2) Inicializar aleatoriamente el prestatario m_j como el centroide de cada uno de los *k* agrupamientos.
3) Repetir:
   a. Para cada prestatario x_i, calcular la distancia entre él y cada uno de los centroides m_j, para asignar a x_i al agrupamiento con el centroide más cercano, típicamente se emplea la fórmula de la distancia Euclideana para calcular la distancia, en esta propuesta se aplica la fórmula de la distancia de Manhattan como lo muestra la ecuación IV.30:

$$d(x_i, m_j) = \sum_{k=1}^{D} |x_{ik} - m_{jk}| \qquad \text{Ecuación IV-30}$$

   b. Empleando la formula IV.31 para calcular el centroide de cada agrupamiento:

$$m_j = \frac{1}{n} \sum_{\forall x_i \in C_j} x_i \qquad \text{Ecuación IV-31}$$

Hasta que no se satisfagan las condiciones de paro.

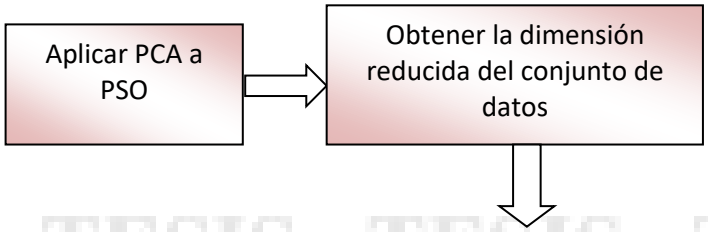
En donde la simbología empleada en el algoritmo se definió como sigue:

- D denota la dimensión del conjunto de datos de entrada.

- *k* denota el número de agrupamientos a definir.
- x_i denota el *i-ésimo* vector que representa al *i-ésimo* prestatario.
- m_j denota el *j-ésimo* vector que representa el centroide del *j-ésimo* agrupamiento.
- n_j denota el número de prestatarios pertenecientes al *j-ésimo* agrupamiento.
- C_j denota el subconjunto de observaciones del *j-ésimo* agrupamiento.

La propuesta de este trabajo presentó una mejora a k-medias que consistió en emplear los resultados del análisis de componentes principales y la hibridación de algoritmos PSO para el cálculo de los agrupamientos acorde a los atributos del conjunto de prestatarios. El uso de métodos heurísticos y metaheurísticos se ha vuelto más apropiado sobre todo para la resolución de problemas de gran tamaño cuando el problema de agrupación se convierte en NP-completo. Esto se produce porque las funciones criterio de similitud no son convexas ni lineales de manera que el problema de la agrupación resultante puede tener soluciones de mínimos locales. Por otra parte, los problemas de agrupamiento muestran una complejidad exponencial en términos del número de agrupamientos que generan, teniéndose que el problema de agrupamiento se vuelve de igual forma NP-completo cuando el número de agrupamientos excede a tres (Welch, 1982).

La matriz de puntuación generada por PCA se explotó como inicio para definir los datos de agrupamiento y la cantidad de agrupamientos a emplear, que es uno de los requisitos para uso de K-medias. Este procedimiento no solo provoca una reducción de la dimensión original de los datos, sino que también atenúa el problema de multico-linealidad entre los atributos mejorando la densidad de la muestra y originando resultados que resultados más precisos y a la vez confiables. La figura IV.5 muestra gráficamente el fundamento de la propuesta.



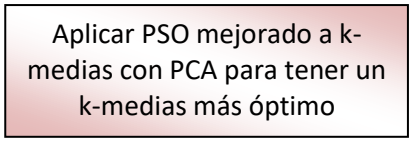


Figura IV-5 Modelo K-medias mejorado mediante la hibridación de PCA, PSO.

En este modelo de agrupamiento con PSO, una partícula representa los *k* centroides de los agrupamientos, es decir, la definición de cada partícula está dada por IV.32.

$$z_p = (m_{p1}, \cdots, m_{pj}, \cdots, m_{pk}) \qquad \textbf{Ecuación IV-32}$$

Donde m_{pj} es el *j-ésimo* prestatario centroide de la *p-ésima* partícula en el agrupamiento C_{pj}. Por lo tanto, un enjambre de partículas representa una serie de agrupamientos candidatos para las observaciones actuales. El valor de la función de aptitud se obtiene a través de los errores de cuantificación como lo muestra la ecuación IV.33.

$$J_e = \frac{\sum_{j=1}^{k}\left[\sum_{\forall x_i \in C_{pj}} \frac{d(x_i, m_{pj})}{|C_{pj}|}\right]}{k} \qquad \textbf{Ecuación IV-33}$$

Donde $|C_{pj}|$ es el número de prestatarios que pertenecen al agrupamiento C_{pj}.

Así, el proceso del algoritmo hibridizado de K-medias, PCA y PSO se puede describir como se muestra en la tabla IV.8:

Tabla IV-8 Algoritmo mejorado de agrupamiento a través de la hibridación de PCA, PSO y K-medias.

- Entradas:
  a. Conjunto de Datos a partir de PCA X = [d_1, d_2,…., d_n], done d_i= prestatario indvidual, *n* = # de prestatarios.
  b. Centroides de Agrupamientos *M* = [m_1, m_2,…., m_k],

c. Número máximo de iteraciones *max_iter.*

1. Seleccionar aleatoriamente el centroide para cada partícula.
2. Inicializar
    a. La posición original de cada partícula.
    b. La velocidad de cada partícula.
    c. La mejor posición individual de cada partícula $l_{best.}$
    d. La mejor posición global de las partículas $g_{best.}$
3. Por cada observación d_i, encontrar el centroide m_j asignando d_i al agrupamiento con el centroide m_j más cercano.
4. Repetir los pasos 5-9 hasta que se cumpla el criterio de paro o no se encuentren nuevos centroides.
5. Actualizar
    a. La mejor posición local de cada partícula empleando la ecuación IV.3
    b. La mejor posición global empleando la ecuación IV.4
    c. La velocidad de cada partícula empleando la ecuación IV.5
    d. La posición de cada partícula empleando la ecuación IV.2
6. Recalcular los centroides para cada agrupamiento C_j con $(1 \leq j \leq k)$
7. Para cada partícula d_i obtener la distancia del centroide m_j del agrupamiento actual más cercano
8. Si la distancia calculada es menor o igual que la distancia previa calculada entonces el punto se queda en el agrupamiento previo
9. Si no, calcular la distancia de la observación a cada uno de los nuevos centroides y asignarla al agrupamiento más cercano en base a las distancias de los centroides
10. Ejecutar algoritmo k-medias
    a. Asignar los *k* centroides generados a partir del algoritmo PSO.
    b. Calcular la distancia de cada prestatario d_n y el *k* centroide m_k empleando distancia Manhattan.
    c. Recalcular los centroides para cada agrupamiento C_j con $(1 \leq j \leq k)$
    d. Asignar la observación al agrupamiento correspondient.

El diseño del modelo propuesto de agrupación de clientes se puede observar en la Figura IV.6.

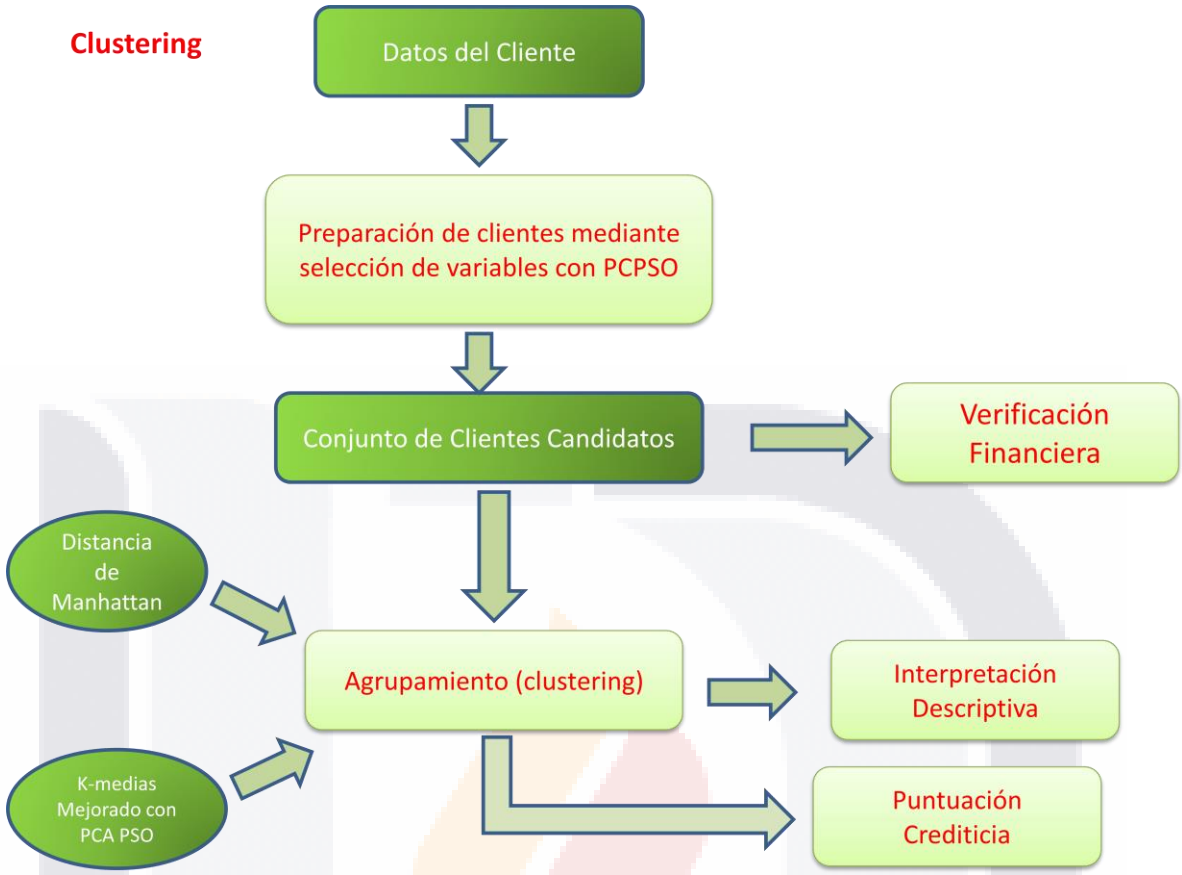


Figura IV-6 Modelo de Agrupación de Clientes.

Una vez que definidos los agrupamientos, la puntuación crediticia se alcanzó a partir de la probabilidad de que el nuevo cliente pertenezca a cada grupo. Por tanto, el resultado final se derivó de la media ponderada de las calificaciones de riesgo de crédito obtenidas a partir de las agrupaciones resultantes, donde la ponderación se estableció como una la probabilidad de pertenencia a cada una de ellas. Para obtener la probabilidad de pertenecer a un grupo sencillo fue necesario hacer una conversión dado que la salida es una medida de distancia. Para activar la distancia que resulta en una probabilidad, se utilizó la ecuación IV.34.

$$Pertenencia\ Cluster\ i \Longrightarrow pc_i = \frac{\frac{1}{distancia\ \ al\ cluster\ i}}{\sum_{i=1}^{k} \frac{1}{distancia\ \ al\ cluster\ i}} \qquad \textbf{Ecuación IV-34}$$

Por su parte, los valores ponderados de pertenencia a un grupo se obtuvieron a través de la ecuación IV.35.

$$valor\ ponderado = \frac{\sum_{i=1}^{k}((w_1\hat{y}_1 + w_2\hat{y}_2) \times pc_i)}{k} \qquad \text{Ecuación IV-35}$$
$$s.t. \quad w_1 + w_2 = 1$$

Donde $w_1\hat{y}_1 + w_2\hat{y}_2$ corresponde a la función de cálculo de puntuación crediticia como se enuncia en la ecuación IV.25, pc_i corresponde a la probabilidad de pertenencia a un agrupamiento según lo establecido por la ecuación IV.34 y k es el total de agrupamientos generados. De esta forma se obtiene la puntuación crediticia a partir de los agrupamientos generados para cada uno de los prestatarios.

IV.3.3.4 Complejidad Computacional de los modelos propuestos.

La complejidad computacional estudia la eficiencia de los algoritmos estableciendo su efectividad de acuerdo al tiempo de corrida y al espacio requerido en la computadora o almacenamiento de datos, ayudando a evaluar la viabilidad de la implementación práctica en tiempo y costo.

El análisis asintótico permite conocer la eficiencia de un algoritmo con base en el tiempo de corrida cuando el tamaño de los datos de entrada es suficientemente grande de tal forma que las constantes y los términos de menor orden no afectan. Los tiempos de corrida de un algoritmo pueden dividirse en: el mejor, el probabilístico y, el peor. Sin embargo, para analizar un algoritmo normalmente se considera el peor caso, es decir, el tiempo más largo para cualquier entrada de tamaño n. Las notaciones para el tiempo de corrida asintótico de un algoritmo se definen en términos de funciones cuyo dominio es el conjunto de números naturales. Las notaciones se utilizan para describir la función del tiempo de corrida, $T(n)$, normalmente definidas en tamaños de entrada enteros. Este tipo de análisis permite facilitar la elección del mejor algoritmo entre varios y, por supuesto, es más conveniente aplicar medidas para la eficiencia que implementarlo y medir la eficacia después de cada corrida.

En el análisis asintótico la Notación O representa una cota asintótica superior. Sea una función $g(n)$ se denota por O ($g(n)$) el conjunto de funciones:

$$O\ (g(n)) = \{f\ (n):\ \exists \text{ constantes positivas } c \text{ y } n_0 \mid 0 \leq f\ (n) \leq\ cg(n)\ \forall\ n \geq n^0\ \}$$

La notación O se utiliza para acotar el peor caso del tiempo de corrida de un algoritmo.

De esta forma la complejidad computacional empleando la Notación O de los algoritmos empleados en esta tesis fue la que se muestra a continuación.

- Regresión Lineal Múltiple. De acuerdo a la ecuación III.1, para poder llevar a cabo su cálculo se requiere reescribir utilizando notación matricial, quedando de la siguiente forma: $Y = X\beta + \epsilon$, de donde los estimadores mínimo cuadráticos se obtienen a partir de la ecuación IV.36.

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad \textbf{Ecuación IV-36}$$

  Debido a que el cálculo de regresión múltiple implica una inversión matricial que tiene una complejidad computacional de grado O (n^3) qué es el máximo grado de todas las operaciones entonces la complejidad de MLR es de O (n^3).

- Regresión Logística. Partiendo de la ecuación III.2 expresada de la forma $\pi(x) = E(y|x)$ que representa la media condicional de $y = 1$ dado x, donde $\pi(x)$ representa la probabilidad de que ocurra $y = 1$, se espera una relación curvilínea con propiedades de una función de distribución acumulada, la cual se puede expresar mediante la función de distribución acumulada de la distribución logística dada por la ecuación IV.37.

$$\pi(x) = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} \qquad \textbf{Ecuación IV-37}$$

  La transformación de la ecuación IV.37 empleando la transformación logit lleva a la representación logarítmica de la ecuación III.2 la cual produce una complejidad computacional para construir el modelo de regresión logística es de O (nv^2c), donde n denota el número de ejemplos de entrenamiento v el numero de atributos y c el numero de clases.

- Análisis de componentes principales. El costo computacional para encontrar vectores propios de la matriz de auto correlación M es de un grado de complejidad cúbica $O(d^3)$. Donde es la dimensionalidad de los datos por la cantidad de registros contenidos en el data set, en el caso de estudio se tienen 29 variables originales por 8162 registros de prestatarios, siendo el valor de d = 29*8162. De la teoría matricial se tiene que si el número de registros en el conjunto N es menor que la dimensión de la matriz de

covarianza *M*, entonces *M* es singular y sus filas mayor que *N*. Los vectores propios a continuación, pueden ser determinados por una combinación lineal de la formación *N* de registros. La complejidad computacional se reduce a $O(N^3)$. En conclusión, la complejidad de cálculo puede expresarse simplemente como $O(r^3)$, donde $r = min(n, d)$.

- Optimización por cúmulo de partículas. Este es un modelo simple computacional cuyo grado de complejidad es lineal y está definido por O (*n*), según lo ilustra la literatura.
- K-medias. Respecto a la complejidad computacional, el agrupamiento K-medias para problemas en espacios de *d* dimensiones es:
    a. NP-completo en un espacio euclidiano general *d* incluso para 2 grupos.
    b. NP- completo para un número general de grupos *k* incluso en el plano.

  Así, para la fase 1 del cálculo de centroides y la redistribución de los puntos en los agrupamientos, implica una complejidad computacional de O (*nKld*), donde *n* es el total de punto, *K* es el número de agrupamientos definidos, *l* es el número de iteraciones y *d* la cantidad de atributos (variables) que maneja el sistema.

La tabla IV.9 muestra el conglomerado de la complejidad computacional de los algoritmos empleados en esta propuesta.

Tabla IV-9Complejidad Computacional de los algoritmos empleados en las propuestas de tesis.

Algoritmo	Complejidad Computacional
MLR	O (n^3)
LR	O (nv^2c)
PCA	O (r^3) r= min [*n,d*]
PSO	O (*n*)
K-medias	O (*nKld*)

La complejidad computacional que manejan los algoritmos empleados es polinomial como se puede observar, por lo tanto el grado de complejidad de los modelos propuestos en el trabajo de tesis es de igualmente polinomial y está sujeto a los valores correspondientes al método de k-medias, por lo tanto el modelo con un mayor grado de complejidad computacional de los aquí propuestos es el del modelo de establecimiento de puntuación crediticia con agrupamientos.

Las fases restantes del modelo CRISP para su aplicación en este trabajo se analizan en los siguientes capítulos.

V

Capítulo V Evaluación y Resultados a partir de la Solución Propuesta

V. 1 Variables seleccionadas

La selección de variables se realizó a partir de los datos de una microfinanciera con 8190 prestatarios registrados y un total de 138 variables, de las cuales solamente 60 resultaron válidas para aplicarse al contexto del problema. El manejo de una cantidad de variables tan grande resulta en un problema intratable con la tecnología computacional disponible motivo por el cual se procedió a su simplificación, partiendo de los métodos de selección y pre-tratamiento se obtiene una cantidad de 30 variables, que son las que cumplieron los requisitos del coeficiente de Kaiser-Meyer-Olkin (KMO), una medida de aplicación muestral que contrasta si las correlaciones parciales entre las variables son suficientemente pequeñas, un valor menor a 0.5 indica que un análisis factorial no debe realizarse con los datos de la muestra empleada, ver tabla A0-1 del anexo A como ejemplo del formato de los datos. La información concentrada de los datos, como en la gran mayoría de los casos de minería de datos estaban incompletos y con irregularidades por lo que se requirió el empleo de un pretratamiento y de normalización de la información acorde a las ecuaciones IV.11 y IV.12, además de una adecuación de los valores nulos o inválidos. Una muestra de los datos derivados de este procedimiento se muestra en la tabla A0-2 del anexo A. El listado de variables resultantes se muestra en la tabla V.1 en donde se consideraron 29 variables definidas como x_i y una variable de salida y para cada ejemplo listado en la tabla.

Tabla V-1 Definición de Variables.

Variable	*Índice*	*Cálculo*
x_1	CODIGO	VALOR ACTUAL
x_2	SEXO	FEMENINO=1, MASCULINO=2
x_3	CALLE	VALOR ACTUAL
x_4	TELEFONO	TIENE TELEFONO=1, NO TIENE TELEFONO=2
x_5	EDOCIVIL	CASADO=1, SOLTERO=2, DIVORCIADO=3, UNIÓN LIBRE=4, VIUDO=5
x_6	NIVESCOLAR	PRIMARIA=1, SECUNDARIA=2, BACHILLERATO=3, LICENCIATURA=4, COMERCIAL=5, RELIGIOSO=6,NINGUNA=7
x_7	NACIMIENTO	VALOR ACTUAL
x_8	NACIOMU	CODIGO DEL MUNICIPIO EN EL PAIS
x_9	NACIOLO	CODIGO DE LA LOCALIDAD EN EL PAIS
x_{10}	NODEPEND	VALOR ACTUAL
x_{11}	ALTA	VALOR ACTUAL
x_{12}	RESTRICCION	NINGUNA=0, GENERAL= 1, NOMINAL=2
x_{13}	REGMARITAL	BIENES SEPARADOS=1, BIENES MANCOMUNADOS=2, SOLTERO=3
x_{14}	CDGPRPE	VALOR ACTUAL
x_{15}	CDGOCPE	VALOR ACTUAL
x_{16}	SOLICITUD	VALOR ACTUAL
x_{17}	PERIODICIDAD	SEMANAL=1, QUINCENAL=2, MENSUAL=3
x_{18}	CANTAUTOR	VALOR ACTUAL
x_{19}	CANTENTRE	VALOR ACTUAL
x_{20}	TASAINI	VALOR ACTUAL
x_{21}	DURACINI	VALOR ACTUAL
x_{22}	TASARECFIJ	VALOR ACTUAL
x_{23}	EMPLSOSTH	VALOR ACTUAL
x_{24}	EMPLOSTM	VALOR ACTUAL
x_{25}	MODOAPLIRECA	VALOR ACTUAL
x_{26}	TASA	VALOR ACTUAL
x_{27}	ABONOS	VALOR ACTUAL
x_{28}	CANTENTRE_1	VALOR ACTUAL
x_{29}	FECHA_TERMINO_ PRESTAMO	VALOR ACTUAL
Y	SALDO	ADEUDO=1 BIEN=2

Con los datos limpios después del pretratamiento se realizó el cálculo de análisis de componentes principales, mediante la aplicación de las ecuaciones IV.14 a IV.19 con la cual se obtuvieron 8 diferentes componentes principales, considerándose solamente los 4 primeros que alcanzan una tasa de contribución acumulada superior al 82%, la tabla A0-3 del anexo A muestra esta información, el cual de manera empírica es considerado un valor adecuado para lograr una buena representatividad de los datos originales. La Figura V.1 muestra el aporte de cada eigenvalor obtenido, para definir la cantidad de componentes a considerar.

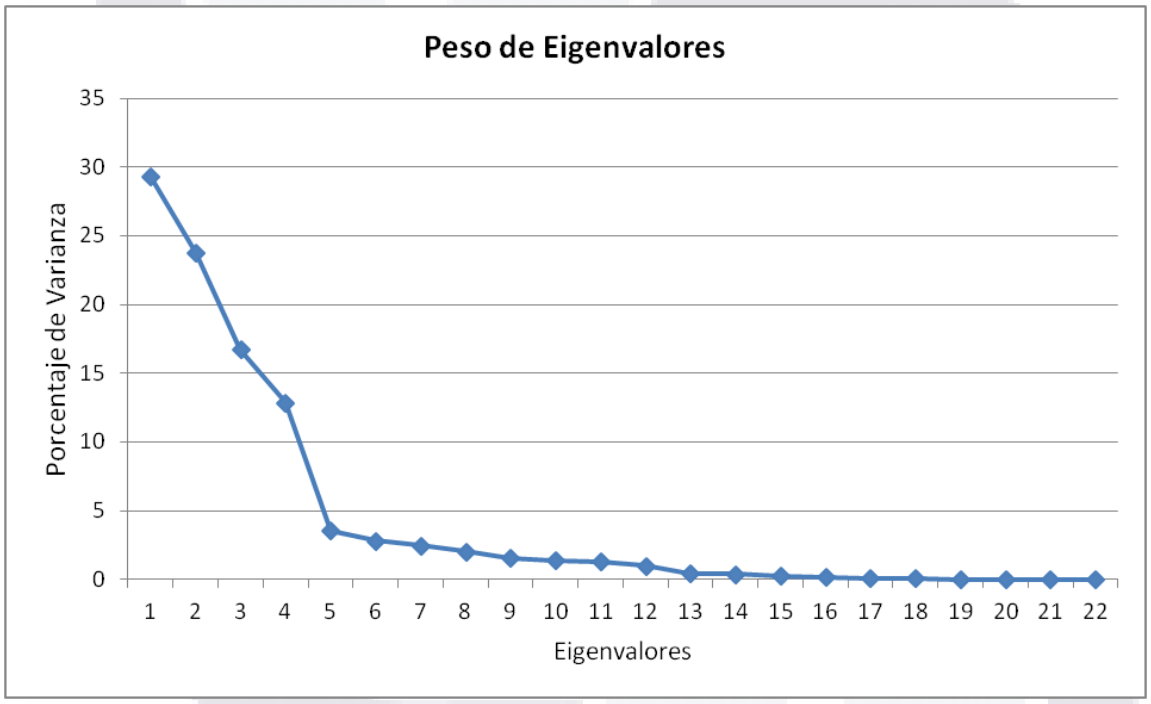


Figura V-1. Porcentajes de contribución de los eigenvalores calculados.

La tabla V.2 muestra los resultados relevantes obtenidos. Los valores obtenidos del error estándar en los cuatro componentes principales mostraron que la confiabilidad del modelo es aceptable, además de que al no presentarse un alto grado de dispersión en los datos se pudo descartar la existencia de anomalías dentro de la información que dieran una gran variabilidad a la muestra, teniéndose de igual forma una homogeneidad adecuada como lo sugirieron los resultados de los cocientes de varianza de los mismos componentes principales.

Tabla V-2. Valores representativos de la aplicación de PCA.

	Componente Principal 1	Componente Principal 2	Componente Principal 3	Componente Principal 4
Error estándar del componente principal	4.29	3.78	2.25	1.68
El cociente de varianzas	0.272	0.208	0.146	0.108
Tasa de contribución acumulada	27.289	20.733	18.715	15.857

El peso de las 29 variables independientes se muestra en la tabla V.3, debe tenerse en cuenta que las variables representativas para cada componente son aquellas que se aproximan más al valor de 1, de esta forma las variables que se consideran son las representativas de cada componente para poder realizar la definición final mediante la aplicación de PSO.

Tabla V-3. Valor de las variables para los componentes principales.

Variable	Componente Principal 1	Componente Principal 2	Componente Principal 3	Componente Principal 4
X1	-0.162132	-0.260985	0.196457	0.41135
X2	0.246675	0.742675	-0.21168	-0.1405
X3	0.182852	-0.12269	-0.18286	0.56283
X4	-0.00135	-0.33043	-0.23193	0.20807
X5	-0.23818	-0.42548	0.102875	0.181875
X6	0.33485	-0.21753	0.169625	0.19125
X7	0.1397	-0.1016	0.3366	0.8038
X8	0.0026	0.211525	0.230225	0.369
X9	0.10923	0.319154	0.166437	-0.64593
X10	-0.24078	0.145425	0.350725	-0.13638
X11	0.342025	0.7769	0.715475	-0.5113
X12	-0.20633	0.167975	-0.62988	0.6294
X13	-0.14733	-0.51513	0.324925	0.37975
X14	0.234125	-0.37058	0.2087	0.27975
X15	0.22565	-0.1317	0.2325	0.1435
X16	-0.15513	0.784125	0.807675	0.265
X17	0.34835	-0.19435	0.38625	-0.17388
X18	0.71715	-0.24115	0.26325	0.1092
X19	0.14185	0.11533	-0.40123	0.04052
X20	-0.38625	0.288475	-0.0948	0.034
X21	0.3442	0.326975	-0.1295	0.3548

Variable	Componente Principal 1	Componente Principal 2	Componente Principal 3	Componente Principal 4
X22	-0.56588	0.8177	0.706425	0.22175
X23	0.146275	0.752625	-0.1418	-0.1533
X24	-0.24628	-0.24263	0.14175	0.15325
X25	0.37005	-0.1341	-0.3763	-0.054
X26	0.334132	-0.67063	0.0872	0.27981
X27	0.7207	-0.32373	0.134	0.868525
X28	-0.04635	-0.74267	0.1418	0.15336
X29	0.03315	-0.10186	-0.1555	0.25543

Una muestra de la representación de la partícula empleada se ejemplifica en la Figura V.2, en donde cada uno de los componentes del vector que define al prestatario corresponde a las variables seleccionadas para realizar la optimización de la selección de variables asignándose un valor de 1 en los casos donde la variable será considerada, mientras que un valor distinto significa que la variable no se tomará en cuenta.

x_2	x_7	x_{11}	x_{12}	x_{13}	x_{16}	x_{18}	x_{21}	x_{22}	x_{23}	x_{27}
1	1	0	1	1	1	0	0	0	1	1

Figura V-2. Representación de una partícula correspondiente a un Prestatario.

En el caso de la partícula presentada se tomarían en cuenta las variables *1,7, 11, 12, 13, 16, 21, 23 y 27*. Los resultados arrojados por la aplicación de PCA sirvieron como punto de partida para llevar a cabo su optimización mediante la ejecución de PSO binario debido a la codificación de la partícula es esta forma. La función de velocidad de la partícula definida por la ecuación IV.5 dado que se utilizó la versión mejorada de PSO, se empleó en este caso como distribución de probabilidad base para el cálculo de la ecuación de la posición de la partícula, como lo define la ecuación V.1.

$$X_i^{t+1} = \begin{cases} 1, & aleatorio < \dfrac{1}{1 + e^{-v_i^{t+1}}} \\ 0, & aleatorio > \dfrac{1}{1 + e^{-v_i^{t+1}}} \end{cases}$$

Ecuación V-1

Los parámetros empleados para la función de velocidad después de calibrar el algoritmo fueron los siguientes:

- Tamaño de enjambre es 25.
- Grado de conocimiento individual (φ_1) se fija en 2.
- Grado de conocimiento grupal (φ_2) de igual forma se fija en 2.
- Los coeficientes de aceleración β_1 y β_2 se mantienen con un límite máximo β_{max} de 1.
- Peso inercial *w* para aplicar PSO mejorado es de 0.6
- Condición de paro es de 90 iteraciones.
- Se realizan 10 repeticiones para cada enjambre seleccionado.

En el anexo B se realiza una explicación de la calibración de parámetros del algoritmo PSO.

La función de aptitud de la ecuación IV.23 se empleó para conducir la evolución del modelo en términos de su capacidad para maximizar la separación de clases de prestatarios, alcanzándose la selección de clases más adecuadas de las variables, así como produciendo una distribución del espacio de búsqueda para la optimización que se ilustra de forma parcial en la figura V.3.

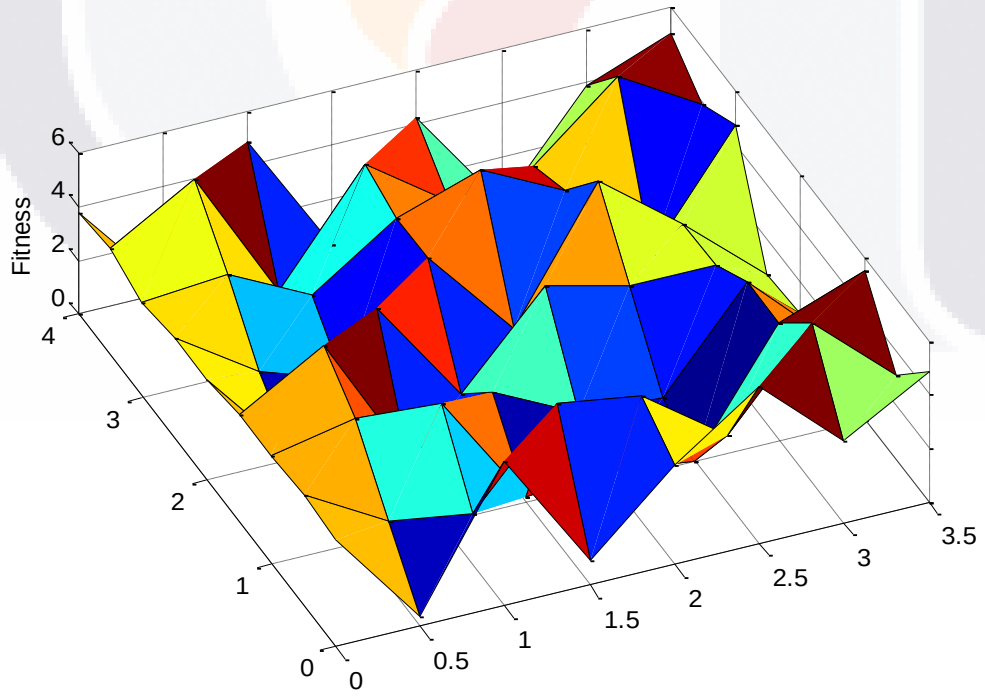


Figura V-3 Espacio de optimización en la función de aptitud en la selección de variables.

De la figura se puede observar que los mayores valores de la función de aptitud son muy cercanos a 6, así que las partículas próximas a este valor fueron las candidatas a definir la selección de variables. El valor mayor alcanzado de la función de aptitud a través de todas las ejecuciones fue el de 5.88173, aunque el valor promedio fue el de 5.84046, que fue el valor considerado para elegir la partícula representativa que entregó la cantidad de variables a emplear en el resto del desarrollo de la propuesta. La tabla V.4 muestra los valores con las variables seleccionadas de acuerdo a la los mejores valores de la función de aptitud arrojados para cada una de la 10 repeticiones del algoritmo.

Tabla V-4.Resultados generados en cada una de la ejecución del método.

Repetición	*fitness*	*x2*	*x7*	*x11*	*x12*	*x16*	*x17*	*x18*	*x21*	*x22*	*x23*	*x27*
1	5.85009	1	1	1	1	1	1	1	1	1	0	1
2	5.819266	1	1	1	1	1	1	0	1	1	1	1
3	5.801346	1	0	1	1	1	1	1	1	1	1	1
4	5.811911	0	1	1	1	1	1	1	1	1	1	1
5	5.839543	1	1	1	1	1	1	1	1	1	0	1
6	5.855017	1	1	1	1	1	1	1	1	0	1	1
7	5.88173	1	1	1	1	1	1	1	0	1	1	1
8	5.863452	1	1	1	1	1	1	1	1	1	1	0
9	5.852926	1	1	1	0	1	1	1	1	1	1	1
10	5.837325	1	1	1	1	1	1	1	1	1	0	1

La aplicación de PSO con los parámetros definidos genera una selección de variables como la que se muestra en la Figura V.4, tomando el criterio de elegir el promedio de la función de aptitud obtenida.

x_2	x_7	x_{11}	x_{12}	x_{13}	x_{16}	x_{18}	x_{21}	x_{22}	x_{23}	x_{27}
1	1	1	1	1	1	1	1	1	0	1

Figura V-4. Variables seleccionadas al aplicar PSO.

Así la variable x_{23} se descartó del conjunto de variables que se emplearon para trabajar en el modelo de puntuación crediticia propuesto. Esta cantidad de variables resultó adecuada para poder implementar los modelos de regresión lineal y regresión logística en el cálculo, como lo ilustra la siguiente sección. En la tabla C1 del anexo C se puede observar el código central del cálculo de PCA en Java empleando la librería Efficient Java Matrix Library, además en las tablas C2, C3, C4 y C5 del mismo anexo se presenta el código del algoritmo

PSO, para la actualización de la mejor posición de la partícula p_{best}, la actualización de la mejor partícula global g_{best}, el cálculo de inercia del algoritmo mejorado y el movimiento de partículas respectivamente.

V. 2 Puntuación Crediticia

La probabilidad de rechazo en el otorgamiento de un crédito por parte de las IMFs es muy baja y en un alto porcentaje de los casos se entrega la cantidad solicitada al prestatario, así que no dentro de la información de 8190 registros todos correspondían a clientes a los que se les asignó un crédito, para solucionar este conflicto se optó por considerar los buenos pagadores y los malos pagadores, entendiéndose por malos pagadores aquellos clientes que definitivamente no liquidaron su deuda, o bien tuvieron un retraso mayor a 90 días en el finiquito de sus pagos, de acuerdo a lo establecido por el concilió de Basilea (Basel, 2010), resultando una clasificación del 58.33% de malos pagadores y un 42.66% de buenos pagadores. Así que se realizó una nueva selección de los datos para ajustar la proporción de prestatarios buenos y malos, empleando un método de muestreo aleatorio estratificado. Primeramente se dividió el conjunto total de datos en los grupos de buenos pagadores y de malos pagadores, obteniéndose así dos grupos de los cuales de manera aleatoria se extrajeron 1000 ejemplos respectivamente, con lo que se logró una proporción de buenos pagadores y de malos pagadores muy cercana a 1:1. De los 2000 ejemplos conseguidos para el modelo, se dejaron 1000 para entrenamiento compuestos por 500 clientes buenos y 500 clientes malos y los otros 1000 sirvieron para prueba con una proporción igual de 500 prestatarios buenos y malos, considerando para cada uno de los ejemplos solamente las variables seleccionadas mediante el proceso de elección de variables.

A partir de las adecuaciones realizadas en la información se desarrolló el modelo de puntuación crediticia mediante un enfoque ensamblado como se comenta en el capítulo IV. Primero, se construyeron los dos modelos simples empleando regresión lineal múltiple y regresión logística como herramientas de puntuación crediticia a la cuales se les fijó el valor crítico de 0.5, es decir aquellos pronósticos mayores a 0.5 se consideraron como buenos pagadores, mientras que los otros se interpretaron como clientes malos. Debido al enfoque ensamblado de bagging, cada uno de los modelos simples se debieron de entrenar

con un conjunto de datos distintos, por lo cual de los 1000 registros de entrenamiento, 500 se tomaron para calcular el valor de MLR con una proporción 1:1 entre buenos y malos clientes, mientras que para calcular la ecuación de LR se consideraron los 500 restantes registros de entrenamiento. Con los resultados arrojados por estos modelos MLR y LR el segundo paso consistió en construir el modelo ensamblado mediante la combinación de los dos modelos y el algoritmo PSO para buscar los pesos correspondientes a MLR y LR en el modelo ensamblado definido por la ecuación IV.25.

V.2.1 Modelo MLR.

El modelo de regresión lineal múltiple se construyó a partir de la ecuación III.1, la cual como se menciona en el capítulo III se elaboró estimando los valores de los coeficientes beta del modelo de regresión empleando el paquete estadístico OpenStat. Es necesario indicar que estos coeficientes no son independientes entre sí. De hecho, el valor concreto estimado para cada coeficiente se ajusta teniendo en cuenta la presencia del resto de variables independientes. Además, el signo del coeficiente de la regresión parcial de una variable puede no ser el mismo que el del coeficiente de correlación simple entre esa variable y la dependiente, debido a los ajustes que se llevan a cabo para poder obtener la mejor ecuación posible. La tabla V.5 muestra el valor de los coeficientes obtenidos.

Tabla V-5. Relación de coeficientes asignados a las variables tras la aplicación de MLR.

	Coeficientes
Intercepción	0.468173556
Variable X_2	-0.027215577
Variable X_7	0.224961349
Variable X_{11}	0.737697493
Variable X_{12}	-0.330128951
Variable X_{13}	0.71210323
Variable X_{16}	0.185942371
Variable X_{18}	-0.14704751
Variable X_{21}	0.385804526
Variable X_{22}	1.308894181
Variable X_{23}	-0.406910718
Variable X_{27}	0.249521251

Cada uno de los coeficientes forma parte de la ecuación en puntuaciones directas, siendo el modelo obtenido para MLR el que se ilustra en la ecuación V.2

$$y = 0.47 - 0.03x_2 + 0.22x_7 + 0.74x_{11} - 0.33x_{12} + 0.71x_{13} + 0.19x_{16} - 0.15x_{18} + 0.39x_{21} + 1.31x_{22} - 0.41x_{23} + 0.25x_{27}$$

Ecuación V-2

Para verificar el ajuste y la significancia de la ecuación y los coeficientes obtenidos se emplearon el coeficiente de determinación de R^2 para el ajuste y las pruebas estadísticas de ANOVA y T para la significancia. Los resultados obtenidos para de los coeficientes de ajuste se expresan en la tabla V.6.

Tabla V-6. Coeficientes de ajuste del modelo MLR.

Estadísticas de la regresión	
Coeficiente de correlación múltiple R	0.86101771
Coeficiente de determinación R^2	0.83723161
R^2 ajustado	0.806526358
Error típico	0.07087346
Observaciones	500

El coeficiente de correlación múltiple *R* es el equivalente al coeficiente de Pearson, entre las variables involucradas, mientras que el coeficiente de determinación R^2 expresa la proporción de varianza de la variable dependiente que esta explicada por las diversas variables independientes. Por su parte R^2 ajustado es una corrección a la baja de R^2 basado en el número de casos y variables independientes. Finalmente el error típico de estimación es la desviación típica de los residuos, es decir, la desviación típica de las distancias existentes entre las puntuaciones en la variable dependiente (y_i) y los pronósticos efectuados con las rectas de regresión ($\hat{y}_i$), aunque divididos entre (*n-2*), donde *n* son los casos considerados (Box, Hunter, & Hunter, 2008). *R* toma un valor relativamente alto (su máximo es 1); y el valor de R^2 indica que el 83.7% de la variación de la variable dependiente es explicada por el modelo. La tabla V.7 muestra los valores resultantes de los estadísticos Anova.

Tabla V-7 .Significancia de la Ecuación obtenida por MLR mediante el estadístico F.

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	17	4.9823	0.29338	11.5637	0.000583
Residuos	483	107.0916	0.22172		
Total	500	112.0739			

Los resultados de la tabla V.7 proporcionaron la existencia o inexistencia de una relación significativa entre las variables. El estadístico F permite contrastar la hipótesis nula de que el valor poblacional de R es cero, y por tanto permite decidir si existe relación lineal significativa entre la variable dependiente y el conjunto de variables independientes tomadas todas juntas. El valor del nivel crítico de F =0.00058, indica una relación lineal significativa dado que todo valor por debajo de 0.001 (Box, et al., 2008), indica que si existe una relación lineal significativa. Así se pudo afirmar, por tanto, la ecuación de regresión obtiene un buen ajuste a la nube de puntos.

El valor del estadístico T se empleó para contrastar las hipótesis nulas de que los coeficientes de regresión valen cero en la población. Estos estadísticos se distribuyen según el modelo de probabilidad *t de Studen*t con *n-2* grados de libertad. Por tanto, pueden ser utilizados para decidir si un determinado coeficiente de regresión es significativamente distinto de cero y, en consecuencia, si la variable independiente está significativamente relacionada con la pendiente. Se puede observar que las variables empleadas alcanzaron coeficientes significativamente distintos de cero. Por tanto, todas ellas contribuyen de forma significativa a explicar lo que ocurre en la variable dependiente. En la tabla V.8 se muestran los valores resultantes de la aplicación del estadístico T.

Tabla V-8 .Significancia de los coeficientes obtenidos por MLR mediante el estadístico T.

	Estadístico t	*Probabilidad*
Intercepción	0.472292	0.6402
Variable X_2	-4.160003	0.0095
Variable X_7	6.628418	0.00534
Variable X_{11}	11.11274	0.00274
Variable X_{12}	-8.74831	0.0046
Variable X_{13}	11.440777	0.0016
Variable X_{16}	5.335102	0.0073
Variable X_{18}	-3.284106	0.0077
Variable X_{21}	6.587999	0.0056
Variable X_{22}	7.175794	0.0024
Variable X_{23}	-10.090579	0.0028
Variable X_{27}	11.41113	0.0016

La información de la ecuación V.2 resultó bien evaluada por los procesos de ajuste y significancia por lo que se aceptó dicho modelo como válido. Empleando un valor crítico de 0.5, empleando los 1000 clientes de prueba previamente establecidos para definir como buenos prestatarios aquellos por encima del valor crítico y como malos prestatarios los que obtuvieron una puntuación por debajo del valor crítico, se generó la matriz de clasificación definida en la tabla II.1, obteniéndose como resultado la tabla V.9 que se ilustra a continuación.

Tabla V-9. Matriz de clasificación para MLR.

Modelo	Real	
	1	0
1	146 (29.2%)	92 (18.4%)
0	74 (14.8%)	188 (37.6%)

La confiabilidad del modelo se obtuvo empleando la ecuación IV.27 con el resultado mostrado a continuación.

$$CP = \frac{146 + 188}{146 + 92 + 74 + 188} = 0.668$$

Así, la confiabilidad del modelo de regresión lineal múltiple se estableció en 66.8% con la información que se empleó en este caso.

V.2.2 Modelo LR.

La ecuación III.2 se consideró como el punto de partida para el desarrollo del modelo LR, el cual se implementó a través de la técnica de paso a paso (stepwise por su término en inglés), en donde a cada iteración o paso el modelo va ajustando el número de variables que son agregadas, siendo para este caso los valores de los coeficientes β_i obtenidos, los que se muestran en la tabla V.10, debiéndose tomar en cuenta que este modelo no es una regresión lineal pura, sino en una escala logarítmica, es decir la diferencia de probabilidad de que ocurra un suceso respecto de que no ocurra guarda una relación lineal logarítmica.

Tabla V-10. Relación de coeficientes asignados a las variables tras la aplicación de LR.

	Coeficientes
Intercepción	-7.0469
Variable X_2	1.0010
Variable X_7	2.0457
Variable X_{11}	-0.9555
Variable X_{12}	-0.6083
Variable X_{13}	3.6459
Variable X_{16}	0.9935
Variable X_{18}	4.2394
Variable X_{21}	5.7981
Variable X_{22}	-3.8255
Variable X_{23}	0.00031
Variable X_{27}	-0.3369

Así, la interpretación de los coeficientes de regresión depende tanto del valor de la variable x_i donde se produzca el incremento como del valor del resto de las variables, debido a que la pendiente de la curva de regresión cambia. Además, la forma de codificación de las variables como los casos en que son continuas, o bien catalogadas con valores como 0 o 1, es otro aspecto apreciado para indicar la ausencia o presencia de una determinada característica. Para ayudar a interpretar los coeficientes de LR se ha definido el coeficiente de probabilidades (conocido como Odds Ratio) entre que ocurra un suceso con respecto a de que no ocurra, como lo muestra la ecuación V.3.

$$Odd\ Ratio = \frac{P(Y=1)}{P(y=0)} = \frac{P}{1-P} \qquad \textbf{Ecuación V-3}$$

Siendo P la probabilidad del suceso definido en el modelo, en este caso el que el cliente sea un mal prestatario.

Se puede observar que el valor del coeficiente de la x_{23} resultó muy cercana a 0, lo que indica que la variable no afecta de manera sustancial la ocurrencia del suceso, por tal motivo no se consideró en la construcción del modelo. Así mismo, un coeficiente negativo indica que a medida que el valor de la variable va aumentando, el logaritmo del cociente de probabilidades va a ir disminuyendo y al revés si es positivo. El modelo alcanzado con la aplicación de LR se muestra en la ecuación V.4

$$y = -7.05 + x_2 + 2.06x_7 - 0.96x_{11} - 0.61x_{12} + 3.65x_{13} + 0.99x_{16} - 4.24x_{18} + 5.8x_{21} - 3.83x_{22} - 0.34x_{27}$$

Ecuación V-4

En el modelo LR, el contraste de regresión no se realizó sobre la descomposición de la suma de cuadrados como se llevó a cabo en la regresión lineal sino sobre el incremento de la verosimilitud (conocida como likelihood, por su término en inglés), más exactamente sobre la disminución de -2LL (-2 log likelihood), una relación logarítmica con la que se adecuo su puntuación. La construcción del contraste se definió acorde a la ecuación V.5.

$$C2LL = -2LL(b_0) - (-2LL(b_0, b_1, b_2, \dots, b_k))$$

Ecuación V-5

La diferencia de verosimilitudes se distribuye de acuerdo con una distribución χ_j^2, donde j es la diferencia del número de parámetros en el modelo (Box, et al., 2008). La tabla V.11 muestra la verosimilitud lograda en la última iteración del proceso paso a paso.

Tabla V-11 . Muestra de la iteración para el cálculo de contraste del método.

Iteración		*-2LL*
Paso 12	1	55.6371
	2	40.2992
	3	39.0415
	4	38.9618
	5	38.9613
	6	38.9613

Los resultados arrojaron que el algoritmo terminó correctamente porque se logró el criterio de parada, es decir, un cambio de todos los coeficientes estimados inferior al 0.001, lo cual se interpretó como que la significancia de los coeficientes resulta adecuada para el modelo.

El modelo LR no emplea el coeficiente de determinación *R* para mostrar la bondad del ajuste, sino que calcula el incremento de la verosimilitud, pese a lo cual reciben el nombre de R^2 aunque el significado geométrico generado no es el mismo que en la regresión lineal. La bondad de ajuste que se consideró fue la prueba de Hosmer y Lemeshow que es una constante de distribución χ^2 y cuyos resultados generales del modelo se muestran en la tabla V.12.

Tabla V-12. Resultados Generales de la Prueba de Homer y Lemeshow.

Chi-cuadrada	*Grados de Libertad*	*Significancia*
0.7667	2	0.6819

La significancia de que obtuvo el modelo de 0.7667, con una probabilidad del 68.2% indicó que el modelo no tuvo falta de ajuste en la regresión, no teniéndose diferencias sustanciales entre los valores observados y los valores pronosticados, por lo que se aceptaron los resultados del modelo entregados por la ecuación V.3. Para el modelo LR de pronóstico se fijó nuevamente en 0.5, de donde se desprendieron los resultados mostrados en la matriz de clasificación de la tabla V.13.

Tabla V-13. Matriz de clasificación para LR.

Modelo	Real	
	1	0
1	208(41.6%)	49 (9.8%)
0	96 (19.2%)	147 (29.4%)

La confiabilidad del modelo después de aplicar la ecuación IV.27 se estableció en un 71.0% según se muestra a continuación.

$$CP = \frac{208 + 147}{208 + 49 + 96 + 147} = 0.71$$

Con los resultados obtenidos se realizó el proceso de ensamble hibridado con PSO como se describe en la sub-sección siguiente.

V.2.3 Cálculo Ensamblado de la Puntuación Crediticia.

La combinación de los modelos MLR y LR y la utilización de la ecuación IV.25 generaron los pesos de aporte de cada uno de los modelos, a través del algoritmo PSO mejorado de la ecuación IV.5 como ajuste de velocidad y IV.2 para definir el desplazamiento de la partícula en uso. Los parámetros aplicados al algoritmo después de calibrarse el modelo se muestran en la tabla V.14.

Tabla V-14. Parámetros de configuración de PSO para estimación de pesos.

Parámetro	*Valor*
Tamaño de Enjambre	25
Grado de conocimiento individual (φ_1)	2
Grado de conocimiento grupal (φ_2)	2
Coeficiente de aceleración β_1	~(0,1)
Coeficiente de aceleración β_2	~(0,1)
Peso inercial *w* para aplicar PSO mejorado	[0.9->0.6]
Condición de paro	70 iter.
Cantidad de simulaciones	20

El peso inercial de la ecuación IV.5 se adecuó para que tuviese un decrecimiento lineal de 0.9 a 0.6 según se definió en la ecuación IV.26, para lograr un mejor desempeño del algoritmo. La función de aptitud del modelo representada en la ecuación IV.29 minimizó el comportamiento del algoritmo en el intento de disminuir el error Tipo II del modelo, en donde se definieron los siguientes valores de sus parámetros:

- n_1 y n_2 que representa el valor de prestatarios no deudores y deudores respectivamente tomaron el valor de 250 cada uno, de los 500 datos de muestreo adicionales a los empleados por LR.
- *M* se fijó en 100, para poder percibir el cambio en el valor de aptitud de la partícula.
- *K* después del proceso de calibración se graduó en un valor de 11.

Los valores pronosticados de MLR y LR, se tomaron como el vector de entrada del algoritmo PSO, así que las partículas resultantes son pequeñas como lo ilustra la Figura V.5, lo que justificó el uso de un enjambre de tamaño 25 en la ejecución del algoritmo. La información referente a los procesos de calibración de parámetros se puede referenciar en el Anexo B.

Pronóstico de MLR	*Pronóstico de LR*
0.635	0.747

Figura V-5. Representación de una partícula del modelo ensamblado de puntuación crediticia.

Al finalizar la aplicación del modelo ensamblado de puntuación crediticia propuesto, los pesos óptimos w_1 y w_2 encontrados por PSO, y asignados a los modelos simples LR y MLR respectivamente fueron los siguientes:

$$w_1 = 1.69457, \qquad w_2 = -0.69457$$

La tabla V.15 muestra los resultados arrojados para cada una de las 20 repeticiones realizadas sobre el algoritmo PSO para establecer los pesos en el proceso de ensamblado. Por su parte en la tabla C6 del anexo C muestra el código en Java del proceso de puntuación crediticia ensamblada.

Así el modelo ensamblado se definió a partir de la ecuación IV.25 como:

$$f = 1.69475\hat{y}_1 - 0.69475\hat{y}_2$$

Al igual que en los modelos simples se fijó el valor crítico de 0.5, para poder pronosticar los buenos y malos prestatarios, calificándose los 1000 prestatarios de prueba, con los resultados obtenidos se construyó la matriz de clasificación, obteniéndose la tabla V.15.

Tabla V-15. Matriz de clasificación para el modelo Ensamblado.

	Real	
Modelo	1	0
1	410(41.0%)	88 (8.8%)
0	124 (12.4%)	378 (37.8%)

La confiabilidad resultante del modelo, obtenido a partir de la ecuación IV.27 es del 78.8%, de acuerdo al cálculo realizado.

Con los datos obtenidos se calculó la curva ROC del modelo que se ilustra en la Figura V.6, para validar su desempeño, mediante la sensibilidad, la especificidad y el valor predictivo positivo de la tabla II.2, aplicando las funciones correspondientes con lo que se alcanzaron los valores que se muestran a continuación:

$$Se = \frac{410}{(410+124)} = 0.7667, \qquad Ep = \frac{378}{(378+88)} = 0.8116, \qquad VPP = \frac{410}{(410+88)} = 0.8239$$

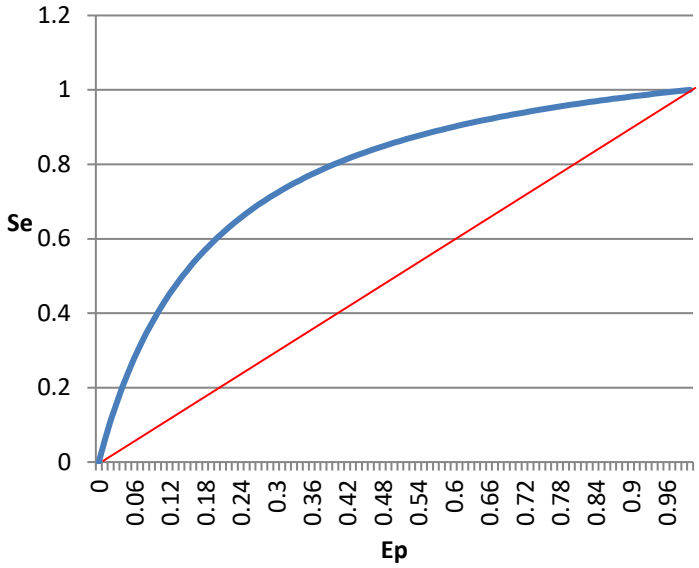


Figura V-6. Curva ROC del modelo propuesto de puntuación crediticia.

La curva ROC mostró una tendencia más cercana a 1 del área bajo la curva, o bien un acercamiento al borde superior izquierdo de la gráfica, recordando que estas situaciones indican un desempeño más apropiado del modelo, por lo que también quedó por este medio validado el modelo de puntuación crediticia propuesto.

La última variante de la propuesta de este trabajo se basó en el establecimiento de agrupamientos para el cálculo de puntuación crediticia de los clientes.

V. 3 Agrupamiento de Prestatarios

El agrupamiento de prestatarios se llevó a cabo mediante el empleo de los algoritmos de PCA, PSO y k-medias de una manera hibrida, según se especificó en el capítulo IV. El proceso de hibridación se estableció para solventar los problemas que se presentan durante el proceso de agrupamiento, como lo son la alta dimensión de los datos, la múltiple colinealidad y la densidad de los datos empleados, así como la velocidad de convergencia hacia la definición del los puntos centrales de cada uno de los agrupamientos.

La problemática que se presenta con los efectos de la alta dimensión, la colinealidad y la densidad de los datos durante los procesos de agrupamiento, se redujo con el empleo de la

matriz de puntaje obtenida del empleo de PCA. Así, los resultados arrojados por la aplicación de PCA en la sección V.1 se tomaron como punto de partida para la realización del agrupamiento de clientes. PCA generó una reducción del espacio dimensional del problema a solo 11 variables dependientes definidos en 12 componentes principales de los cuales solo se eligieron 4 por el criterio de varianzas acumulada. Los cuatro componentes seleccionados mostraron la existencia de 2 clases de prestatarios, por lo que se definió la cantidad de 4 centroides de agrupamiento para la aplicación del algoritmo de k-medias. La cantidad de prestatarios empleada para el desarrollo de los agrupamientos fue de 2067, es decir, se redujo el número de casos empleando el estadístico F para optimizar el proceso en el cálculo de agrupamientos.

Los resultados producidos tras la aplicación de PCA se emplearon como información inicial del algoritmo PSO, a través de una estrategia hibridada donde de acuerdo a la ecuación IV.32, una partícula representa los *k* centroides de los agrupamientos, siendo la idea de la aplicación de PSO la obtención óptima de la posición de dichos valores y con ello se llevó a cabo el proceso de agrupamiento, la Figura V.7 muestra un ejemplo típico de una partícula del planteamiento propuesto.

Centroide primer agrupamiento	Centroide segundo agrupamiento	*Centroide tercer agrupamiento*	*Centroide cuarto agrupamiento*
6.38293528	1. 8714848	2. 52781311	4. 33619086

Figura V-7. Representación de una partícula para la aproximación de centroides en el modelo de agrupamientos propuesto.

Los centroides se definieron inicialmente de forma aleatoria y el formato de la partícula establecido se aplicó el algoritmo PSO considerando la ecuación IV.5 para el cálculo de la velocidad de la partícula y IV.26 para el ajuste del peso inercial de la misma, el cual se hizo variar en un rango entre 0.8 y 0.6 después de los ajustes correspondientes, además del uso de la ecuación IV.3 para el establecimiento de la posición. Los parámetros estimados para la aplicación de las funciones de PSO se ilustran en la tabla V.16. Los procedimientos para el cálculo de los parámetros se pueden consultar en el Anexo B.

Tabla V-16. Parámetros de configuración de PSO para estimación de pesos.

Parámetro	*Valor*
Tamaño de Enjambre	40
Grado de conocimiento individual (φ_1)	1.47
Grado de conocimiento grupal (φ_2)	1.47
Coeficiente de aceleración β_1	˜(0,1)
Coeficiente de aceleración β_2	˜(0,1)
Peso inercial *w* para aplicar PSO mejorado	[0.8->0.6]
Condición de paro	1000 iter.
Cantidad de repeticiones	25

Con la condición de paro fijada en 1000 iteraciones para el proceso completo de agrupamiento, se uso el promedio del intervalo de confianza de 25 repeticiones en la obtención de los centroides que sirvieron para la implementación hibridada en el proceso de K-medias. La Figura V.8 muestra los valores obtenidos.

Centroide primer agrupamiento	Centroide segundo agrupamiento	*Centroide tercer agrupamiento*	*Centroide cuarto agrupamiento*
2.65893528	7.4787148	3.13115278	4.90863361

Figura V-8. Representación los valores de la partícula promedio para la fijación de centroides en el proceso hibridado de K-medias.

Con los centroides obtenidos por PSO, lo que sustituyó el paso inicial del algoritmo estándar K-medias, consistente en inicializar los centroides de manera aleatoria, con los centroides se agruparon los datos a través de la distancia de Manhattan reduciendo la distancia de los prestatarios pertenecientes al mismo grupo y se obtuvieron los nuevos centroides que definieran de mejor forma el grupo de centro de masa que es quien representa los puntos de cada grupo. En la tabla C7 del anexo C se presenta el código en Java del proceso de asignación de prestatarios a los agrupamientos.

El criterio de paro del algoritmo K-medias se fijó en 60 iteraciones que en promedio fue el número de repeticiones en los que el cambio en los centroides de los agrupamientos fue imperceptible y no se presentaron cambios en las membrecías de los mismos agrupamientos.

Una muestra de la evolución desarrollada por los agrupamientos se muestra en la Figura V.9, en donde la representación no se realiza con respecto a dos variables como ocurre en la mayoría de la literatura, pues se cuenta con 11 dimensiones, sino en base a los valores de la función de aptitud.

Figura V-9. Fases del desarrollo de los agrupamientos. (a) muestra la distribución de los prestatarios antes de iniciar el proceso de agrupamientos, (b) se muestran la distribucion de la información en la iteración 15 del proceso K-medias, (c) corresponde a la iteración 25, (d) muestra la información de los agrupamientos después de 50 repeticiones.

El desempeño del modelo de agrupamiento empleado se evaluó a través de la tasa de error, el criterio del error cuadrático medio además de las distancias dentro de los agrupamientos y entre los agrupamientos. La tabla V.17 muestra los resultados obtenidos por el modelo propuesto.

Tabla V-17. Resultados de validación del proceso de agrupamiento.

Tasa de Error	*Error Cuadrático Medio*	*Distancia intra Agrupamiento*	*Distancia inter Agrupamiento*
9.8%	*3.25*	*0.89±0.206*	*2.25±0.437*

Con los agrupamientos generados se realizaron los cálculos de disancia de cada prestatario hacia los 4 distintos agrupamientos generados para realizar el prónostico de puntuación creditica propuesta para agrupamientos. La tabla V.18 muestra la matriz de clasificación generada a partir de los resultados correspondientes.

Tabla V-18. Matriz de clasificación para modelo de agrupamiento.

Modelo	Real	
	1	0
1	819 (39.6%)	171 (8.3%)
0	242 (11.7%)	835 (40.4%)

La confiabilidad del modelo se obtuvo empleando la ecuación IV.27 con el resultado mostrado a continuación.

$$CP = \frac{819 + 835}{819 + 835 + 242 + 171} = 0.800$$

La tabla V.19 muestra un resumen del pronóstico de resultados obtenidos para los errores Tipo I y Tipo II de los modelos simples y el modelo ensamblado. Los valores se obtuvieron del cálculo del índice de falso negativo (α) e índice de falso positivo (β) ilustrados en la Tabla II.2, así como de los valores de confiabilidad realizada de manera independiente para cada uno de ellos.

Tabla V-19. Pronóstico de errores Tipo I, Tipo II y Confiabilidad con los modelos empleados.

Modelos	*Valores pronosticados con los datos de prueba*		
	Error Tipo I (α)	*Error Tipo II (β)*	*Confiabilidad*
MLR	20.97%	25.75%	66.80%
LR	31.58%	25.00%	71.00%
Ensamblado	20.97%	18.84%	78.80%
Agrupamiento	22.80%	16.99%	80.01%

La Figura V.10 muestra una comparativa entre los valores resultantes en cuanto a los valores resultantes en las matrices de clasificación de cada uno de los modelos empleados (VP, VN, FN y FP).

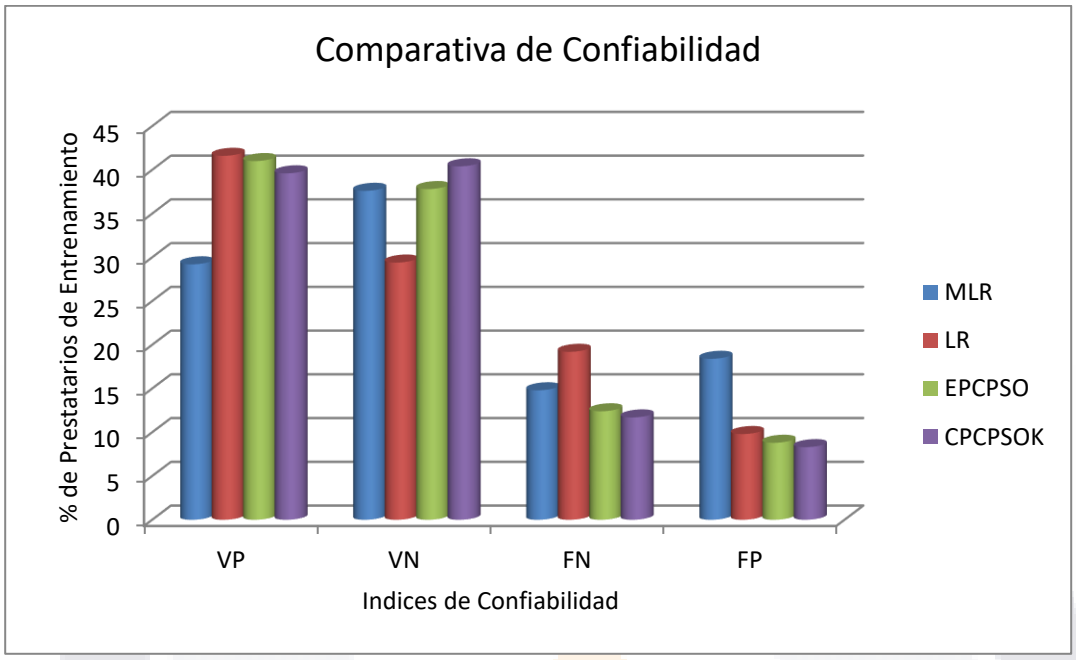


Figura V-10. Índices de confiabilidad entre los modelos de clasificación empleados.

Los resultados arrojan una ventaja considerable de los modelos propuestos con respecto a los modelos tradicionales empleados en el sector financiero.

Con los resultados obtenidos para cada una de las fases del trabajo propuesto en el capítulo VI se realiza una análisis y discución de dichos resultados que permitan derivar las conclusiones del trabajo desarrollado.

VI

Capítulo VI Análisis y Discusión

En este capítulo se discuten las ventajas y desventajas de los modelos formulados y se desarrolla una síntesis de los experimentos realizados con la intención de alcanzar un mejor entendimiento de cómo se desempeñan las propuestas en el apoyo de las actividades de las IMFs. La discusión del trabajo realizado se inicia con la justificación del empleo de PSO como heurística de optimización hibridada o ensamblada en las diferentes fases del trabajo.

VI.1 Justificación de uso de PSO

Es importante resaltar el uso del algoritmo PSO en las tres fases de aplicación de la tesis con la idea de presentar el rendimiento del método, aún en problemas a los que no está orientado, pero que en combinación de otras técnicas a partir de la hibridación o del ensamblado puede obtener resultados favorables, teniéndose en cuenta la simplicidad de implementación, pues solo requiere de una ecuación para actualizar las soluciones a diferencia de otras metaheurísticas donde se requiere métodos de representación, mutación, cruza, especiación y selección. Otras ventajas de PSO se enlistan a continuación:

1. PSO se basa en la inteligencia. Se puede aplicar en entornos de investigación científica y usos ingenieriles.
2. PSO no requiere de información adicional. La búsqueda se realiza a través de la velocidad de la partícula. Durante el desarrollo de varias generaciones, solo la partícula más óptima puede transmitir información al resto de las partículas, lográndose una alta velocidad de búsqueda.
3. El cálculo de PSO es muy simple. Comparado con el cálculo de otras técnicas, emplea su mayor capacidad de optimización alcanzando el valor óptimo fácilmente.
4. PSO adopta el valor real codificado directamente de la solución. La dimensionalidad empleada es constante para toda la solución.

5. PSO cuenta con una complejidad computacional baja. Tiene un buen desempeño en espacios de búsqueda grandes, pues converge rápidamente.

Evidentemente PSO no es un algoritmo que pueda enfrentar todo tipo de problemas y por lo tanto presenta algunas desventajas con respecto a algunas otras metodologías desarrolladas, las más significativas se mencionan a continuación:

1. El método fácilmente sufre de optimización parcial, provocando menor exactitud en la regulación de la velocidad y dirección adecuada.
2. PSO no puede enfrentar problemas de optimización dispersa.
3. El método no puede solucionar problemas de sistemas no coordinados, como la solución en el área energética.

El uso de PSO se justifica debido a las bondades que maneja el algoritmo, tratando de suplir sus deficiencias con el empleo de métodos alternos de manera combinada, la siguiente sección analiza el desempeño de la selección de variables. En este trabajo en particular de manera hibridada se logro la optimización de los siguientes procesos:

- Selección de variables mediante hibridación de PCA y PSO.
- Estimación de puntuación crediticia con ensamble de MLR y LR hibridado con PSO.
- Estimación de puntuación crediticia con hibridación de PCA, PSO y K-medias.

VI.2 Análisis de selección de variables

El pretratamiento de la información es la actividad inicial en todo proceso de minería de datos, que en términos promedio involucra un 75% del esfuerzo de un proyecto de minería, pero que es un gran soporte para la obtención de resultados confiables. El pretratamiento es una actividad que va más allá de la limpieza de los datos mediante el manejo de ruido, valores incompletos, casos extremos (en caso de ser necesario), etc., pues de igual forma involucra la transformación de los datos partiendo de sí éstos se tienen que entender como categóricos o puntuales. Cuando la variable tiene algún valor no numérico, será entendida automáticamente como categórica y todos sus valores como modalidades. Si todos sus

valores son números y algunos de ellos tienen decimales será entendida como continua. Finalmente si los valores son enteros se podrá elegir entre categórica y continua, siendo habitual definir variables categóricas en las que sus modalidades son representadas por enteros, por esta razón, es necesario que se pueda configurar el número máximo de enteros diferentes hasta el cual se entenderá tal variable como categórica por defecto. La selección de variables corresponde al paso final del pretratamiento de la información.

La selección de variables es una operación recomendable en todo trabajo de minería de datos debido a la alta densidad de información con la que se cuenta, en donde la gran cantidad de variables reflejan parte de la información general hasta cierto punto, pero existe una cierta correlación entre ellas, denominada colinealidad. La información estadística derivada del análisis factorial generalmente arroja superposición de la información. La superposición de datos redundantes acarrea la problemática de la dimensionalidad que es la razón de ser de la minería de datos. Por ello en esta tesis como lo muestra la literatura se reduce la dimensionalidad de los datos empleando PCA de forma hibridada con PSO, transformando los atributos de las variables originales en unos pocos atributos integrados. Es decir, de la combinación lineal de las X variables originales, resultan los Z componentes principales derivados de la aplicación de PCA, conservando la información principal de las variables y sin ninguna correlación existente entre ellas, evitándose así el problema de multicolinealidad entre las variables y con mejores atributos que las variables originales.

Con el filtrado de los datos como un paso de pre-procesamiento en la selección de variables, además de los beneficios ya mencionados, es que se alcanza una selección de variables genéricas facilita el empleo de PSO como seleccionador final de mayor complejidad y no lineal. Así, con los datos obtenidos de la aplicación de PCA se emplea el algoritmo PSO binario para explorar efectivamente el espacio de solución para el óptimo subconjunto de variables. La búsqueda heurística de PSO se ajusta iterativamente a través de la función de aptitud definida en términos de maximizar la separación entre las diferentes clases de variables conseguidas.

En general, existe una gran variedad de métodos para implementar la selección de variables, en este trabajo se emplean varios de ellos según lo muestra la tabla VI.1.

Tabla VI-1. Breve descripción de los métodos empleados en el proceso de selección de variables. Selección de Variables (SV), Pre-procesamiento (Pre) y Pos-procesamiento (POS).

Método	*Función*	*Descripción*
pc-extracción	SV	Extracción de variables con PCA
Pearson	SV	Clasificación de variables de acuerdo al coeficiente de correlación de Pearson
Filtro-Z	SV	Clasificación de variables de acuerdo a un filtro heurístico
normalización	Pre	Normalización de la información de los registros de la tabla de datos
estandarización	Pre	Estandarización de las variables
bias	Pos	Busca el mejor umbral para la salida de los clasificadores

La realización de una comparativa del modelo propuesto de selección de variables con estudios representados en la literatura, no es muy exacta pues los objetos de estudio son diferentes y no se tiene un punto de comparación bien definido.

La sección VI.3 realiza un razonamiento del modelo de puntuación crediticia a través del ensamblado de dos de las técnicas de puntuación más empleadas en los entornos financieros como lo son MLR y LR acorde a lo planteado en este trabajo.

VI.3 Análisis de puntuación crediticia

El modelo de puntuación crediticia emplea un método ensamblado de MLR y LR, en donde la intervención del algoritmo PSO tiene como finalidad calcular los pesos de contribución de cada uno de los dos métodos del ensamble.

La idea del ensamblaje es tratar de fomentar la diversidad. Este concepto de diversidad se puede ejemplificar claramente con el problema de asignación de crédito. Sí el comité de asignación en su conjunto hace una predicción errónea, ¿cuánto de este error se debe atribuir a cada miembro? Más concretamente, ¿qué grado de la confiabilidad de predicción del comité se debe a la precisión individual, y cuanto se debe a su interacción conjunta? Idealmente se debería tratar el error de ensamblado como dos componentes distintos: un término para la precisión de cada miembro del comité evaluador, más un término de sus interacciones, es decir, su diversidad. En el caso propuesto la clasificación de puntuación crediticia de MLR y LR, los cuales de manera conjunta pueden complementar un resultado más apropiado. El ensamblado se realiza mediante el método de Bagging (Bootstrap

Aggregating), que promueve la diversidad mediante la presentación de cada modelo base con un subconjunto diferente de ejemplos de entrenamiento o diferentes distribuciones de peso en los ejemplos. De esta forma, se ejecuta cada algoritmo de aprendizaje en diferentes subconjuntos de ejemplo, la formación puede producir muy diferentes clasificadores que pueden combinarse para producir un resultado eficaz. Las predicciones de los modelos base se combinan mediante el voto mayoritario que se emplea frecuentemente en los problemas de clasificación empleando Bagging y consiste en como su nombre lo indica asignar la solución a lo que la mayoría de los modelos empleados designa como se muestra en la figura VI.1.

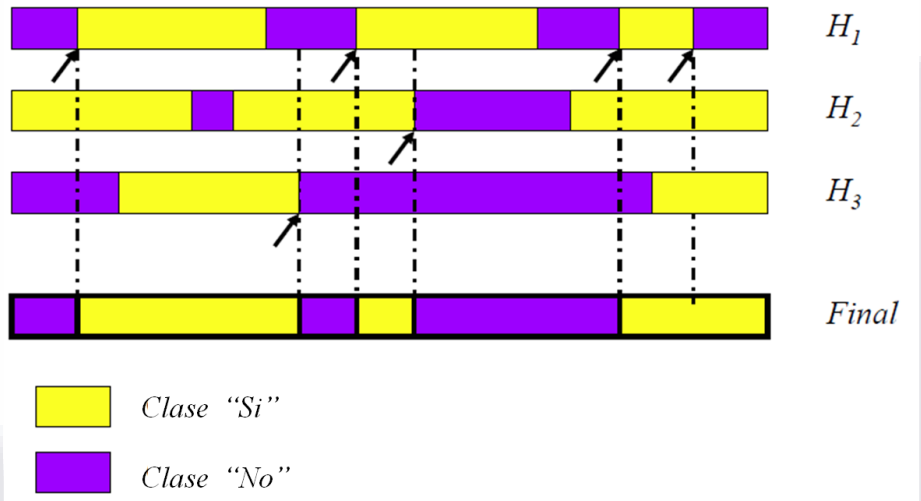


Figura VI-1. Ejemplo de funcionamiento de voto mayoritario.

En el proceso de voto mayoritario los clasificadores de entrada, pueden relacionarse de manera arbitraria, en donde el clasificador final simplemente clasifica los ejemplos de acuerdo con la mayoría de votos. En los casos donde se refina el proceso, los votos de los modelos se ponderan en función de la exactitud individual, como ocurre en el caso propuesto en este trabajo, donde esta refinación se lleva a cabo mediante el voto mayoritario calibrado a través de PSO.

La tabla VI.2 muestra un comparativo de los resultados obtenidos con los de otros trabajos realizados en puntuación crediticia, sobre una base de datos crediticia alemana, que es muy empleada para procesos comparativos (Yun, Qiu-yan, & Hua, 2011).

Tabla VI-2. Comparativa de puntuación crediticia con otras propuestas realizadas.

Modelo	*Error Tipo I (α)*	*Error Tipo II (β)*	*Confiabilidad Total*	AUC
Análisis Discriminante	32.51%	35.09%	65.91%	66.2
Árbol de Decisión	50.07%	21.45%	70.35%	64.24
Función de Base Radial	60.53%	13.31%	68.5%	62.88
Máquina de Soporte Vectorial (SVM)	72.37%	2.42%	71%	62.60
Perceptrón Multicapa (ANN)	44.74%	14.52%	74%	70.37
Ensamble con Redes Neuronales	73.57%	3.63%	74.67%	61.4
Ensamble con SVM	32.19%	11.92%	79.64%	76.39
Propuesta de Tesis (EPCPSO)	22.97%	18.84%	76.80%	67.83
Propuesta de Tesis (CPCPSPK)	23.80%	16.99%	78.01%	75.43

Los resultados comparativos muestran un desempeño competitivo del modelo propuesto con respecto a otras metodologías empleadas para afrontar el pronóstico de puntuación crediticia, pese a no ser una comparativa adecuada por haberse realizado con bases de datos distintas. En la siguiente sección se analizan los resultados de la fase de agrupamiento en este trabajo.

VI.4 Análisis de agrupamiento

El análisis de agrupamiento empleado combina las técnicas de los métodos multivariados de reducción y agrupamiento con el algoritmo de PSO. Los métodos de reducción o factoriales intentan disminuir la dimensión del análisis reduciendo el número de variables necesarias para considerar una cantidad menor que igualmente capte una gran proporción de la variabilidad de los datos, como ya se mencionó PCA es una técnica correspondiente a este tipo de modelos. Por otra parte los métodos de agrupamiento intentan formar grupos de variables sobre la base de las medidas disponibles, siendo un modelo representativo el algoritmo de agrupamiento de *K-medias*.

La intención de emplear los dos métodos multivariados es que se complementan entre sí, pues PCA como procedimiento de reducción se adapta de manera adecuada a la exploración de grandes tablas de datos individuales, resultando su producto sumamente útil para la investigación. Sin embargo, no es suficiente para formar una visión totalmente satisfactoria de la relación entre los datos. No sólo los resultados no se vinculan con una parte de la información, sino que ellos son a veces muy complejos como para interpretarlos

fácilmente. Ante estas circunstancias, las técnicas de agrupamiento pueden complementar y matizar los resultados de los algoritmos de reducción. El complemento entre los dos modelos multivariados concierne en la comprensión de la estructura de datos y la ayuda práctica en la fase de interpretación de los resultados. Al ser PCA un modelo factorial el uso de K-medias sobre el total del espacio o sobre un sub-espacio definido por los primeros vectores propios, como ya se mencionó son los más significativos. Los agrupamientos consideran la dimensión relativa del total de la información, el número de prestatarios considerados en este trabajo, corrigiendo ciertas deformaciones originadas a la operación de proyección. Además la creación de agrupamientos permite de manera natural la detección de anomalías que pueden ser descartadas, o pueden ser relevantes según la interpretación que se esté realizando.

Por contraparte, los agrupamientos no siempre tienen éxito al mostrar la importancia de ciertas tendencias o de factores latentes continuos. Para observar la organización especial de los grupos, el posicionamiento de éstos sobre los ejes factoriales resulta indispensable. Así, los agrupamientos descubren la existencia de grupos de individuos, y el proceso factorial pone en evidencia factores latentes no atendidos. El descubrimiento de tales fenómenos o dimensiones escondidas es el principal objetivo de estas dos metodologías, por lo tanto, su uso complementario resulta fundamental en la consecución de dicho objetivo.

Con la idea de enriquecer la complementariedad de los métodos multivariados se realiza la propuesta de emplear el algoritmo PSO como enlace entre la combinación de resultados de PCA con K-medias. Los datos obtenidos por PCA sirven como entrada de la ejecución de PSO y con ello implementar una mejora en los resultados de K-medias. La combinación de los resultados de PCA y PSO hibridados con K-medias propone una mejora en el desempeño de agrupamiento. PSO es un algoritmo de búsqueda global, el cual tiene la habilidad de encontrar un resultado óptimo global. Sin embargo, la velocidad de convergencia cerca de la solución es muy lenta. El algoritmo K-medias, por el contrario, converge rápidamente en un resultado de óptimo local, pero su habilidad para encontrar una solución global es pobre. Así, el proceso de agrupamiento se inicia con la aplicación de

PSO para intensificar la búsqueda a través de todo el espacio una solución global de los centroides de los agrupamientos, la cual se obtiene cuando el óptimo global comunitario de las partículas g_{best} permanece sin cambio para un número definido de iteraciones (15 en esta propuesta). Cuando la región global se localiza por PSO se continúa el proceso de agrupamiento empleando K-medias, tomando como centroides iníciales los valores generados por PSO, siendo el algoritmo K-medias el encargado de finalizar el proceso de agrupamiento después de 60 iteraciones.

El cotejo del modelo de agrupamiento planteado en este trabajo se presenta en la tabla VI.3, al igual que en las comparativas de las fases pasadas realizadas en las secciones anteriores, el alcance de la comparación puede verse reducido debido a que los datos analizados en cada caso son diferentes.

Tabla VI-3. Comparativa de agrupamiento con otras propuestas realizadas.

Modelo de agupamiento	*Base de Datos Empleada*	*Tasa de Error*	*Error Cuadrático Medio*	*Distancia intra Agrupamiento*	*Distancia inter Agrupamiento*
Propuesta	IMF mexicana	9.8%	*3.25*	*0.89±0.206*	*2.25±0.437*
K-medias	IMF mexicana	11.85%	*3.59*	*1.02±0.457*	*2.58±0.651*
K-medias	Iris	13.2%	*0.54*	*3.973±0.398*	*1.314±0.315*
PSO	Iris	12%	*0.52*	*3.306±0.202*	*0.857±0.096*
PSO-KM	Iris	10.5%	*0.52*	*3.262±0.216*	*0.918±0.083*

El análisis de los resultados obtenidos en las diferentes fases del modelo, así como su contrastación con los productos derivados de otras propuestas da la pauta para realizar las conclusiones de la propuesta en el capítulo VII.

VII

Capítulo VII Conclusiones

VII.1 Conclusiones de resultados.

En la actualidad se ha realizado bastante investigación en el área de puntuación crediticia, teniéndose varios modelos para implementar dicho proceso a través del uso de diferentes herramientas y técnicas, impulsados en mayor medida por las regulaciones internacionales acordadas por los acuerdos de Basilea, que obligan a toda institución financiera a contar con varias herramientas que le impidan caer en problemas de falta de liquidez que puedan propiciar complicaciones financieras como los que han sucedido en las últimas 2 décadas, empezando por el efecto Tequila ocasionado por el alto grado de cartera vencida en los bancos Mexicanos y pasando por crisis bancarias como la estadounidense, la griega y más recientemente la española y la chipriota que han provocado grandes resquebrajamientos en el entorno financiero internacional

Sin embargo, hasta hace muy pocos años la gran mayoría del desarrollo de modelos de puntuación crediticia ha estado dirigido a instituciones financieras bancarias de gran tamaño, quedando relegadas instituciones como las microfinancieras, que pese a su aparente tamaño pequeño en el entorno macroeconómico son de gran impacto en economías en desarrollo como la mexicana. Así, el principal objetivo de esta tesis ha sido estudiar y proponer métodos inteligentes alternativos que permitan a las IMFs una evaluación y medición del riesgo de impago, que permitan un desarrollo productivo de estas entidades.

El inicio de las tareas en esta tesis doctoral consistió en el análisis del estado del arte y la identificación de los principales métodos empleados por otros investigadores para enfrentar esta problemática. De este estudio inicial se desprendió la identificación de las principales propuestas, las cuales son consideradas técnicas de la minería de datos y que son las siguientes. De juicio que en términos generales son las que emplean la IMFs, las estadísticas que son a las que más recurren aún en la actualidad muchos de los sectores

financieros, y sobre todo el personal del área de finanzas, y finalmente los métodos no paramétricos que incluyen métodos de investigación de operaciones y de inteligencia artificial, métodos que parecen tener la solución para afrontar la cantidad de información que se manipula en el sector financiero.

Al ser un problema de minería de datos el primer obstáculo se presentó para conseguir la base de datos (dataset) con la información correspondiente a una IMF, pues es información que generalmente cualquier institución financiera maneja como confidencial. El uso de los dataset empleados para evaluación comparativa (benchmarking) por muchos investigadores como la base de datos financiera alemana, tampoco presentó una solución viable, ya que la información no es la misma que la manejada por las IMFs, se pudo comprobar lo que la literatura relata en cuanto a la pobreza y baja calidad de la información manejada por las IMF, adicionalmente no es fácil de conseguir pues se requiere el establecimiento de contactos personales que permitan el acceso a dicha información.

Una vez conseguido el dataset el proceso del desarrollo del proyecto se llevó a cabo mediante la implementación de la metodología CRISP-DM, por lo que se recurrió a la asesoría de los desarrolladores del sistema informático y entender en el mayor grado posible la situación de operación del sistema, así como la información que lo conforma. La preparación de los datos se realizó mediante la limpieza, transformación y selección de las variables. El proceso de selección de variables es por sí mismo un área de interés para la investigación, existiendo también mucho desarrollo y métodos para llevarlo a cabo. En particular en esta propuesta se llevó a cabo mediante la hibridación de PCA y PSO.

La aplicación de este modelo a partir del dataset consiste en seleccionar las variables que mejor se adecuan al establecimiento de un modelo de puntuación crediticia, buscando lograr un error bajo de clasificación. Habiéndose probado solamente con los datos analizados en esta propuesta, el modelo parece adecuado para adaptarse a una gran variedad de conjuntos de datos debido a que no se requiere un conocimiento del dominio ni tampoco conocimiento de aprendizaje de máquina. La simplicidad del PSO y su rendimiento comprobado, comparable a la de los algoritmos evolutivos, sitúan a PSO muy adecuado para la selección de variables. Aún cuando PSO es un método de búsqueda global, la

aplicación de PCA permite acelerar su convergencia en problemas de alta dimensionalidad, logrando una reducción en el tiempo de complejidad. La hibridación de estos dos métodos es un paso hacia la inteligencia de enjambre topológica que tiene el potencial para utilizar una formulación mucho más enriquecida. El grado de reducción de variables alcanzado al emplear este modelo así como la confiabilidad de una representatividad adecuada de las variables seleccionadas en el entrono global permiten, además de las razones ya expuestas, considerarlo como buen algoritmo de selección de variables.

La aportación más considerable en este punto radica en que el modelo de selección variables implementada con PCA y PSO es frecuentemente empleada en la selección de características de imágenes pero raramente en la selección de atributos de un dataset, por lo que aunque el tema sea el mismo la adecuación cambia en el aspecto del manejo de los valores de los datos, que es parte del trabajo que se desarrolló en este modelado.

La selección de variables resultante se emplea para implementar el modelo de un clasificador ensamblado para el análisis de puntuación crediticia. Debido a la incertidumbre encontrada en los dataset crediticios reales, los patrones generados clasifican erróneamente las muestras de entrada limitando generalmente la utilidad del clasificador construido, situación que se acentúa en el caso de las IMFs. Partiendo de la aplicación de un método de aprendizaje ensamblado de MLR y LR para mejorar el desempeño de clasificación, empleando Bagging y PSO para encontrar los pesos óptimos de los modelos ensamblados y aplicar un voto mayoritario ponderado controlando el grado de Error Tipo II mediante la función de aptitud. El ensamblado es una buena alternativa para tratar de aprovechar las ventajas de los diversos modelos ensamblados y solventar sus debilidades complementándose estos. Los resultados obtenidos se equiparan y superan a muchos de los alcanzados por otros modelos encontrados en la literatura, por lo que se puede considerar que el modelo propuesto tiene un desempeño significativamente mejor que el de la mayoría de los métodos convencionales de clasificación. Es importante resaltar la importancia de una estrategia útil de selección de variables, que sirve de base para las reglas de asociación de la minería de datos.

En resumen, la contribución de este modelo de puntuación radica en el uso de ensamblado e hibridación mediante MLR, LR y PSO, sin considerar el procedimiento de selección de variables, para permitir que los oficiales de crédito de las IMFs pueden tener la ayuda práctica en su tarea de aprobación de crédito diariamente con una confiabilidad alta, cercana a la de las mejores propuestas de puntuación desarrolladas para la banca comercial según se observa en los resultados obtenidos y con la gran ventaja de que los procedimientos y valores empleados para el cálculo de la puntuación son transparentes y de fácil interpretación a diferencia de métodos clasificadores puros como SVM y ANN.

Finalmente el desarrollo de agrupamientos mediante la hibridación de PCA, PSO y K-medias. El uso de PCA con PSO además de reducir la dimensionalidad no afecta el funcionamiento hibridado de PSO, es decir se puede hibridar PSO con algún otro método y el desempeño de PCA con PSO no se ve afectado, teniendo esto en cuenta a partir de la aplicación de PCA se hibrida PSO con K-medias, este método combinado tiene la ventaja de los dos modelos sin hacer inherentes sus desventajas. El algoritmo PSO realiza las búsquedas exitosas sobre todo el espacio, así que durante las fases iníciales del proceso hibridado, la búsqueda global se desarrolla con PSO. Conforme las partículas en el enjambre se empiezan a aproximar al óptimo global, el algoritmo cambia a K-medias que en óptimos locales converge más rápido que PSO, detectándose el punto adecuado de cambio entre un algoritmo y otro mediante la función de aptitud. Los resultados entregados demuestran que la propuesta tiene un desempeño superior a la mayoría de los modelos analizados en la literatura. Una vez calculados los agrupamientos se procede a la determinación de puntuación crediticia a través del empleo del modelo ensamblado propuesto de MLR, LR y PSO mediante una asignación ponderada de acuerdo al grado de pertenencia de todos los prestatarios de forma particular a cada uno de los agrupamientos definidos.

El aporte en el modelado de agrupamientos es sustancial, en primer lugar hay pocas referencias literarias que empleen los tres métodos combinados para la realización de agrupamientos, este procedimiento reduce la complejidad computacional de k-medias, debido a que se reduce la dimensionalidad del espacio muestral mediante el uso de PCA,

pero además se predefinen de manera óptima el número de agrupamientos y los valores aproximados de los centroides, que son 3 causales del incremento polinomial en algunos casos exponencial de k-medias. Adicionalmente en el proceso de puntuación crediticia la hibridación de todos los métodos empleados es una propuesta innovadora en su conjunto, que proporciona valores confiables, simples y comparables a las mejores propuestas desarrolladas en la literatura.

La tabla VII.I muestra los trabajos más relevantes desarrollados a partir de la realización de esta propuesta.

Tabla VII-1Productividad desarrollada a partir del trabajo doctoral.

Tema	Trabajo	Tipo
Minería de Datos	• Explaining Diverse Application Domains Analyzed from Data Mining Perspective. • New Implementations of Data Mining in a Plethora of Human Activities. • Analysis of Cyber-bullying in a virtual social networking.	• Capítulo de Libro. (Ochoa, et al., 2012) • Capítulo de Libro. (Ochoa, Ponce, Elias, Ornelas, et al., 2011) • Artículo en Journal. (Ochoa, Ponce, Elias, Jaramillo, et al., 2011)
Puntuación Crediticia	• Credit Scoring for Microfinance Institutions in México an Ensemble and Hybridized Approach	• Artículo en Journal. (Elias, et al., 2012)
Agrupamiento	• Outlier analysis for plastic card fraud detection a hybridized and multi-objective approach. • Algoritmo Inteligente para agrupamiento de Clientes en Microfinancieras • Credit Scoring Mechanisms for MFIs	• Artículo en Journal. (Elías, et al., 2011) • Cartel. (Elias & Padilla, 2012) • Ponencia. (Elias, et al., 2012)

Es significativo enfatizar la importancia del uso del algoritmo PSO en esta propuesta siendo su principal ventaja que solo necesita una sola ecuación para la actualización de soluciones, en comparación con los algoritmos evolutivos donde los métodos para la representación, la mutación, cruzado, especiación y selección tienen que ser considerados.

Finalmente, no queda más que establecer que los resultados aquí expuestos son particulares al trabajo desarrollado, sin embargo hay otros tipos de resultados que no se expresan y son

los referentes al enriquecimiento derivado en lo personal y en lo profesional que pese a quedar fuera del entorno es pertinente mencionarlo.

En la siguiente sección se exponen las conclusiones relacionadas con las preguntas de investigación, así como el tratamiento referente a las hipótesis planteadas para este trabajo.

VII.2 Conclusiones de preguntas de investigación.

El trabajo desarrollado permite poder responder de manera objetiva, cualitativa y cuantitativa las preguntas de investigación formuladas como punto original de esta propuesta, obteniéndose las conclusiones que se enuncian a continuación para cada una de las preguntas.

i. ¿Pueden lograr propuestas de inteligencia artificial en la puntuación crediticia un impacto similar en las instituciones microfinancieras al obtenido en instituciones de banca comercial y de las hipotecarias?

   Las propuestas de puntuación crediticia implementadas con modelos no paramétricos de inteligencia artificial pueden lograr un impacto semejante en las IMFs, pues los resultados obtenidos demuestran los valores de los parámetros de la matriz de confiabilidad de la tabla II.1 (VP, FP, FN y VN), así como a los índices de confiabilidad (ICC, IEC, Error tipo I, Error tipo II, Se, SPEF, VPP, VPN, Conf) y la misma AUC, son muy similares a los conseguidos por los sistemas comerciales y propietarios de las otras instituciones financieras. Con esto se acepta Hi1.

   Sin embargo, el éxito en la implementación cotidiana requiere de la implementación de tecnologías y logísticas de funcionamiento adecuados, de tal forma que los oficiales de crédito tomen como un punto a considerar al momento de su decisión los resultados entregados por los modelos de puntuación crediticia.

ii. ¿Resulta relevante el uso de PSO para asignación de los pesos del método de voto mayoritario en una propuesta ensamblada dentro del entorno de puntuación crediticia en IMFs?

   La existencia de varias técnicas metaheurísticas y de aprendizaje de máquina, sin que una de ellas sea considerada mejor en todas las condiciones que el resto, evidencia que el uso de PSO no garantiza un mejor resultado en la asignación de pesos para la

propuesta de ensamblado, no obstante, la relevancia de su aplicación en esta propuesta se deriva de las ventajas desde la perspectiva de simplicidad de implementación que presenta PSO con respecto a otras metaheurísticas, además de la poca aplicación que a nivel regional e incluso nacional ha tenido el uso de PSO según los consultado en la literatura, lo que brinda una oportunidad de desarrollo. Finalmente los resultados de confiabilidad obtenidos tras las pruebas desarrolladas del modelo ensamblado son competitivos con otras propuestas, según se observa en la tabla VI.2, así como en los demás modelos implementados, lo que justifica la relevancia del uso de PSO no solo en el proceso de ensamblaje, sino como parte central de todos los modelos desarrollados durante este trabajo.

iii. ¿Ante la dificultad de la recolección de información una propuesta semi-supervisada presenta mejores resultados de confiabilidad de pronóstico y reducción de error tipo II en la puntuación crediticia que una propuesta supervisada?

Los procesos de clasificación y predicción según lo establece la literatura son procedimientos netamente supervisados, sin embargo estos métodos suelen sufrir a menudo de la falta de suficientes datos de entrenamiento etiquetados. Una solución en apariencia inmediata se presenta al pensar en el uso de algunos datos etiquetados en conjunto con la mayoría de los datos etiquetados, entiéndase procedimientos semi-supervisados. En esta propuesta la idea de aplicar el enfoque semi-supervisado, más allá de proporcionar un apoyo a los métodos supervisados de clasificación es brindar una herramienta alterna de puntuación crediticia, tratando de obtener un mayor grado de confiabilidad, sin embargo con el afán de responder la pregunta de investigación se realizaron experimentos adicionales que para poder realizar la comparativa específica. Los cuales se pueden apreciar en el anexo D. Los resultados del experimento demuestran que el resultado obtenido por el modelo semi-supervisado es mejor a los modelos supervisados, sin embargo hay que tomar en cuenta que el modelo semi-supervisado es hibridado, lo que mejora su rendimiento y además fortalece la aportación planteada del modelo propuesto. El razonamiento expresado en esta respuesta justifica la aceptación de la hipótesis Hi2.

iv. ¿Puede alcanzar la propuesta una confiabilidad como la que logran las máquinas de soporte vectorial (SVM por sus siglas en inglés Support Vector Machines) y las redes neuronales (ANN por sus siglas en inglés Artificial Neural Networks)?

La respuesta a esta pregunta es negativa, los resultados obtenidos por el modelo no superan los alcanzados por propuestas implementadas por máquinas de soporte vectorial y redes neuronales, que son clasificadores por sí mismos, mientras que PSO es principalmente un optimizador. Sin embargo, queda abierta la propuesta para probar el modelo con PSO sobre los mismos dataset en los que se han ejecutado SVM y ANN, debido a que las correlaciones alcanzadas por los datos de la IMF analizada son bajas y esta situación empobrece el desempeño del modelo. Adicionalmente la propuesta realizada tiene la gran ventaja de que emplea MLR y LR que son técnicas que pueden interpretar perfectamente los oficiales de crédito, contra el concepto de caja negra que se emplea en SVM y ANN. La aceptación de la hipótesis alternativa Ha3 queda demostrada con esta respuesta.

VII.3 Trabajo Futuro.

El trabajo realizado hasta el momento de aceptación de este documento ha sido arduo y muy productivo, sin embargo puede perfeccionarse realizando las siguientes tareas:

i. Resulta imperativo realizar un "benchmark" a través del uso de los dataset reportados en la literatura que puedan validar no solo los resultados obtenidos de este trabajo, sino con otros trabajos realizados e ilustrados en la literatura y den mayor sustento a las aportaciones realizadas. En este punto debe considerar que los métodos aquí implementados están orientados al dataset obtenido de una IMF, cuyos campos y estructura es muy diferente a la de los dataset mencionados en la literatura, lo que puede afectar el comportamiento de la propuesta y variar sus resultados.

ii. Orientar el trabajo de agrupamientos no solo a la fijación de puntuación crediticia, sino al establecimiento de propuestas que permitan mayores oportunidades de negocio a las IMFs, a partir del descubrimiento de conocimiento que se genere del manejo de agrupamientos como actividad primordial de la minería de datos.

iii. Dar un enfoque multi-objetivo al planteamiento del problema, para que la propuesta coadyuve en tres de los puntos medulares de una institución financiera: otorgar créditos, generar ganancias y mantener a los clientes buenos.

iv. Obtener la calibración de los parámetros del algoritmo PSO a través de un diseño de experimentos estricto y formal, que apruebe de manera irrestricta los parámetros fijados.

v. Trabajar en nuevas aportaciones que cubran las áreas de interés personal.

BIBLIOGRAFÍA

Bibliografía

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications, 35*(3), 1275-1292.

Adnan, K. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications, 37*(9), 6233-6239.

Adnan, K. (2011). Credit risk evaluation using neural networks: Emotional versus conventional models. *Applied Soft Computing, 11*(8), 5477-5484.

Ahmadyfard, A., & Modares, H. (2008, 27-28 Aug. 2008). *Combining PSO and k-means to enhance data clustering.* Paper presented at the Telecommunications, 2008. IST 2008. International Symposium on.

Ahn, C. W., An, J., & Yoo, J.-C. (2010). Estimation of particle swarm distribution algorithms: Combining the benefits of PSO and EDAs. [In Press]. *Information Sciences*(0).

Angeline, P. (1998). Evolutionary optimization versus particle swarm optimization: Philosophy and performance differences. In V. Porto, N. Saravanan, D. Waagen & A. Eiben (Eds.), *Evolutionary Programming VII* (Vol. 1447, pp. 601-610): Springer Berlin / Heidelberg.

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance, 48*(4), 733-755.

Antonakis, A. C., & Sfakianakis, M. E. (2009). Assessing naïve Bayes as a method for screening credit applicants. *Journal of Applied Statistics, 36*(5), 537-545.

Baesens, B., Egmont-Petersen, M., Castelo, R., & Vanthienen, J. (2002). *Learning Bayesian Network Classifiers for Credit Scoring Using Markov Chain Monte Carlo Search*. Paper presented at the Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3 - Volume 3.

Bahrammirzaee, A., Ghatari, A., Ahmadi, P., & Madani, K. (2011). Hybrid credit ranking intelligent system using expert system and artificial neural networks. *Applied Intelligence, 34*(1), 28-46.

Banks, A., Vincent, J., & Anyakoha, C. (2007). A review of particle swarm optimization. Part I: background and development. *Natural Computing, 6*(4), 467-484.

International convergence of capital measurement and capital standards – a revised framework (2005, comprehensive version 2006).

International framework for liquidity risk measurement, standards and monitoring (2010).

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2, Part 2), 3302-3308.

Bengoetxea, E., & Larrañaga, P. (2010). EDA-PSO: A Hybrid Paradigm Combining Estimation of Distribution Algorithms and Particle Swarm Optimization. In M. Dorigo, M. Birattari, G. Di Caro, R. Doursat, A. Engelbrecht, D. Floreano, L. Gambardella, R. Groß, E. Sahin, H. Sayama & T. Stützle (Eds.), *Swarm Intelligence* (Vol. 6234, pp. 416-423): Springer Berlin / Heidelberg.

Blackwell, T. (2007). Particle Swarm Optimization in Dynamic Environments. In S. Yang, Y.-S. Ong & Y. Jin (Eds.), *Evolutionary Computation in Dynamic and Uncertain Environments* (Vol. 51, pp. 29-49): Springer Berlin / Heidelberg.

Blackwell, T., Branke, J., & Li, X. (2008). Particle Swarms for Dynamic Optimization Problems Swarm Intelligence. In C. Blum & D. Merkle (Eds.), (pp. 193-217): Springer Berlin Heidelberg.

Blum, C., & Li, X. (2008). Swarm Intelligence in Optimization

Swarm Intelligence. In C. Blum & D. Merkle (Eds.), (pp. 43-85): Springer Berlin Heidelberg.

Boguslauskas, V., & Mileri, R. (2009). Estimation of Credit Risk by Artificial Neural Networks Models. *Economics of Engineering Decisions 4*, 7-14.

Box, G. E. P., Hunter, S., & Hunter, W. G. (2008). *Estadistica para investigadores/ Statistics for Investigators: Diseno, innovacion y descubrimiento/ Design, Innovation and Discovery*: Reverte Editorial Sa.

Boyle, M., Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). Methods for credit scoring applied to slow payers. In L. C. Thomas, J. N. Crook & D. B. Edelman (Eds.), *Credit Scoring and Credit Control* (pp. 75-90): Clarendon Press.

Breiman, L. (2001). Random Forests. . *Machine Learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*: Wadsworth International Group.

Capotorti, A., & Barbanera, E. Credit scoring analysis using a fuzzy probabilistic rough set model. [In Press, Corrected proof]. *Computational Statistics & Data Analysis*(0 doi: 10.1016/j.csda.2011.06.036).

Carlisle, A., & Dozier, G. (2000, 13-16 Jul). *Adapting Particle Swarm Optimization to Dynamic Environments.* Paper presented at the The 2000 International Conference on Artificial Intelligence (ICAI'2000) on, Las Vegas, NE.

Carter, C., & Catlett, J. (1987). Assessing Credit Card Applications Using Machine Learning. *IEEE Expert Systems, 2*(3), 71-79.

Clerc, M., & Kennedy, J. (2002). The particle swarm - explosion, stability, and convergence in a multidimensional complex space *Evolutionary Computation, IEEE Transactions on, 6*(1), 58-73.

Coffman, J. Y. (1986). *The proper role of tree analysis in forecasting the risk behaviour of borrowers*.

Colorni, A., Dorigo, M., & Maniezzo, V. (1991). *Distributed Optimization by Ant Colonies.* Paper presented at the Artificial Life European Conference on, Paris, France.

Conde Bonfil, C. (2000). *Pueden ahorrar los pobres? : ONG y proyectos gubernamentales en México*. Zinacantepec, Estado de México: El Colegio Mexiquense : Unión de Esfuerzos para el Campo La Colmena Milenaria.

Consoli, S., Moreno-Pérez, J., Darby-Dowman, K., & Mladenović, N. (2010). Discrete Particle Swarm Optimization for the minimum labelling Steiner tree problem. *Natural Computing, 9*(1), 29-46.

Cui, X., Charles, J., & Potok, T. (2009). A Simple Distributed Particle Swarm Optimization for Dynamic and Noisy Environments. In N. Krasnogor, M. Melián-Batista, J. Pérez, J. Moreno-Vega & D. Pelta (Eds.), *Nature Inspired Cooperative Strategies for Optimization (NICSO 2008)* (Vol. 236, pp. 89-102): Springer Berlin / Heidelberg.

Cui, X., & Potok, T. E. (2005). Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm. *Journal of Computer Sciences* (Special Issue), 27-33.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Survey., 41*(3), 1-58.

Chen, H.-L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S.-J., et al. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. *Knowledge-Based Systems, 24*(8), 1348-1359.

Chen, H., Buntin, P., She, L., Sutjahjo, S., Sommer, C., & Neely, D. (1994). Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment on Greyhound Racing. *IEEE Expert, 9*(6), 21-27.

Chen, M.-C., & Huang, S.-H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications, 24*(4), 433-441.

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications, 36*(4), 7611-7616.

Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics, 4*(1), 43-51.

Dennis, W. L. (1995). Fair lending and credit scoring. *Mortgage Banking, 56*(2), 25, 38.

Derelioğlu, G., Gürgen, F., & Okay, N. (2009). A Neural Approach for SME's Credit Risk Analysis in Turkey. In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition.* (Vol. 5632, pp. 749-759): Springer Berlin / Heidelberg.

Diallo, B. (2006). *Un Modele de "Credit Scoring" Pour Une Institution de Micro-Finance Africaine: Le Cas De Nyesigiso au Mali*. Orléans: Laboratoire d'Economie d'Orléans.

Dinh, T. H. T., & Kleimeier, S. (2007). A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis, 16*(5), 471-495.

Dong, G., Lai, K. K., & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science, 1*(1), 2463-2468.

Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. Boston, MA: MIT Press.

Dréo, J., Pétrowski, A., Siarry, P., & Taillard, E. (2006). *Metaheuristics for Hard Optimization: Methods and Case Studies*: Springer.

Du, W., & Li, B. (2008). Multi-strategy ensemble particle swarm optimization for dynamic optimization. *Information Sciences, 178*(15), 3096-3109.

Eberhart, R. C., Kennedy, J., & Shi, Y. (2001). *Swarm intelligence*. Burlington, MA.: Elsevier.

Eberhart, R. C., & Shi, Y. (2000). *Comparing inertia weights and constriction factors in particle swarm optimization* Paper presented at the Evolutionary Computation, 2000. Proceedings of the 2000 Congress on La Jolla, CA.

Eberhart, R. C., Simpson, P., & Dobbins, R. (1996). *Computational Intelligence PC Tools*. San Diego, CA: Academic Press Professional, Inc.

Eberlein, E., Frey, R., Kalkbrenner, M., & Overbeck, L. (2007). Mathematics in financial risk management. *Jahresbericht der Deutschen Mathematiker-Vereinigung 109*, 24.

Elías, A., Ochoa-Zezzatti, A., Padilla, A., & Ponce, J. (2011). Outlier analysis for plastic card fraud detection a hybridized and multi-objective approach *Hybrid Artificial Intelligent Systems* (pp. 1-9): Springer.

Elias, A., & Padilla, A. (2012). Algoritmo Inteligente para Agrupamiento de Clientes en Microfinancieras. In UAA (Ed.). Aguascalientes.

Elias, A., Padilla, A., & Ochoa, A. (2012). *Credit Scoring Mechanisms for IMFs* Paper presented at the Doctoral Consortium in MICAI 2012.

Elías, A., Padilla, A., & Padilla, F. (2012). Credit Scoring for Microfinance Institutions in México an Ensemble and Hybridized Approach. *Lecture Notes in Information Technology, 21*.

Escalona Cortés, A. (2011). *Uso de los Modelos Credit Scoring en Microfinanzas.* Colegio de Postgraduados, Texcoco, Edo. de México.

Fakhfakh, M., Cooren, Y., Sallem, A., Loulou, M., & Siarry, P. (2010). Analog circuit design optimization through the particle swarm optimization technique. *Analog Integrated Circuits and Signal Processing, 63*(1), 71-82.

Farquad, M. A. H., Ravi, V., Sriramjee, S., & Praveen, G. (2011). Credit Scoring Using PCA-SVM Hybrid Model. In V. V. Das, J. Stephen & Y. Chaba (Eds.), *Computer Networks and Information Technologies* (Vol. 142, pp. 249-253): Springer Berlin Heidelberg.

Fayyad, U., & Uthurusamy, R. (2002). Evolving data into mining solutions for insights. *Commun. ACM, 45*(8), 28-31.

Feigenbaum, E. A., Buchanan, B. G., & Lederberg, J. (1970). *On generality and problem solving: a case study using the DENDRAL program* (Technical Report). Stanford, CA.: Stanford University.

Fındık, O., Babaoğlu, İ., & Ülker, E. (2010). A color image watermarking scheme based on hybrid classification method: Particle swarm optimization and k-nearest neighbor algorithm. *Optics Communications, 283*(24), 4916-4922.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics, 7*(2), 179-188.

Fogel, L. J., Owens, A. J., & Walsh, M. J. (1966). *Artificial intelligence through simulated evolution*. Chinchester, UK: John Wiley & Sons, Ltd.

Friedberg, R. M. (1958). A Learning Machine: Part I. *IBM Journal of Research and Development, 2*(1), 2-13.

Fu, H., & Liu, X. (2011). A Hybrid Model for Credit Evaluation Problem. In Y. Tan, Y. Shi, Y. Chai & G. Wang (Eds.), *Advances in Swarm Intelligence.* (Vol. 6728, pp. 626-634): Springer Berlin / Heidelberg.

Gao, H., & Xu, W. (2011). Particle swarm algorithm with hybrid mutation strategy. *Applied Soft Computing, 11*(8), 5129-5142.

Ghodselahi, A. (2011). A Hybrid Support Vector Machine Ensemble Model for Credit Scoring. *International Journal of Computer Applications, 17*(5), 1-5.

Ghodselahi, A., & Amirmadhi, A. (2011). Application of Artificial Intelligence Techniques for Credit Risk Evaluation. *International Journal of Modeling and Optimization 1*(3), 243-249.

Glover, F. (1990). Improved Linear Programming Models for Discriminant Analysis*. *Decision Sciences, 21*(4), 771-785.

Glover, F. E., & Kochenberger, G. A. (2003). *Handbook of Metaheuristics*: Kluwer Academic Pub.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning* (1st ed.). Boston, MA: Addison-Wesley Longman Publishing Co., Inc. .

Grablowsky, B. J., & Talley, W. K. (1981). Probit and discriminant functions for classifying credit applicants: a comparison. *Journal of Economics and Business, 33*, 254-261.

Greene, W. H. (1992). *A Statistical Model for Credit Scoring, Working Papers (92-29)*. New York, NY: New York University, Leonard N. Stern School of Business.

Grosan, C., & Abraham, A. (2007). Hybrid Evolutionary Algorithms: Methodologies, Architectures, and Reviews. In A. Abraham, C. Grosan & H. Ishibuchi (Eds.), (Vol. 75, pp. 1-17): Springer Berlin / Heidelberg.

Hájek, P. Credit rating analysis using adaptive fuzzy rule-based systems: an industry-specific approach. [In Press]. *Central European Journal of Operations Research*, 1-14 doi10.1007/s10100-10011-10229-10100.

Hájek, P., & Olej, V. (2011). Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning. *Neural Computing & Applications, 20*(6), 761-773.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann Publishers.

Hand, D. J. (1981). *Discrimination and classification* (First ed.): J. Wiley.

Handl, J. K. (2006). *Multiobjective approaches to the data-driven analysis of biological systems.* University of Manchester, Manchester.

Hasan, I., & Zazzara, C. (2006). Pricing risky bank loans in the new Basel II environment. *Journal of Banking Regulation*, 243-267.

Hashemi, A., & Meybodi, M. (2009). Cellular PSO: A PSO for Dynamic Environments

Advances in Computation and Intelligence. In Z. Cai, Z. Li, Z. Kang & Y. Liu (Eds.), (Vol. 5821, pp. 422-433): Springer Berlin / Heidelberg.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and Their Applications, IEEE [see also IEEE Intelligent Systems], 13*(4), 18-28.

Hebb, D. O. (2002). *The organization of behavior: a neuropsychological theory* (Reprinted ed.): L. Erlbaum Associates.

Henley, W. E., & Hand, D. J. (1996). A k-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk. *Journal of the Royal Statistical Society, 45*(1), 77-95.

Hernández Romero, O., & Almorín Albino, R. (2005). *Las Microfinanzas en México, Tendencias y Perspectivas* (1 ed.). Cuernavaca, Morelos, México: Fundación Ayuda en Acción.

Holland, J. H. (1962). Outline for a Logical Theory of Adaptive Systems. *Journal of the ACM, 9*(3), 297-314.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (Reprinted ed.): The MIT Press.

Hsieh, N.-C., & Hung, L.-P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications, 37*(1), 534-545.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems., 37*(4), 543-558.

Hui, K., Xuchun, L., Lei, W., Earn Khwang, T., Jian-Gang, W., & Venkateswarlu, R. (2005, 31 July-4 Aug. 2005). *Generalized 2D principal component analysis.* Paper presented at the Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on.

Ionescu, S. A., Murgoci, C. S., Gheorghe, C. M., & Ionescu, E. (2009). Towards Profitability on the Financial Markets; A Discriminant Analysis Approach. *WSEAS Transactions on Business and Economics, 6*(1), 99-111.

Jackson, J. E. (1991). *A user's guide to principal components* (Vol. 244): Wiley-Interscience.

Jentzsch, N. (2007). *Financial Privacy: An International Comparison of Credit Reporting Systems* (2nd ed.). Heidelgerg, Germany: Springer Gmbh & Co.

Jia, D., Zheng, G., Qu, B., & Khan, M. K. (2011). A hybrid particle swarm optimization algorithm for high-dimensional problems. *Computers & Industrial Engineering, 61*(4), 1117-1122.

Jia, W., Vadera, S., Dayson, K., Burridge, D., & Clough, I. (2010, 17-19 Sept. 2010). *A comparison of data mining methods in microfinance.* Paper presented at the Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on.

Jiang, Y. (2009). *Credit Scoring Model Based on the Decision Tree and the Simulated Annealing Algorithm .* Paper presented at the Computer Science and Information Engineering, 2009 WRI World Congress on Los Angeles, CA.

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications, 13*(2), 97-108.

Jun, X., & Chang, H. (2009, 20-22 Sept). *The Discrete Binary Version of the Improved Particle Swarm Optimization Algorithm* Paper presented at the Management and Service Science, 2009. MASS '09. International Conference on Wuhan, China.

Kaelbling, L. P., Littman, M. L., & Moore, A. P. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research, 4*, 237-285.

Kamosi, M., Hashemi, A., & Meybodi, M. (2010). A New Particle Swarm Optimization Algorithm for Dynamic Environments. In B. Panigrahi, S. Das, P. Suganthan & S. Dash (Eds.), *Swarm, Evolutionary, and Memetic Computing* (Vol. 6466, pp. 129-138): Springer Berlin / Heidelberg.

Karlan, D., & Zinman, J. (2011). Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science, 332*(6035), 1278-1284.

Kennedy, J., & Eberhart, R. (1995, Nov/Dec 1995). *Particle swarm optimization.* Paper presented at the Neural Networks, 1995. Proceedings., IEEE International Conference on, Perth, WA, Australia.

Kennedy, J., & Eberhart, R. C. (1997, 12 - 15 Oct). *A discrete binary version of the particle swarm algorithm* Paper presented at the Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on Orlando, FL, USA.

Keramati, A., & Yousefi, N. (2011). *A Proposed Classification of Data Mining Techniques in Credit Scoring.* Paper presented at the 2011 International Conference on Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia.

Khanesar, M. A., Teshnehlab, M., & Shoorehdeli, M. A. (2007, 27-29 Jun). *A novel binary particle swarm optimization* Paper presented at the Control & Automation, 2007. MED '07. Mediterranean Conference on Athens, GR.

Kim, J.-C., Kim, D.-H., Kim, J.-J., Ye, J.-S., & Lee, H.-S. (2000). Segmenting the Korean housing market using multiple discriminant analysis. *Construction Management and Economics, 18*(1), 45-54.

Kinda, O., & Achonu, A. (2012). Building a Credit Scoring Model for the Savings and Credit Mutual of the Potou Zone. *Consilience - The Journal of Sustainable Development, 7*, 23-31.

Kocenda, E., & Vojtek, M. (2009). Default Predictors and Credit Scoring Models for Retail Banking. *SSRN eLibrary*.

Kolesar, P., & Showers, J. L. (1985). A Robust Credit Screening Model Using Categorical Data. *Management Science, 31*(2), 123-133.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.

Kulkosky, E. (1996). Credit scoring could have a downside, experts say.(Mortgage Banks Association of America panel discussion). *American Banker, 16*, 208-215.

Kumar, R., Sharma, D., & Sadu, A. (2011). A hybrid multi-agent based particle swarm optimization algorithm for economic power dispatch. *International Journal of Electrical Power & Energy Systems, 33*(1), 115-123.

Lahsasna, A., Ainon, R. N., & Wah, T. Y. (2010). Credit Scoring Models Using Soft Computing Methods: A Survey. *Int. Arab J. Inf. Technol.*, 115-123.

Lai, K., Yu, L., Wang, S., & Zhou, L. (2006). Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model Artificial Neural Networks – ICANN 2006. In S. Kollias, A. Stafylopatis, W. Duch & E. Oja (Eds.), (Vol. 4132, pp. 682-690): Springer Berlin / Heidelberg.

Lara Rubio, J. (2010). *La Gestión del Riesgo de Crédito en las Instituciones de Microfinanzas.* Universidad de Granada, Grananda, España.

Lee, S., Soak, S., Oh, S., Pedrycz, W., & Jeon, M. (2008). Modified binary particle swarm optimization. *Progress in Natural Science, 18*(9), 1161-1166.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*(4), 1113-1130.

Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245-254.

Li, F.-C. (2009, Aug 14-16). *The Hybrid Credit Scoring Strategies Based on KNN Classifier* Paper presented at the Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on Tianjin, China.

Li, X.-s., & Guo, Y.-h. (2006). Personal Credit Scoring Models on Naive Bayesian Classifier. *Computer Enginnering and Applications, 42*(1), 197-201.

Lin, T.-L., Horng, S.-J., Kao, T.-W., Chen, Y.-H., Run, R.-S., Chen, R.-J., et al. (2010). An efficient job-shop scheduling algorithm based on particle swarm optimization. *Expert Systems with Applications, 37*(3), 2629-2636.

Liu, C., & Wechsler, H. (2000). Evolutionary pursuit and its application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22*(6), 570-582.

Liu, L., Wang, D., & Yang, S. (2008). Compound Particle Swarm Optimization in Dynamic Environments. In M. Giacobini, A. Brabazon, S. Cagnoni, G. Di Caro, R. Drechsler, A. Ekárt, A. Esparcia-Alcázar, M. Farooq, A. Fink, J. McCormack, M. O'Neill, J. Romero, F. Rothlauf, G. Squillero, A. Uyar & S. Yang (Eds.), *Applications of Evolutionary Computing* (Vol. 4974, pp. 616-625): Springer Berlin / Heidelberg.

Lozano, M., & García-Martínez, C. (2010). Hybrid metaheuristics with evolutionary algorithms specializing in intensification and diversification: Overview and progress report. *Computers & Operations Research, 37*(3), 481-497.

Luo, L., Xiong, J., & Zhou, Q. (2011). Credit Risk Model and Bayesian Improvement for Companies in China. In D. D. Wu & Y. Zhou (Eds.), *Modeling Risk Management for Resources and Environment in China.* (pp. 265-273): Springer Berlin Heidelberg.

Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on, 28*(2), 129-137.

Makowski, P. (1985). Credit scoring branches out: decision tree - recent technology. *Credit World, 75*(1), 30-37.

Maldonado, S., & Paredes, G. (2010). A Semi-supervised Approach for Reject Inference in Credit Scoring Using SVMs. In P. Perner (Ed.), *Advances in Data Mining. Applications and Theoretical Aspects.* (Vol. 6171, pp. 558-571): Springer Berlin / Heidelberg.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research, 136*(1), 190-211.

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega, 31*(2), 83-96.

Mangasarian, O. L. (1965). Linear and Nonlinear Separation of Patterns by Linear Programming. *Operations Research, 13*(3), 444-452.

Mansell Carstens, C. (1995). *Las Finanzas Populares en México, el redescubrimiento de un sistema financiero olvidado* (1 ed.). Centro de Estudios Monetarios Latinoamericanos: Editorial Milenio, Instituto Tecnológico Autónomo de México.

Marinakis, Y., & Marinaki, M. (2010). A Hybrid Multi-Swarm Particle Swarm Optimization algorithm for the Probabilistic Traveling Salesman Problem. *Computers & Operations Research, 37*(3), 432-442.

Marinakis, Y., Marinaki, M., Doumpos, M., & Zopounidis, C. (2009). Ant colony and particle swarm optimization for financial classification problems. *Expert Systems with Applications, 36*(7), 10604-10611.

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & Baesens, B. (2008). Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. In J. Diederich (Ed.), *Rule Extraction from Support Vector Machines.* (Vol. 80, pp. 33-63): Springer Berlin / Heidelberg.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology, 5*(4), 115-133.

Mendes, R., Kennedy, J., & Neves, J. (2004). The fully informed particle swarm: Simpler, maybe better. *IEEE Transactions on Evolutionary Computation, 8*(3), 204-210.

Mester, L. J. (1997). What's the point of credit scoring? *Business Review*(September), 3-16.

Milena, E., Miller, M., & Simbaqueba, L. (2005). *The Case for Information Sharing by Microfinance Institutions: Emprirical Evidence of the Value of Credit Bureau-Type in the Nicaraguan Microfinance Sector* (Research Report). New York, NY: The World Bank.

Miller, M., & Rojas, D. (2004). *Improving Access to Credit for SMEs: An Empirical Analysis of the Viability of Pooled Data SME Credit Scoring Models in Brazil, Colombia & Mexico*. New York: The World Bank.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM, 42*(11), 30-36.

Moslehi, G., & Mahnam, M. (2011). A Pareto approach to multi-objective flexible job-shop scheduling problem using particle swarm optimization and local search. [doi: 10.1016/j.ijpe.2010.08.004]. *International Journal of Production Economics, 129*(1), 14-22.

Muhammad, M. S., Selvan, K. V., Masra, S. M. W., Ibrahim, Z., & Abidin, A. F. Z. (2011, 11-15 Apr). *An improved binary particle swarm optimization algorithm for DNA encoding enhancement* Paper presented at the Swarm Intelligence (SIS), 2011 IEEE Symposium on Paris, FR.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*(2, Part 2), 3028-3033.

Niknam, T., & Amiri, B. (2010). An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. *Applied Soft Computing, 10*(1), 183-197.

Nimtawat, A., & Nanakorn, P. (2011). Simple Particle Swarm Optimization for Solving Beam-Slab Layout Design Problems. *Procedia Engineering, 14*(0), 1392-1398.

Novoa-Hernández, P., Corona, C., & Pelta, D. (2011). Efficient multi-swarm PSO algorithms for dynamic environments. *Memetic Computing, 3*(3), 163-174.

Nwulu, N. I., Oroja, S., & Ilkan, M. (2011). Credit Scoring Using Soft Computing Schemes: A Comparison between Support Vector Machines and Artificial Neural Networks. In E. Ariwa & E. El-Qawasmeh (Eds.), *Digital Enterprise and Information Systems.* (Vol. 194, pp. 275-286): Springer Berlin Heidelberg.

Ochoa, A., Azpeitia, D., Elías, A., Salazar, P., García, E., Maldonado, M., et al. (2012). Explaining Diverse Application Domains Analyzed from Data Mining Perspective. In A. Karahoca (Ed.), *Data Mining Applications in Engineering and Medicine*.

Ochoa, A., Ponce, J., Elias, A., Jaramillo, R., Ornelas, F., Hernandez, A., et al. (2011). *Analysis of Cyber-bullying in a virtual social networking.* Paper presented at the Hybrid Intelligent Systems (HIS), 2011 11th International Conference on.

Ochoa, A., Ponce, J., Elias, A., Ornelas, F., Jaramillo, R. n., Zataraín, R. n., et al. (2011). New Implementations of Data Mining in a Plethora of Human Activities. In K. Funatsu (Ed.), *Knowledge-Oriented Applications in Data Mining*.

Olson, D. L. (2009). Data Mining. In C. A. Floudas & P. M. Pardalos (Eds.), *Encyclopedia of Optimization* (pp. 600-607): Springer US.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464-473.

Orgler, Y. E. (1970). A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit and Banking, 2*(4), 435-445.

Orgler, Y. E., & Sciences, T.-A. U. D. o. E. (1971). *Evaluation of Bank consumer loans with credit scoring models*: Tel-Aviv University, Dept. of Envirnonmental Sciences.

Pang, S.-l., & Gong, J.-z. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering - Theory & Practice, 29*(12), 94-104.

Panigrahi, S., Kundu, A., Sural, S., & Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion, 10*(4), 354-363.

Parsopoulos, K. E., & Vrahatis, M. N. (2004). On the computation of all global minimizers through particle swarm optimization. *Evolutionary Computation, IEEE Transactions on, 8*(3), 211-224.

Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research, 112*(2), 310-321.

Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization an Overview. *Swarm Intelligence, 1*(1), 33-57.

ProDesarrollo Finanzas y Microempresa, A. C. (2011). *BENCHMARKING de las microfinanzas en México 2010: Un informe del sector, Reporte Técnico*. México, DF: Microfinance Information Exchange (MIX).

Qi, C. (2011). Application of Improved Discrete Particle Swarm Optimization in Logistics Distribution Routing Problem. *Procedia Engineering, 15*(0), 3673-3677.

Raidl, G. (2006). A Unified View on Hybrid Metaheuristics Hybrid Metaheuristics. In F. Almeida, M. Blesa Aguilera, C. Blum, J. Moreno Vega, M. Pérez Pérez, A. Roli & M. Sampels (Eds.), (Vol. 4030, pp. 1-12): Springer Berlin / Heidelberg.

Rayo Canton, S., Lara Rubio, J., & Camino Blasco, D. (2010). Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science, 15*(28), 89-124.

Rechenberg, I. (1965). *Cybernetic solution path of an experimental problem* (Translation 1122). Farnborough, UK.: Royal Aircraft Establishment.

Reeves, W. T. (1983). Particle Systems\;a Technique for Modeling a Class of Fuzzy Objects. *ACM Transactions on Graphics., 2*(2), 91-108.

Reinke, J. (1998). How to lend like mad and make a profit: a micro-credit paradigm versus the start-up fund in South Africa. *Journal of Development Studies, 34*(3), 44-61.

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (Second ed.). Newark, NJ: Jhon Wiley & Sons.

Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. *SIGGRAPH Comput. Graph., 21*(4), 25-34.

Rezazadeh, H., Ghazanfari, M., Saidi-Mehrabad, M., & Jafar Sadjadi, S. (2009). An extended discrete particle swarm optimization algorithm for the dynamic facility layout problem. *Journal of Zhejiang University - Science A, 10*(4), 520-529.

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society, 56*(3), 409-456.

Robati, A., Barani, G. A., Nezam Abadi Pour, H., Fadaee, M. J., & Rahimi Pour Anaraki, J. (2012). Balanced fuzzy particle swarm optimization. *Applied Mathematical Modelling, 36*(5), 2169-2177.

Rosenberg, E., & Gleit, A. (1994). Quantitative Methods in Credit Management: A Survey. *Operations Research, 42*(4), 589-613.

Sadri, J., & Suen, C. Y. (2006, 16-21 Jul). *A Genetic Binary Particle Swarm Optimization Model* Paper presented at the Evolutionary Computation, 2006. CEC 2006. IEEE Congress on Vancouver, BC.

Sarkar, S., & Das, S. (2010). A Hybrid Particle Swarm with Differential Evolution Operator Approach (DEPSO) for Linear Array Synthesis. In B. Panigrahi, S. Das, P. Suganthan & S. Dash (Eds.), *Swarm, Evolutionary, and Memetic Computing* (Vol. 6466, pp. 416-423): Springer Berlin / Heidelberg.

Schebesch, K. B., & Stecking, R. (2008). Using Multiple SVM Models for Unbalanced Credit Scoring Data Sets. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 515-522): Springer Berlin Heidelberg.

Schoeman, I., & Engelbrecht, A. (2006). Niching for Dynamic Environments Using Particle Swarm Optimization. In T.-D. Wang, X. Li, S.-H. Chen, X. Wang, H. Abbass, H. Iba, G.-L. Chen & X. Yao (Eds.), *Simulated Evolution and Learning* (Vol. 4247, pp. 134-141): Springer Berlin / Heidelberg.

Schreiner, M. (1999). *A Scoring Model of the Risk of Costly Arrears at a Microfinance Lender in Bolivia*. St. Louis, MO.: Washington University.

Schreiner, M. (2001). *Credit Scoring for Microfinance: Can It Work?* (Ocacional Report): Microfinance Risk Management and Center for Social Development.

Schreiner, M. (2003). Scoring: The Next Breakthrough in Microcredit?, *Occasional paper 7* (pp. 43). Washington D. C.: Consultative Group to Assist the Poorest.

Schreiner, M. (2004). *Benefits and Pitfalls of Statistitcal Credit Scoring for Microfinance* (Ocacional Report). St. Louis, MO: Washington University.

Schwarz, A., & Arminger, G. (2005). Credit Scoring Using Global and Local Statistical Models. In C. Weihs & W. Gaul (Eds.), *Classification — the Ubiquitous Challenge.* (pp. 442-449): Springer Berlin Heidelberg.

Seni, G., & Elder, J. F. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery, 2*(1), 1-126.

Sharma, M., & Zeller, M. (1996). Repayment performance in group-based credit programs in Bangladesh. *FCND discussion papers, 15,* 1731-1742.

Shi, Y., & Eberhart, R. (1998, 4-9 May 1998). *A modified particle swarm optimizer.* Paper presented at the Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, Anchorage, AK, USA.

Shi, Y., Liu, H., Gao, L., & Zhang, G. (2011). Cellular particle swarm optimization. *Information Sciences, 181*(20), 4460-4493.

Showers, J. L., & Chakrin, L. M. (1981). Reducing Uncollectible Revenue from Residential Telephone Customers. *Interfaces, 11*(6), 21-34.

Siami, M., Gholamian, M., Basiri, J., & Fathian, M. (2011). An Application of Locally Linear Model Tree Algorithm for Predictive Accuracy of Credit Scoring. In L. Bellatreche & F. Mota Pinto (Eds.), *Model and Data Engineering* (Vol. 6918, pp. 133-142): Springer Berlin / Heidelberg.

Sinha, A., & Goldberg, D. E. (2003). *A Survey of Hybrid Genetic and Evolutionary Algorithms* (Technical Report). Urbana-Champaign, ILL: University of Illinois.

Smalz, R., & Conrad, M. (1994). Combining evolution with credit apportionment: A new learning algorithm for neural nets. *Neural Networks, 7*(2), 341-351.

Sohn, S. Y., & Kim, H. S. (2007). Random effects logistic regression model for default prediction of technology credit guarantee fund. *European Journal of Operational Research, 183*(1), 472-478.

Sreekantha, D. K., & Kulkarni, R. V. (2010). Expert system design for credit risk evaluation using neuro-fuzzy logic. *Expert Systems*, no-no.

Stecking, R., & Schebesch, K. (2007). Combining Support Vector Machines for Credit Scoring. In K.-H. Waldmann & U. M. Stocker (Eds.), *Operations Research Proceedings 2006* (Vol. 2006, pp. 135-140): Springer Berlin Heidelberg.

Stefan, T., & Svetlozar T, R. (2009). Chapter 2 - Rating and Scoring Techniques *Rating Based Modeling of Credit Risk* (pp. 11-30). Boston: Academic Press.

Storn, R., & Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization, 11*, 341-359.

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., & Cela-Díaz, F. (2010). Statistical Methods for Fighting Financial Crimes. . *Technometrics, 52*(1), 5-19.

Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications, 36*(3, Part 1), 4736-4744.

Talbi, E. G. (2002). A Taxonomy of Hybrid Metaheuristics. *Journal of Heuristics, 8*(5), 541-564.

Tarsauliya, A., Kala, R., Tiwari, R., & Shukla, A. (2011). Financial Time Series Forecast Using Neural Network Ensembles. In Y. Tan, Y. Shi, Y. Chai & G. Wang (Eds.), *Advances in Swarm Intelligence* (Vol. 6728, pp. 480-488): Springer Berlin / Heidelberg.

Tay, F. E. H., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing, 48*(1-4), 847-861.

Thangaraj, R., Pant, M., Abraham, A., & Bouvry, P. (2011). Particle swarm optimization: Hybridization perspectives and experimental illustrations. [doi: 10.1016/j.amc.2010.12.053]. *Applied Mathematics and Computation, 217*(12), 5208-5226.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149-172.

Thomas, L. C. (2009). Operations research in consumer finance: challenges for operational research. Unpublished Discussion Paper. University of Southampton.

Tolle, K. M., Chen, H., & Chow, H.-H. (2000). Estimating drug/plasma concentration levels by applying neural networks to pharmacokinetic data sets. *Decision Support Systems, 30*(2), 139-151.

Trevino, L. J., & Daniels, J. D. (1995). FDI theory and foreign direct investment in the United States: a comparison of investors and non-investors. *International Business Review, 4*(2), 177-194.

Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing, 10*(2), 374-380.

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*(4), 2639-2649.

Tsai, M.-C., Lin, S.-P., Cheng, C.-C., & Lin, Y.-P. (2009). The consumer loan default predicting model - An application of DEA-DA and neural network. *Expert Syst. Appl., 36*(9), 11682-11690.

Tu, C.-J., & Chuang, L.-Y. (2007). Feature selection using PSO-SVM. *IAENG International journal of computer science, 33*(1), 6.

Turner, M. A., & Varghese, R. (2006). Contribución de la Informaciónn Negativa y Positiva: Los Beneficios de una Mayor participación del Reporte Crediticio en América Latina y los Costos del Status Quo. Unpublished White Paper. Information Policy Institute.

Valdez, F., Melin, P., & Castillo, O. (2011). An improved evolutionary method with fuzzy logic for combining Particle Swarm Optimization and Genetic Algorithms. *Applied Soft Computing, 11*(2), 2625-2632.

Van Gestel, T., Baesens, B., Garcia, J., & Van Dijcke, P. (2003). A support vector machine approach to credit scoring. *Bank en Financiewezen, 2*, 73-82.

Van Gestel, T., Suykens, J. A. K., Baestaens, D.-E., Lambrechts, A., Lanckriet, G., Vandaele, B., et al. (2001). Financial time series prediction using least squares support vector machines within the evidence framework *IEEE Transactions on Neural Networks, 12,* 809-821.

Van Gool, J., Baesens, B., Sercu, P., & Verbeke, W. (2009). *An Analysis of the Applicability of Credit Scoring for Microfinance.* Paper presented at the Academic and Business Research Institute Conference Orlando, FL.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY, USA.: Springer Verlag New York, Inc.

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking & Finance, 22*(10-11), 1421-1439.

Verikas, A., Kalsyte, Z., Bacauskiene, M., & Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft Computing - A Fusion of Foundations, Methodologies and Applications, 14*(9), 995-1010.

Viganò, L. (1993). A Credit Scoring Model for Development Banks: An African case study. *Savings and Development, 17*(4), 441-482.

Vogelgesang, U. (2003). Microfinance in Times of Crisis: The Effects of Competition, Rising Indebtedness, and Economic Crisis on Repayment Behavior. *World Development, 31*(12), 2085-2114.

Voglis, C., Parsopoulos, K. E., Papageorgiou, D. G., Lagaris, I. E., & Vrahatis, M. N. (2012). MEMPSODE: A global optimization software based on hybridization of population-based algorithms and local searches. *Computer Physics Communications, 183*(5), 1139-1154.

Voss, M. S. (2005, 8-10 June 2005). *Principal component particle swarm optimization (PCPSO).* Paper presented at the Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications, 38*(1), 223-230.

Wang, G., & Ma, J. (2012). A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. *Expert Systems with Applications, 39*(5), 5325-5331.

Wang, J., Cai, Y., Zhou, Y., Wang, R., & Li, C. (2011). Discrete particle swarm optimization based on estimation of distribution for terminal assignment problems. *Computers & Industrial Engineering, 60*(4), 566-575.

Wang, Y., Wang, S., & Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk *IEEE Transactions on Fuzzy Systems, 13*(6), 820-831.

Wei-Li, J. (2011). Research and Application of Credit Score Based on Decision Tree Model. In D. Zeng (Ed.), *Applied Informatics and Communication* (Vol. 224, pp. 493-501): Springer Berlin Heidelberg.

Welch, W. J. (1982). Algorithmic complexity: Three NP-hard problems in computational statistics. *Journal of Statistical Computation and Simulation, 15*(1), 17-25.

Widrow, B., Rumelhart, D. E., & Lehr, M. A. (1994). Neural networks: applications in industry, business and science. *Commun. ACM, 37*(3), 93-105.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behaviour. *Journal of Financial and Quantitative Analysis, 15*, 757-770.

Wong, C., & Versace, M. (2012). CARTMAP: a neural network method for automated feature selection in financial time series forecasting. *Neural Computing & Applications*, 1-9.

Xikun, L., & Zhengzheng, Z. (2010, November 7-9). *A Credit Model Based on Multi-Variables Fuzzy Reasoning* Paper presented at the E-Product E-Service and E-Entertainment (ICEEE), 2010 International Conference on. , Henan, China.

Xu, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications, 36*(2, Part 2), 2625-2632.

Ya-qiong, P. (2007). *A Study on Evaluation of Consumer Credit's Risks of Commercial Banks.* Paper presented at the International Conference on Wireless Communication (WiCom 2007), Shangai.

Yang, Z., Wu, D., Fu, G., & Luo, C. (2008). Credit Risk Evaluation Using Neural Networks. In D. L. Olson & D. Wu (Eds.), *New Frontiers in Enterprise Risk Management* (pp. 163-179): Springer Berlin Heidelberg.

Yao, P., Wu, C., & Yao, M. (2009). Credit Risk Assessment Model of Commercial Banks Based on Fuzzy Neural Network In W. Yu, H. He & N. Zhang (Eds.), *Advances in Neural Networks – ISNN 2009* (Vol. 5551, pp. 976-985): Springer Berlin / Heidelberg.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications, 38*(10), 13274-13283.

Yi, J., & Li Hua, W. (2009, 11-13 Dec. 2009). *Credit Scoring Model Based on Simple Naive Bayesian Classifier and a Rough Set.* Paper presented at the Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on.

Yu, L., Lai, K., Wang, S., & Zhou, L. (2007). A Least Squares Fuzzy SVM Approach to Credit Risk Assessment. In B.-Y. Cao (Ed.), *Fuzzy Information and Engineering* (Vol. 40, pp. 865-874): Springer Berlin / Heidelberg.

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications, 34*(2), 1434-1444.

Yu, L., Wang, S., & Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *European Journal of Operational Research, 195*(3), 942-959.

Yu, L., Wang, S., Lai, K. K., & Zhou, L. (2008a). Evolving Least Squares SVM for Credit Risk Analysis *Bio-Inspired Credit Risk Analysis* (pp. 105-131): Springer Berlin Heidelberg.

Yu, L., Wang, S., Lai, K. K., & Zhou, L. (2008b). An Intelligent-Agent-Based Multicriteria Fuzzy Group Decision Making Model for Credit Risk Analysis *Bio-Inspired Credit Risk Analysis* (pp. 197-222): Springer Berlin Heidelberg.

Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications, 37*(2), 1351-1360.

Yuan, X., Nie, H., Su, A., Wang, L., & Yuan, Y. (2009). An improved binary particle swarm optimization for unit commitment problem. *Expert Systems with Applications, 36*(4), 8049-8055.

Yun, L., Qiu-yan, C., & Hua, Z. (2011, 3-4 Dec. 2011). *Application of the PSO-SVM Model for Credit Scoring.* Paper presented at the Computational Intelligence and Security (CIS), 2011 Seventh International Conference on.

Zeller, M. (1998). Determinants of Repayment Performance in Credit Groups: The Role of Program Design, Intragroup Risk Pooling, and Social Cohesion. *Economic Development and Cultural Change, 46*(3), 599-620.

Zhang, D., Hifi, M., Chen, Q., & Ye, W. (2008). *A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines*. Paper presented at the Proceedings of the 2008 Fourth International Conference on Natural Computation - Volume 07.

Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications, 37*(12), 7838-7843.

Zhang, J. L., & Härdle, W. K. (2010). The Bayesian Additive Classification Tree applied to credit risk modelling. *Computational Statistics & Data Analysis, 54*(5), 1197-1205.

Zhang, L.-l., Hui, X.-f., & Wang, L. (2009). *Application of adaptive support vector machines method in credit scoring.* Paper presented at the International Conference on Management Science and Engineering, 2009. ICMSE 2009. , Moscow.

Zhang, L., & Hui, X. (2009). Application of Support Vector Machines Method in Credit Scoring. In H. Wang, Y. Shen, T. Huang & Z. Zeng (Eds.), *The Sixth International Symposium on Neural Networks (ISNN 2009)* (Vol. 56, pp. 283-290): Springer Berlin / Heidelberg.

Zhou, L., & Lai, K. (2009). Adaboosting Neural Networks for Credit Scoring. In H. Wang, Y. Shen, T. Huang & Z. Zeng (Eds.), *The Sixth International Symposium on Neural Networks (ISNN 2009)* (Vol. 56, pp. 875-884): Springer Berlin / Heidelberg.

Zhou, L., Lai, K., & Yu, L. (2009). Credit scoring using support vector machines with direct search for parameters selection. *Soft Computing - A Fusion of Foundations, Methodologies and Applications, 13*(2), 149-155.

Zurada, J. (2010). *Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions? .* Paper presented at the System Sciences (HICSS), 2010 43rd Hawaii International Conference on Honolulu, HI.

Anexo A. Ejemplos del dataset.

Tabla A0-1. Muestra ejemplo de los datos del data set antes de pretratamiento.

CODIGO	SEXO	CALLE	TELEFONO	EDOCIVIL	NIVESCOLAR	NACIMIENTO	NACIOMU	NACIOLO	NODEPEND	ALTA	RESTRICCION	REGMARITAL	CDGPRPE	CDGOCPE	SOLICITUD
168	F	DOM. CONOCIDO GUADALUPE VICTORIA		S	T	21/07/1980			2	08/08/2005			14	14	01/08/2005 11:38
5224	F	MATAMOROS 138		C	P	05/09/1980	21	1	5	18/07/2006	G	M	56	5	31/12/2009 12:44
498	F	CALLE SIN NOMBRE NO. 86		S	C	13/04/1980			3	14/09/2005	B		41	41	14/09/2005 11:03
5266	F	CRISANTEMAS #1315-A	493-14-55	C	P	10/09/1974	19	1	5	21/07/2006			23	23	24/07/2006 15:12
5259	F	J.J. CALVO #6606		C	P	02/05/1965	19	1	5	21/07/2006			23	23	24/07/2006 14:11
4853	F	CIPRES No 811	484-92-10	C	S	22/06/1962			1	05/06/2006			18	18	11/09/2006 12:19
4688	M	AGUILILLAS NO. 11		C	L	07/03/1966	32		5	11/05/2006	G	M	42	42	06/07/2007 17:36
4973	F	PASCUAL OROZCO SUR 423		C	P	26/01/1950	21	1	4	23/06/2006		M	56	41	23/10/2006 09:29
4200	M	AVE 6ı SUR # 946	4826970	S	S	29/11/1959	21	4	2	27/02/2006	N		39	39	27/02/2006 09:56
91	F	FRANCISCO I.MADERO SUR # 203	4859635	D	T	16/07/1971			4	17/06/2005			39	39	13/06/2005 10:34
5061	F	ALABASTROS NO 2533	4829587	C	P	06/05/1973			6	30/06/2006			38	1	26/06/2006 13:29
5016	F	17 # 2007		C	P	25/09/1948	19	1	1	23/06/2006			44	44	27/11/2006
18	M	2da No 606 INT-108	4100101	C	L	30/09/1951			5	25/05/2005	N		8	8	03/01/2008 11:01
4765	F	16 #2706		C	S	03/03/1984	1	515	3	24/05/2006			21	1	04/09/2006 12:26
5128	F	27 1/2 #1312		C	P	27/02/1973	19	1	5	07/07/2006			38	18	11/12/2006 10:39
607	M	PLAZA ALAMEDA No 1931	483-37-82	C	L	20/09/1982			1	13/10/2005			19	19	10/10/2005 16:16
440	F	PROLONGACION SONORA NO. 74	52-32349	C	C	09/07/1961			4	24/08/2005	B	M	41	41	24/08/2005 10:25
4927	F	PUERTA BLANCA No 17302		C	C	05/04/1966			4	15/06/2006			37	37	12/06/2006 11:43
1124	M	HACIENDA EL ENCANTO # 10225		D	B	#########				16/11/2005 08:57			24	32	14/11/2005 12:20
29	F	DOROLFO ARANGO No 14719	4844476	C	S	12/08/1967			3	27/05/2005			19	19	30/05/2005 16:05
4769	F	GUILLERMO VALENCIA NO 3704		C	S	25/07/1980	19	1	1	25/05/2006			18	18	11/09/2006 12:20
4432	F	CALLE BAVICORA 163		C	S	08/08/1957			4	26/03/2006			32	32	18/12/2006 13:32
4310	F	RUBEN JARAMILLO No 44		C	C	06/04/1960			1	10/03/2006			19	19	13/03/2006 13:15
1997	M	PASEO DE LOS TREBOLES 8471		S	B	11/10/1973 12:12				##########			32	32	17/01/2006 18:00
4780	F	CESAR BAEZA No 14713		C	P	13/08/1972			2	25/05/2006			18	18	05/06/2006 17:18
5282	M	LAZARO CARDENAS No 238		C	B	25/02/1970			3	24/07/2006			17	17	24/07/2006 11:54

Tabla A0-1. Continuación.

PERIODICIDAD	CANTAUTOR	CANTENTRE	TASAINI	DURACINI	TASARECFIJ	EMPLSOSTH	EMPLSOSTM	MODOAPLIRECA	TASA	ABONOS	CANTENTRE_1	SALDO_TOTAL	FECHA_TERMINO_PRESTAMO
S	4952	4952	4	16	8	0	1	1	4	2440	4952	4862.25	28/11/2005
M	266.79	266.79	0	12		0	1	2	0	266.79	266.79	926.44	01/01/2011
Q	10000	10000	5.75	12	8	0	1	1	5.75	10759.7	10000	0	30/03/2006
S	2075	2075	6.1	12	8	0	1	1	6.1	1339.56	2075	0	23/10/2006
S	2075	2075	6.1	12	8	0	1	1	6.1	2239	2075	528.43	23/10/2006
S	2100	2100	5.75	16	5	0	1	1	5.75	2111	2100	-19.7	08/01/2007
Q	30300	30300	4.6	12	5	1	0	1	4.6	6311.4	30300	-19.89	16/01/2008
S	3125	3125	5.5	20	5	0	1	1	5.5	4105.5	3125	229.04	19/03/2007
S	7000	7000	6.1	20	8	1		1	6.1	2270	7000	0	24/07/2006
S	10000	10000	5.75	26	8	0	1	1	5.75	7570	10000	1946.92	19/12/2005
S	4100	4100	5.5	16	8	0	1	1	5.5	3864.63	4100	-0.77	30/10/2006
S	5150	5150	5	24	5	0	1	1	5	7093	5150	323.73	21/05/2007
M	23275	23275	3.45	2	5	1		2	3.45	50	23275	-61.1	03/03/2008
S	5075	5075	4.6	12	5		1	1	4.6	3948.4	5075	7352.97	04/12/2006
S	5150	5150	5.5	24	5		1	1	5.5	3222	5150	0	04/06/2007
S	10000	10000	5.75	20	8	1		1	5.75	12081	10000	262.62	09/03/2006
Q	9000	9000	6.1	12	8		1	1	6.1	10735	9000	8680.85	28/02/2006
S	4125	4125	5.9	20	8		1	1	5.9	3755	4125	9679.29	06/11/2006
S	8000	8000	6	16	8	1		1	6	9375	8000	0	13/03/2006
S	5000	5000	5.9	16	8		1	1	5.9	3271.25	5000	-36.8	30/09/2005
S	3100	3100	5.75	16	5		1	1	5.75	4097	3100	4989.29	08/01/2007
S	20175	20175	5	28	5		1	1	5	27988	20175	-0.6	02/07/2007
S	3000	3000	6.1	12	8		1	1	6.1	3740	3000	-13.69	12/06/2006
S	6000	6000	6	12	8	1		1	6	7200	6000	-1.03	17/04/2006
S	3125	3125	5.9	20	8		1	1	5.9	3601	3125	9.29	30/10/2006

Tabla A0-2. Muestra ejemplos de los datos del dataset después del pretratamiento.

CODIGO	sex	TELEFONO	edocivil	nivel	nacimien	NACIOMU	NODEPEND	alta	restriccion	regmarital	CDGPRPE	CDGOCPE	solicitud	periodicidad	CANTAUTOR	CANTENTRE	TASAINI
168	1	1	0.4	0.556	0.143	0.422	0.167	0.5	0.25	1	0.324	0.873	0.5	1	0.0617	4.06E-09	0.25
5224	1	2	0.2	0.667	0.429	0.822	0	0.5	1	0.667	0.451	0.451	0.75	0.667	0.2046	0.2832	1
498	1	1	0.2	0.778	0.571	0.022	0.208	0.5	1	0.333	1	0.62	0	0.333	0.1006	0.2713	1
5266	1	2	0.2	0.222	0.286	0.422	0.167	0.5	1	0.667	0.324	0.324	0.75	0.333	0.101	0.294	1
5259	1	2	0.2	0.222	0.857	0.467	0.083	0.75	1	0.667	0.451	0.451	0.75	0.333	0.1338	0.2832	1
4853	2	2	0.2	0.222	0.286	0.711	0.125	0.5	0.5	0.667	0.028	0.028	0.5	1	0.2978	2.96E-05	0.5
4688	2	2	0.2	0.556	0.286	0.422	0	0.5	1	0.667	0.338	0.451	0.75	0.667	0.236	0.2832	1
4973	1	1	0.2	0.111	0.571	0.822	0.167	1	1	0.333	0.549	0.69	1	0.333	0.0846	0.294	1
4200	1	1	0.4	0.889	0.143	0.022	0.083	0.75	1	1	0.197	0.197	1	0.333	0.114	0.0256	1
91	1	1	0.2	0.111	0.143	0.467	0.208	0.5	0.25	0.667	0.789	0.07	0	1	0.0515	4.06E-09	0.25
5061	1	1	0.4	0.556	0.143	0.467	0.125	0.75	0.75	1	0.577	0.577	1	0.667	0.2019	0.2515	0.75
5016	1	2	0.2	0.111	0.286	0.422	0.208	0.5	1	0.667	0.324	0.324	0.75	0.333	0.0722	0.294	1
18	1	1	0.2	0.111	0.571	0.422	0.208	0.5	1	0.333	0.324	0.324	0.75	0.333	0.0722	0.294	1
4765	1	2	0.2	0.222	0.571	0.711	0.042	0.5	1	0.667	0.254	0.254	0.75	0.333	0.0726	0.2515	1
5128	2	1	0.2	0.444	0.571	0.711	0.208	0.5	0.25	0.667	0.592	0.592	0.5	0.667	0.3525	0.0782	0.25
607	1	1	0.2	0.111	1	0.467	0.167	0.5	1	0.667	0.789	0.577	0.75	0.333	0.0863	0.2143	1
440	2	2	0.4	0.222	0.714	0.467	0.083	0.75	0.5	1	0.549	0.549	1	0.333	0.1486	0.294	0.5
4927	1	1	0.6	0.889	0.286	0.422	0.167	0.75	1	1	0.549	0.549	1	0.333	0.2019	0.2515	1
1124	1	2	0.2	0.111	0.286	0.422	0.25	0.5	1	0.333	0.535	0.014	0.75	0.333	0.1006	0.2143	1
29	1	1	0.2	0.111	1	0.422	0.042	0.5	1	0.333	0.62	0.62	0.75	0.333	0.1172	0.1342	1
4769	2	1	0.2	0.444	0.857	0.822	0.208	0.75	0.5	0.667	0.113	0.113	0.5	1	0.3425	0.0064	0.5
4432	1	2	0.2	0.222	0.143	0.022	0.125	0.5	1	0.667	0.296	0.014	0.75	0.333	0.1159	0.0782	1
4310	1	2	0.2	0.111	0.286	0.422	0.208	0.5	1	0.667	0.535	0.254	0.75	0.333	0.1172	0.2143	1
1997	2	1	0.2	0.444	0.143	0.022	0.042	0.75	1	0.333	0.268	0.268	1	0.333	0.2019	0.2515	1
4780	1	1	0.2	0.556	0.571	0.467	0.167	0.75	0.75	0.667	0.577	0.577	1	0.667	0.1842	0.294	0.75
5282	1	2	0.2	0.556	0.429	0.711	0.167	0.5	1	0.667	0.521	0.521	0.75	0.333	0.101	0.2713	1

Tabla A0-2. Continuación

DURACINI	TASARECFIJ	EMPLSOSTH	EMPLSOSTM	MODOAPLIRECA	ABONOS	abono int	abonoint	SALDO_TOTAL	FECHA_TERMINO_PRESTAMO
0.2766	0.2	0	1	2	0.0674	0.6392	0.093	0.110923033	28/11/2005
0.2766	0.4	0	1	1	0.2472	0.5488	0.07	0.146186137	01/01/2011
0.3617	1	0	1	1	0.1246	1.146	0.204	0	30/03/2006
0.4468	1	0	1	1	0.1067	0.5917	0.081	0	23/10/2006
0.5319	0.4	0	1	1	0.1608	0.6303	0.091	0.367169447	23/10/2006
0.0426	0.4	1	0	1	0.0674	0.6246	0.089	-0.62820229	08/01/2007
0.2766	1	1	0	1	0.2189	0.8388	0.142	-0.29516011	16/01/2008
0.3617	1	0	1	1	0.1094	0.7185	0.113	0.194457885	19/03/2007
0.3617	1	0	1	1	0.0931	0.6317	0.091	0	24/07/2006
0.2766	0.2	0	1	2	0.07	0.9162	0.159	0.192568564	19/12/2005
0.2766	1	0	1	1	0.2038	0.7065	0.11	-0.2126839	30/10/2006
0.2766	1	0	1	1	0.081	0.8861	0.153	0.27449412	21/05/2007
0.2766	1	0	1	1	0.0908	0.5405	0.068	-0.62801678	03/03/2008
0.3617	0.4	0	1	1	0.0894	0.7106	0.111	0.139225844	04/12/2006
0.2766	0.4	1	0	1	0.1421	0.6753	0.102	0	04/06/2007
0.4468	0.4	0	1	1	0.1131	1.2575	0.222	0.237007721	09/03/2006
0.4468	1	1	0	1	0.0912	1.1441	0.204	4.970223698	28/02/2006
0.5745	1	0	1	1	0.1594	0.701	0.109	0.778805545	06/11/2006
0.3617	1	0	1	1	0.1101	1.0399	0.185	0	13/03/2006
0.5319	0.4	0	1	1	0.1528	0.6776	0.103	-0.13204222	30/09/2005
0.0638	0.4	1	0	2	0.0679	0.7181	0.113	0.123314239	08/01/2007
0.2766	0.4	0	1	1	0.1112	3.8401	0.344	-0.45275329	02/07/2007
0.5319	0.4	0	1	1	0.1023	0.7003	0.109	-0.10265189	12/06/2006
0.4468	1	1	0	1	0.2217	0.8928	0.154	-0.18009142	17/04/2006
0.2766	1	0	1	1	0.2035	0.6935	0.107	0.110923033	30/10/2006
0.4468	1	0	1	1	0.1088	0.7529	0.122	0.146186137	18/12/2006

Tabla A0-3. Eigenvalores extraidos y porcentaje de trazo para PCA.

	Eigenvalores Extraidos	Porcentaje de trazo
1	5.264	29.289
2	4.321	23.733
3	3.677	16.715
4	2.949	12.857
5	1.264	3.565
6	0.821	2.814
7	0.676	2.455
8	0.611	2.049
9	0.542	1.564
10	0.474	1.367
11	0.416	1.273
12	0.352	0.98
13	0.185	0.404
14	0.165	0.395
15	0.107	0.274
16	0.085	0.167
17	0.052	0.092
18	0.033	0.066
19	0.009	0.014
20	0.003	0.008
21	0	0
22	0	0
23	0	0
24	0	0
25	0	0
26	0	0
27	0	0
28	0	0
29	0	0

B

Anexo B. Procedimiento de calibración de parámetros para PSO.

La selección de los mejores parámetros de PSO es una tarea definidas como de selección de modelo. Afortunadamente, se han realizado varios estudios empíricos y teóricos acerca del cálculo de estos parámetros de la que puede obtenerse información útil.

Los parámetros w, φ_1 y φ_2 caracterizan el comportamiento de las partículas, y la experiencia demuestra que el éxito o el fracaso de la búsqueda dependen en gran medida de los valores de estos parámetros. Las principales causas de los fracasos de búsqueda son las que se muestran a continuación:

1. La velocidad de las partículas aumenta rápidamente, y las partículas se mueven fuera del espacio de búsqueda.
2. La velocidad de las partículas disminuye rápidamente, y convertirse en partículas inmóviles.
3. Las partículas no pueden escapar de soluciones localmente óptimas.

Varios de los estudios realizados (Blackwell, et al., 2008; Blum & Li, 2008; Clerc & Kennedy, 2002; Y. Shi & Eberhart, 1998) han demostrado qué en PSO el equilibrio entre las capacidades globales y locales de exploración es controlada principalmente por el peso de inercia. Al fijar la velocidad máxima permitida en 2, se encontró que los PSO con una valor inercial en el intervalo [0,9, 1,2], en promedio, tiene un mejor rendimiento, es decir, como una oportunidad más grande para encontrar el óptimo global en un plazo razonable número de iteraciones. Por otra parte, la disminución del peso inercial en el tiempo de 1,4-0 se encuentra que es mejor que una masa de inercia fija. Esto es debido a los pesos de inercia más grandes en el comienzo ayudan a encontrar buenas semillas y más tarde los pequeños pesos inercia facilitar la búsqueda fina.

Estas recomendaciones y otras adicionales se toman en cuenta como punto de partida para calibrar los parámetros de PSO, y posteriormente mediante un análisis empírico se pueden obtener los valores idóneos para cada problema en particular que sea resuelto a través del uso de PSO. Las siguientes tablas muestran los resultados de optimización para el modelado de puntuación crediticia mediante agrupamientos, un proceso similar se desarrolló para la calibración de parámetros en el empleo de PSO en los otros modelos.

La tabla B0-1 muestra los valores obtenidos para la función de aptitud variando el tamaño de enjambre y tomando como base los valores definidos por la literatura para ω= [0.9, 0.4], φ_1= 2, φ_2=*2*, condición de paro 1500 iteraciones y 30 simulaciones.

Tabla B0-1. Valor de la función objetivo empleando diferentes tamaños de enjambre.

	Enjambre					
Simulación	20	25	30	35	40	45
1	8.3276	8.2489	13.251	17.38	15.545	15.856
2	17.675	13.182	17.14	12.968	14.329	20.886
3	11.213	14.981	14.733	14.97	19.974	18.79
4	18.191	9.0681	13.914	9.1023	11.783	12.763
5	9.4672	18.537	18.168	21.788	19.481	18.165
6	5.3151	18.957	21.667	9.0294	11.148	20.186
7	5.8871	19.891	16.596	12.391	19.904	14.55
8	16.858	11.865	20.694	9.8615	16.64	18.001
9	7.5763	7.8451	18.867	13.388	15.551	12.004
10	10.606	12.346	10.785	11.686	16.749	20.686
11	6.4655	23.94	12.377	12.059	14.658	19.972
12	17.104	17.064	11.745	20.234	15.158	15.351
13	9.1943	9.2512	11.541	8.884	12.78	16.45
14	17.411	18.82	17.586	18.707	11.817	15.949
15	5.7352	19.449	14.113	11.507	19.833	20.674
16	5.8708	13.525	12.314	15.653	17.503	12.529
17	7.7866	11.036	11.848	21.262	18.039	22.376
18	9.8698	21.077	19.659	13.818	16.5	21.175
19	8.6705	19.647	21.572	21.548	14.425	12.616
20	9.7895	17.718	11.88	13.399	17.327	19.907
21	10.821	11.574	14.386	20.258	16.287	14.43
22	15.545	16.936	22.286	14.163	16.862	13.188
23	7.3205	11.666	16.116	18.523	14.087	12.663
24	9.4648	12.377	21.297	9.3771	18.923	14.504

	Enjambre					
25	8.733	8.0633	18.274	17.822	17.099	14.976
26	17.951	21.567	21.932	12.786	11.423	20.17
27	17.228	23.135	16.082	9.2226	11.913	21.785
28	13.465	13.436	18.741	16.705	13.559	19.222
29	13.472	10.691	11.912	11.535	13.203	15.977
30	7.8035	7.356	14.944	8.8512	11.507	22.022
Promedio	11.027	14.775	16.214	14.296	15.467	17.261

Con el tamaño de enjambre ya definido se procede en este caso al cálculo de los parámetros de conocimiento individual y conocimiento grupal, manteniendo el resto de los valores inalterados, los resultados se muestran en las tablas B0-2 y B0-3 respectivamente.

Tabla B0-2 . Valor de la función objetivo empleando diferentes valores de φ_1.

	Grado de conocimiento individual φ_1											
Simulación	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3
1	7.195	14.53	14.32	16.23	15.58	11.36	18.75	23.37	11.37	21.66	12.59	10.32
2	11.17	18.11	13.82	24.01	11.76	12.21	19.17	11.26	16.27	14.42	10.48	17.27
3	15.8	18.58	12.5	11.74	13.03	24.26	14.53	10.16	15.03	20.47	18.05	20.43
4	15.81	18.08	16.7	23.4	19.73	20.2	12.45	14.56	14.98	12.19	12.27	27.4
5	14.23	19.24	12.78	14.24	14.74	22.13	6.385	10.07	9.855	10.25	30.67	9.289
6	13.27	17.61	18.04	10.52	18.37	14.55	12.56	13.15	18.7	15.05	9.972	23.92
7	17.03	16.24	16.98	10.99	13.38	21.6	18.28	18.99	9.356	13.28	9.344	16.83
8	14.27	13.6	14.02	12.66	14.09	16.23	6.636	17.48	10.07	10.18	20.42	24.14
9	9.219	11.06	18.04	19.89	19.89	23.59	17.69	15.67	15.93	15.93	6.658	16.11
10	16.24	18.11	15.56	15.35	13.57	23.58	13.99	18.61	18.13	21.06	16.17	25.1
11	10.22	16.29	16.95	20.81	16.01	25.06	11.39	28.81	16.67	15.99	21.7	24.62
12	20.72	14.98	17.52	10.96	22.32	19.42	13.68	24.52	18.09	12.08	23.24	11.16
13	7.579	7.301	15.13	11.32	11.4	23.96	11.54	15.85	14.87	14.84	10.66	13.95
14	7.672	16.01	12.93	15.98	17.17	26.9	9.77	14.14	10.58	17.79	16.99	24.02
15	10.12	8.662	13.65	10.45	11.88	11.76	10.67	17.58	9.28	16.43	29.21	27.91
16	10.27	17.17	16.71	24.08	14.25	19.19	12.14	12.81	16.26	11.15	16.32	23.64
17	8.603	14.14	16.55	10.3	13.67	21.68	6.438	14.04	9.421	16.22	15.12	24.09
18	13.81	10.71	15.01	15.54	11.4	22.85	12.1	25.66	14.57	19.43	20.79	26.53
19	14.83	14.46	15.96	9.786	12.55	10	7.571	22.36	12.63	13.21	14.54	22.45
20	9.676	13.82	12.51	22.12	16.06	19.65	17.93	22.85	11.77	23.73	14.3	8.394
21	13.7	19.71	13.77	12.31	22.79	12.39	14.75	11.97	18.88	17.57	11.7	21.93
22	12.45	15.1	13.4	16.16	15.76	19.99	14.31	9.879	11.63	21.32	28.9	18.33
23	13.35	10.35	16.54	20.89	19.67	12.15	13.05	18.57	18.79	22.46	22.47	17.1

	Grado de conocimiento individual φ_1											
24	20.94	15.12	15.52	21	17.59	17.02	12.22	19.71	16.58	21.19	10.16	18.63
25	7.282	8.195	15.09	20.8	14.99	23.36	14.95	17.44	14.09	24.13	19.55	17.45
26	13.7	18.24	18.75	17.32	18.82	25.34	9.873	22.56	10.63	12.18	26.8	19.05
27	13.2	13.23	18.79	15.42	15.54	19.69	6.755	13.35	12.08	24.08	25.99	17.64
28	12.61	8.819	13.88	20.15	20.02	20.48	9.822	18.49	10.68	17.12	21.62	21.48
29	16.3	15.48	13.26	20.42	19.04	20.93	11.09	25.95	16.58	10.45	23.62	7.216
30	15.35	10.72	13.37	20.82	16.25	24.3	15.03	16.53	13.5	14.65	7.623	26.36
Promedio	12.89	14.46	15.27	16.52	16.04	19.53	12.52	17.55	13.91	16.68	17.6	19.43

Tabla B0-3. Valor de la función objetivo empleando diferentes valores de φ_2.

	Grado de conocimiento grupal φ_2											
Simulación	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3
1	11.28	16.89	17.78	13.37	18.03	25.64	8.129	28.81	8.254	24.55	12.84	14.09
2	14.96	18.55	15.31	21.73	16.16	26.88	14.04	17.12	18.35	23.73	21.21	9.627
3	12.61	9.124	17.05	17.27	22.54	22.74	18.65	15.45	14.77	12.69	8.222	20.21
4	14.03	7.93	16.4	23.76	12.06	12.06	18.66	21.81	15.74	11.77	11.17	18.1
5	13.85	19.68	14.59	12.78	11.58	18.57	12.89	25.22	13.42	17	17.8	21.68
6	7.098	8.287	17.18	17.71	17.83	16.3	7.603	13.99	11.79	21.17	21.73	19.17
7	12.31	19.3	17.12	10.06	17.23	16.21	17.87	12.15	12.33	22.71	27.54	24.43
8	20.28	16.31	15.06	22.2	11.41	12.29	16.47	24.76	11.1	22.98	10.55	17.17
9	11.42	14.38	15.89	23.44	14.68	20.19	14.31	16.05	17.42	17.65	17.59	20.16
10	6.863	16.48	14.82	14.06	14.13	15.4	16.9	11.03	10.13	23.56	30.72	12.87
11	6.322	16.52	17.46	24.86	11.85	23.36	7.69	17.33	12.23	10.28	27.73	12.66
12	14.84	19.04	14	11.74	11.09	10.2	10.38	28.93	9.869	17.29	28.4	24.42
13	20.85	10.32	13.33	12.77	12.24	16.14	11.74	10.71	13.28	11.39	24.09	10.73
14	6.959	14.29	18.47	19.14	15.89	17.9	10.06	26.86	10.35	11.67	14.7	26.03
15	7.081	9.277	14.58	10.81	11.76	10.4	11.55	20.34	16.87	24.14	10.66	19.32
16	16.1	13.67	13.41	12.11	13.25	23.49	11.94	15.09	10.85	20	9.666	9.772
17	6.756	15.18	18.46	12.31	17.13	15.1	16.96	23.79	17.72	24.08	23.17	21.72
18	8.709	8.631	18.22	18.66	14.03	25.13	7.681	28.83	14.45	15.01	15.14	16.97
19	7.117	8.87	15.79	13.17	14.16	22.26	9.913	23.4	14.44	21.6	9.043	19.09
20	12.51	12.08	17.45	12.7	17.39	16.79	13.13	28.54	12.68	19.96	23.56	14.13
21	14.64	15.56	18.7	17.01	11.45	23.17	11.17	10.87	13.13	17.34	6.497	21.18
22	9.328	11.32	13.11	9.964	20.57	25.78	8.194	16.16	14.2	14.65	19.08	19.49
23	13.99	16.03	12.99	16.02	12.96	15.96	17.43	22.66	10.32	13.54	7.749	7.742
24	16.06	11.87	12.11	18.45	16.49	10.14	11.91	14.09	13.3	11.43	8.424	9.129
25	15.78	14.51	18.35	20.97	16.04	26.74	10.07	26.02	10.4	20.09	22.99	19.98
26	14.01	7.273	13.22	14.31	21.78	10.41	12.23	19.16	17.24	10.93	16.8	19.82
27	9.978	17.71	15.63	23.71	11.64	14.4	18.8	19.72	16.19	19.32	11	15.76

	Grado de conocimiento grupal φ_2											
28	13.23	9.224	13.73	18.08	18.46	10.9	10.38	24.3	17.06	18.42	14.92	26.25
29	7.797	18.01	12.05	15.32	14.31	13.73	19.27	20.1	12.67	15.88	12.57	10.86
30	18.13	12.15	16.48	12.77	21.11	16.88	18.59	21.14	16.86	12.54	20.59	15.32
Promedio	12.16	13.62	15.62	16.38	15.31	17.84	13.15	20.15	13.58	17.58	16.87	17.26

La tabla B0-4 muestra el número de iteraciones a las cuales converge el algoritmo, empleando 1500 iteraciones de prueba, y los valores predefinidos y calculados para el resto de los parámetros.

Tabla B0-4 .Cantidad de iteraciones requeridas para alcanzar el punto de convergencia.

Simulación	Iteración	Simulación	Iteración
1	911	14	1088
2	978	15	1097
3	1122	16	867
4	1035	17	1043
5	1073	18	940
6	851	19	938
7	1084	20	905
8	1063	21	1044
9	955	22	1109
10	896	23	976
11	995	24	948
12	1096	25	983
13	1080		
		Promedio	1003.08

Los resultados obtenidos y evaluando el mejor comportamiento de la función de aptitud se puede definir el mejor valor para cada uno de los parámetros del modelo. Es evidente que la obtención de parámetros es totalmente empírica y falta sustentarla mediante un diseño de experimentos para implementar técnicas estadísticas que permitan validar los valores de los parámetros.

C

Anexo C. Extractos de código en Java, para la implementación del modelo.

Tabla C0-1.Base para calcular PCA empleando Efficient Java Matrix Library.

```
/**
    * Computes a basis (the principle components) from the most dominant eigenvectors.
    *
    */
   public void computeBasis( int numComponents ) {
      if( numComponents > A.getNumCols() )
         throw new IllegalArgumentException("Más components requeridos que la longitud
         de de datos.");
      if( sampleIndex != A.getNumRows() )
         throw new IllegalArgumentException("No se han agergado todos los datos");
      if( numComponents > sampleIndex )
         throw new IllegalArgumentException("Se requieren más datos para calcular la cantidad de
        components requeridos");

      this.numComponents = numComponents;

      // compute the mean of all the samples
      for( int i = 0; i < A.getNumRows(); i++ ) {
         for( int j = 0; j < mean.length; j++ ) {
            mean[j] += A.get(i,j);
         }
      }
      for( int j = 0; j < mean.length; j++ ) {
         mean[j] /= A.getNumRows();
      }

      // subtract the mean from the original data
      for( int i = 0; i < A.getNumRows(); i++ ) {
         for( int j = 0; j < mean.length; j++ ) {
            A.set(i,j,A.get(i,j)-mean[j]);
         }
      }

      // Compute SVD and save time by not computing U
      SingularValueDecomposition<DenseMatrix64F> svd =
            DecompositionFactory.svd(A.numRows, A.numCols, false, true, false);
      if( !svd.decompose(A) )
```

```
            throw new RuntimeException("SVD falló");

        V_t = svd.getV(null,true);
        DenseMatrix64F W = svd.getW(null);

        // Singular values are in an arbitrary order initially
        SingularOps.descendingOrder(null,false,W,V_t,true);

        // strip off unneeded components and find the basis
        V_t.reshape(numComponents,mean.length,true);
    }
```

Tabla C0-2. Actualización de pbest

```
for(int i=0; i<SWARM_SIZE; i++) {
                                    if(fitnessValueList[i] < pBest[i]) {
                                            pBest[i] = fitnessValueList[i];
                                            pBestLocation.set(i, swarm.get(i).getLocation());
                                    }
                            }
```

Tabla C0-3. Actualización de gbest

```
int bestParticleIndex = PSOUtility.getMinPos(fitnessValueList);
                            if(t == 0 || fitnessValueList[bestParticleIndex] < gBest) {
                                    gBest = fitnessValueList[bestParticleIndex];
                                    gBestLocation = swarm.get(bestParticleIndex).getLocation();
                            }
```

Tabla C0-4. Cálculo de inercia para PSO-mejorado.

```
w = W_UPPERBOUND - (((double) t) / MAX_ITERATION) * (W_UPPERBOUND - W_LOWERBOUND);
```

Tabla C0-5. Movimiento de partículas

```
for(int i=0; i<SWARM_SIZE; i++) {
        double r1 = generator.nextDouble();
        double r2 = generator.nextDouble();

        Particle p = swarm.get(i);

        // actualización de velocidad
        double[] newVel = new double[PROBLEM_DIMENSION];
        newVel[0] = (w * p.getLocation().getLoc()[0]) +  (r1 * C1) *
                        (pBestLocation.get(i).getLoc()[0] - p.getLocation().getLoc()[0]) +
                        (r2 * C2) * (gBestLocation.getLoc()[0] - p.getLocation().getLoc()[0]);
        newVel[1] = (w * p.getLocation().getLoc()[1]) +  (r1 * C1) *
                        (pBestLocation.get(i).getLoc()[1] - p.getLocation().getLoc()[1]) +
                        (r2 * C2) * (gBestLocation.getLoc()[1] - p.getLocation().getLoc()[1]);

        Velocity vel = new Velocity(newVel);
        p.setVelocity(vel);

        // actualización de posición
        double[] newLoc = new double[PROBLEM_DIMENSION];
        newLoc[0] = p.getLocation().getLoc()[0] + newVel[0];
        newLoc[1] = p.getLocation().getLoc()[1] + newVel[1];
        Location loc = new Location(newLoc);
        p.setLocation(loc);
}
```

Tabla C0-6. Planteamiento de problema de puntuación crediticia ensamblada.

```
public class ProblemSet {
	public static final double LOC_X_LOW = 1;
	public static final double LOC_X_HIGH = 4;
	public static final double LOC_Y_LOW = -1;
	public static final double LOC_Y_HIGH = 1;
	public static final double VEL_LOW = -1;
	public static final double VEL_HIGH = 1;

	public static final double ERR_TOLERANCE = 1E-20;

	public static double evaluate(Location location, Weight weight) {
		double result = 0;
		double w1 = weight.getWeig()[0]; // peso de MLR
		double w2 = weight.getWeig()[0]; // peso de LR
		double x = location.getLoc()[0]; // Puntuación de MLR
		double y = location.getLoc()[1]; // Puntuación  de LR

		result = w1*x+w2*y

		return result;
	}
}
```

D

Anexo D. Procedimiento de Puntuación crediticia ensamblada.

El procedimiento de cálculo de puntuación crediticia de manera simplificada se explica a continuación: sean 12 prestatarios mostrados en la tabla D0-1.

Tabla D0-1.Muestra de prestatarios para ejemplificar el procedimiento de puntuación ensamblado.

x_2	x_7	x_{11}	x_{12}	x_{13}	x_{16}	x_{18}	x_{21}	x_{22}	x_{23}	x_{27}	Y
2	0.714	1	0.333	0.183	0.333	0.134	0.4	1	0	0	0
1	0.429	1	1	0.183	0.333	0.283	1	0	1	0	0
1	1	1	1	0.197	0.333	0.271	1	0	1	0.588	1
1	0.286	1	0.333	0.69	0.333	0.294	1	0	1	-0.411	1
2	0.571	1	0.667	0.197	0.333	0.294	1	1	0	0.221	0
1	0.286	0.75	0.667	0.127	0.333	0.252	0.4	0	1	0.947	0
2	0.571	1	0.667	0.268	0.333	0.252	1	1	0	0.274	0
1	0.429	1	1	0.451	1	0.214	1	0	1	-0.373	1
1	0.429	0.75	1	0.592	0.667	0.294	1	0	1	-0.748	1
2	0.714	1	1	0.085	1	0.006	0.4	1	0	0	1
1	0.714	1	0.667	0.268	0.333	0.294	1	0	1	0.053	0
1	1	1	1	0.127	0.667	0.214	0.4	0	1	0.705	1

Los cuales se puede observar que 6 son buenos clientes y los 6 restantes son malos pagadores (la última columna define a los buenos pagadores con 0 y a los malos clientes con 1), para tener un control para emplear el proceso de ensamble mediante bagging. 6 de ellos se usan para el método MLR y los 6 restantes para LR, con una relación 1:1 entre buenos y malos prestatarios, como lo ilustra la tabla D0-2 para el método MLR y la tabla D0-3 para el modelo LR.

Tabla D0-2. Muestra ejemplo de prestaratrios para modelo MLR.

x_2	x_7	x_{11}	x_{12}	x_{13}	x_{16}	x_{18}	x_{21}	x_{22}	x_{23}	x_{27}	Y
2	0.714	1	0.333	0.183	0.333	0.134	0.4	1	0	0	0.3902
1	0.429	1	1	0.183	0.333	0.283	1	0	1	0	0.4465
1	1	1	1	0.197	0.333	0.271	1	0	1	0.588	0.6045
1	0.286	1	0.333	0.69	0.333	0.294	1	0	1	-0.411	0.7797
2	0.571	1	0.667	0.197	0.333	0.294	1	1	0	0.221	0.4915
1	1	1	1	0.127	0.667	0.214	0.4	0	1	0.705	0.5595

Tabla D0-3. Muestra ejemplo de prestaratrios para modelo LR.

x_2	x_7	x_{11}	x_{12}	x_{13}	x_{16}	x_{18}	x_{21}	x_{22}	x_{23}	x_{27}	Y
1	0.286	0.75	0.667	0.127	0.333	0.252	0.4	0	1	0.947	0.41574
2	0.571	1	0.667	0.268	0.333	0.252	1	1	0	0.274	0.32988
1	0.429	1	1	0.451	1	0.214	1	0	1	-0.373	0.82899
1	0.429	0.75	1	0.592	0.667	0.294	1	0	1	-0.748	0.75011
2	0.714	1	1	0.085	1	0.006	0.4	1	0	0	0.89749
1	0.714	1	0.667	0.268	0.333	0.294	1	0	1	0.053	0.36315

A los elementos de la tabla D0-2 se les calcula su puntuación crediticia aplicando la ecuación V.2, y se normaliza para obtener un valor entre 0 y 1 para valorar su puntuación con respecto al modelo MLR y calificarlo como buen o mal prestatario bajo el criterio de qué calificaciones por encima de 0.5 son buenos clientes y por debajo son malos clientes. El mismo procedimiento se realiza para los clientes de la tabla D0-3 pero empleando la ecuación V.4 con la finalidad de emplear el modelo LR, los resultados correspondientes se pueden observar en la tabla D0-4.

Tabla D0-4. Resultados esperados y pronosticados para MLR y LR.

Modelo MLR		*Modelo R*	
Y	$\hat{y}_1$	Y	$\hat{Y}_2$
0.3902	0.4046	0.41574	0.4312
0.4465	0.5662	0.32988	0.3975
0.4915	0.3234	0.36315	0.3603
0.6045	0.5352	0.82899	0.7232
0.7797	0.6287	0.75011	0.7090
0.5595	0.5987	0.89749	0.4477

Con estos valores se puede crear el ensamble empleando la ecuación IV.25, siendo las variables w_1 y w_2 valores supuestos que sirven como elementos de la partícula para emplear

PSO, así sí se suponen los valores de w_1 = 3.47 y w_2 = -2.47 (debido a la condición de la ecuación V.25 de $w_1 + w_2 = 1$), entrega la función de ensamble siguiente:

$$f_1 = 3.47 * 0.4046 - 2.47 * 0.4312 = 0.338898$$

La tabla D0-5 muestra los valores pronosticados empleando el valor de ensamble, recordando que los valores mostrados se encuentran normalizados y adecuados al criterio de clasificación empleado.

Tabla D0-5. Resultados obtenidos para el modelo ensamblado.

Modelo ensamblado	
Y	$\hat{y}_1$
0.3902	0.3388
0.4465	0.9828
0.4915	0.2322
0.6045	0.070
0.7797	0.4303
0.5595	0.97167

Estos valores son los que se emplean en el modelo PSO, teniéndose así que la función de aptitud de la ecuación IV.28 en este caso tendría los siguientes valores en sus variables

$$F = 100 * \left(\frac{1}{3} * ((0.3902 - 0.3389)^2 + (0.5595 - 0.9716)^2 + (0.5595 - 0.9716)^2) + 11 * \frac{1}{3} * ((0.4465 - 0.9829)^2 + (0.6045 - 0.07084)^2 + (0.7797 - 0.4004)^2)\right)$$

$$F = 653.513407$$

Este valor de la función de aptitud es el que se considera para actualizar p_{best} y p_{gbest} en caso de que se cumpla la condición de optimización en este caso el de minimizar, prosiguiendo con el proceso del algoritmo PSO actualizando la velocidad y posición de la partículas.

Experimento comparativo de entre métodos supervisados y semi-supervisados.

El funcionamiento de los modelos supervisados presenta problemas derivados del proceso de entrenamiento y la cantidad de elementos destinados a este fin, pero de igual forma el

éxito de sus resultados depende de los parámetros seleccionados para las funciones nucleares de los métodos, así como la puesta a punto de estos parámetros.

Con estas ideas en mente el experimento consistió en realizar la simulación de los métodos supervisados (MLR, LR ensamblado) y semi-supervisado (agrupamiento) con una menor cantidad de clientes de entrenamiento y modificando indiscriminadamente los valores de los coeficientes de las funciones objetivo de MLR (ecuación V.2) y LR (ecuación V.3). Y posteriormente realizar la comparativa con los resultados obtenidos en la hacer la comparativa con los resultados generados de los modelos en su estado ideal en cuanto a los valores de Error Tipo II y el ICC. Los resultados derivados de este procedimiento se muestran en las Figuras D0.1 y D0.2.

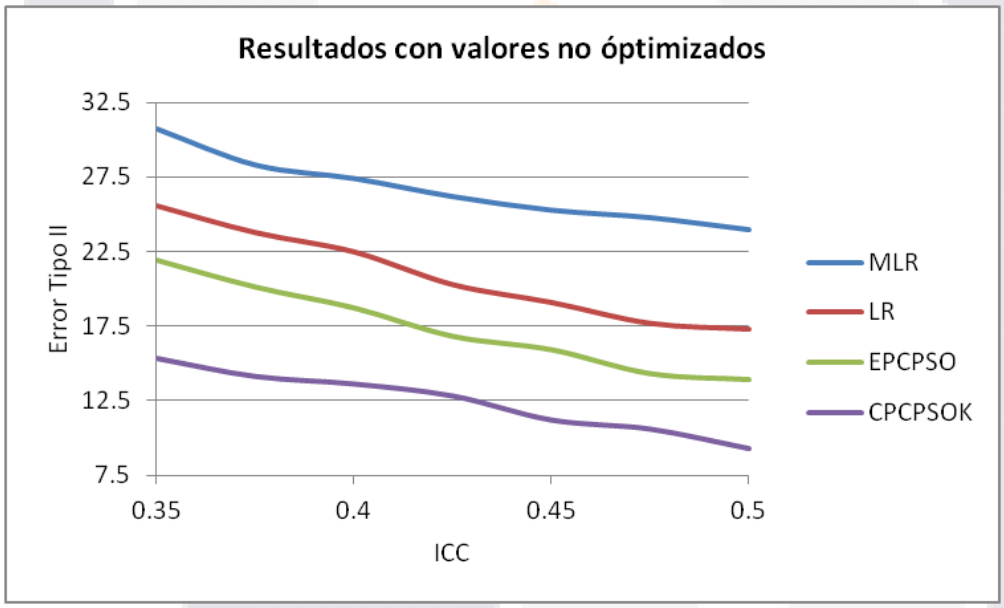


Figura D0-1.Valores de índice de clasificación correcta contra error tipo II en la aplicación de los modelos de puntuación crediticia empleando 100 clientes de entrenamiento y valores alterados en las ecuaciones V.2 y V.4

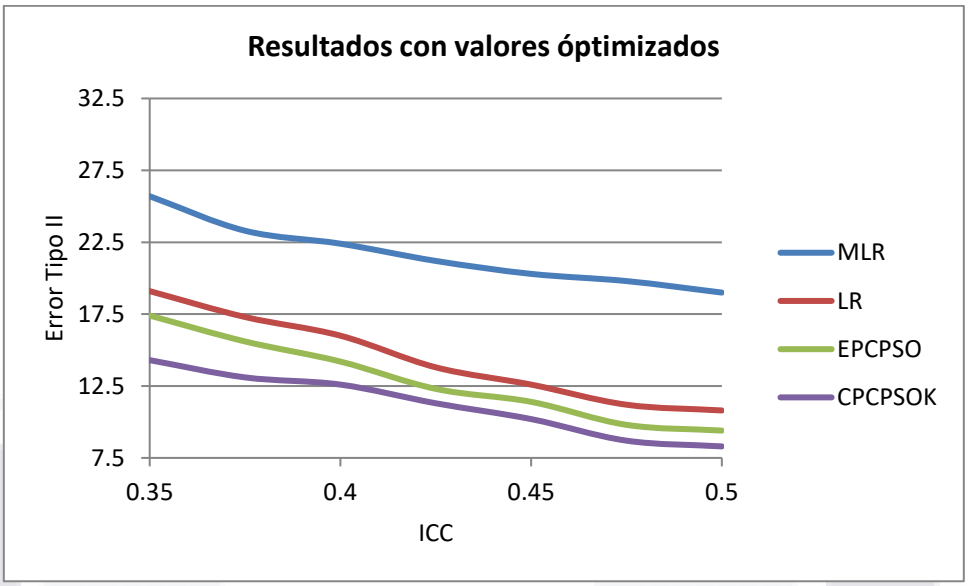


Figura D0-2. Valores de índice de clasificación correcta contra error tipo II en la aplicación de los modelos de puntuación crediticia con la calibración óptima empleada en el trabajo de tesis.

Los resultados experimentales muestran que aunque el modelo semi-supervisado tiene mejores resultados, no son tan grandes como pudiera esperarse.

Anexo E. Cartas de aceptación de trabajos realizados.

Cartas de aceptación de trabajos de autoría principal por orden de relevancia.

Paper Id:	HAIS11-SS01-1785
Title:	Outlier Analysis for Plastic Card Fraud Detection A Hybridized and Multi-objective Approach
Status:	**Tentatively accepted, subject to revision** **This paper has been tentatively accepted. Authors have to review the paper taking into account the reviewers´ comments.**
Abstract:	Nowadays, plastic card fraud detection is of great importance to financial institutions. This paper presents a proposal for an automated credit card fraud detection system based on the outlier analysis technology. Previous research has established that the use of outlier analysis is one of the best techniques for the detection of fraud in general. However, to establish patterns to identify anomalies, these patterns are learned by the fraudsters and then they change the way to make of fraud. The approach applies a multi-objective model hybridized with particle swarm optimization of typical cardholder?s behavior and to analyze the deviation of transactions, thus finding suspicious transactions in a non supervised scheme.
Sesion/Topic:	SPECIAL SESSION: HYBRID INTELLIGENT SYSTEM ON LOGISTICS AND INTELLIGENT OPTIMIZATION
Other Author:	Dr., Alberto, Ochoa, megamax8@hotmail.com

Uploaded Files

Type	Version	Name	Size (bytes)	Upload time	Download
Main	version 0	Paper_HAIS_Arturo_Elias.pdf	283496	2011-JAN-12 01:49:23	

Session Chair Comment

Date	Comment
2011-02-02	-

11:43:12	

REVISION 1	
REVISED	2011-01-31
RECOMENDATION	ACCEPT SUBJECT TO REVISION
COMMENT	The readability of the paper should be improved: 1.- The figures are not readable. The equations are not in format. 2.- The Section 3 does not make clear how the hypothesis are to be tested. The fitness functions are also not included. 3.- There is no experimentation stage, so interested readers can not understand the effectiveness of the proposal. How conclusions are extracted if no experiments are carried out?

REVISION 2	
REVISED	2011-01-31
RECOMENDATION	ACCEPT SUBJECT TO REVISION
COMMENT	An interesting approach. Numerical experiments-involving comparisons- are necessary to prove the system effectiveness and efficiency. language revision is also necessary.

Asunto ICTAM 2012 acceptance notification for paper 77
Remitente ICTAM 2012
Destinatario Arturo Elías
Fecha 2012-04-05 19:17

Thank you for your submission to ICTAM 2012. We are pleased to inform you that, according to the reports from anonymous reviewers, the following distinguished work from you has been accepted for ICTAM 2012, with the publisher of Lecture Notes in Information Technology (ISSN: 2070-1918), which will be indexed by ISTP, and submitted for indexing by EI.

You are kindly reminded with the following important notes:
1. Open the link and find a lot of information of paper submission and registration.
http://www.icmta-conf.com/reg

2. In order to make high quality of Proceedings, the camera-ready version should follow format. Kindly download from here
http://www.icmta-conf.com/reg

3. After finishing the final Paper, you can prepare a Copyright Release Form. The copyright should download, print, write author names, paper title, sign a name and date, and scanned it to PDF format
http://www.icmta-conf.com/reg/COPYRIGHT%20FORM.doc

4. Kindly download the registration form and pay for it, http://www.icmta-conf.com/reg and send both registration form and a scanned receipt from your bank to above Email ictam2012@163.com before April 20, 2012. If you have not paid for your paper in that time, your paper will not be published.

5. The e-copy official acceptance Letter could be download though http://www.icmta-conf.com/reg/Acceptance_letter.doc Write you write author names, paper title and print it. Kindly send Final paper (doc format), copyright, registration form and a scanned receipt to ictam2012@163.com before April 20, 2012.

Sincerely,
Program Committee of ICTAM 2012
E-mail: ictam2012@163.com
URL: http://www.icmta-conf.com

Answer the following question in scale (1-7, 1-low and 7 high)
Quality of the abstract: 4
Relevance to the conference: 5
Introduction and motivation: 5
Presentation of the "state of the art": 4
Description, originality of the own contribution: 5
Presentation of the results: 4
Conclusions and future work: 3
Readability, quality of the English: 4
Quality of the figures: 4

Quality of format: 4
Overall Paper Recommendation (1-7, 1 strong reject, 7 strong accept)
accepted as
regular paper: 5

Please modify your paper according to ICTAM 2012 Format strictly. Otherwise, we will not publish your paper in the proceedings. If you are not a native speaker (not familiar with in English environment), please check your sentences and/or English one more time to improve the quality of the final camera-ready paper.

NUMBER: 77
TITLE: Credit Scoring for Microfinance Institutions in México an ensemble and hybridized approach
AUTHOR: Arturo Elías

Asunto	Doctoral Consortium 2012
Remitente	Oscar Herrera
Destinatario	aeliasr@correo.uaa.mx
Fecha	2012-09-06 11:33

Dear Arturo Elías Ramírez,

We are pleased to inform you that your work has been accepted for oral presentation and publication in the proceedings of the Doctoral Consortium of MICAI 2012, to be held on (Tuesday) October 30, 2012 at San Luis Potosí, México.

Please consider the following suggestions/observations before sending us the final version of your paper (deadline: September 28):

- Follow the Springer LNCS instructions (as specified in http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0).

- It will be excellent if you can present some first results concerning the ontological framework.
- Be careful in fitting your paper into 4 pages.

Best regards,

Oscar Herrera and Miguel González

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

MC ARTURO ELIAS RAMIREZ
UNIVERSIDAD AUTONOMA DE AGUASCALIENTES
P R E S E N T E

Estimado(a) MC ARTURO ELIAS RAMIREZ:

Por este medio me permito informarle que su trabajo titulado "ALGORITMO INTELIGENTE PARA AGRUPAMIENTO DE CLIENTES EN MICROFINANCIERAS" fue ACEPTADO para participar en la modalidad CARTEL dentro del Tercer Congreso Internacional "La Investigación en el Posgrado" que se llevará a cabo los días 17, 18 y 19 de Octubre del presente año en la Unidad de Estudios Avanzados de la Universidad Autónoma de Aguascalientes.

Le informamos que se ofrecerán talleres sin costo con cupo limitado, por lo que le sugerimos estar al pendiente para que pueda registrarse con tiempo. Los talleres son:

-Redacción Científica
-Estrategias Docentes
-Investigación Cualitativa
-Uso y Manejo del SPSS en la investigación Científica

Sin más por el momento me despido de usted, aprovechando la oportunidad para enviarle un cordial saludo.

A T E N T A M E N T E
"SE LUMEN PROFERRE"
Aguascalientes, 30 de Agosto de 2012

DR. FERNANDO JARAMILLO JUÁREZ
DIRECTOR GENERAL DE INVESTIGACIÓN Y POSGRADO

F

Anexo F. Citas de artículos en bases de datos en Internet

Listado de citas de trabajos de autoría y coautoría reconocidos por las bases de datos DBLP y Google Académico.

Información contenida en DBLP

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Refine by AUTHOR
Arturo Elías(2) Julio César Ponce Gallegos(2) Francisco Ornelas(1) Arturo Hernández(1) [top 4] [all 10]
Refine by VENUE
HAIS(1) HIS(1)
Refine by YEAR
2011(2)

hide facet boxes

	2011
c2	Arturo Elías, Alberto Ochoa-Zezzatti, Alejandro Padilla, Julio Ponce: **Outlier Analysis for Plastic Card Fraud Detection a Hybridized and Multi-Objective Approach.** . HAIS (2) 2011: 1-9
c1	Alberto Ochoa, Julio Ponce, Rubén Jaramillo, Francisco Ornelas, Alberto Hernandez, Daniel Azpeitia, Arturo Elías, Arturo Hernández: **Analysis of Cyber-bullying in a virtual social networking.** HIS 2011: 229-234

Información contenida en Google Académico

Cambiar foto

Arturo Elias

Profesor de sistemas electrónicos

inteligencia artificial - redes de computadoras

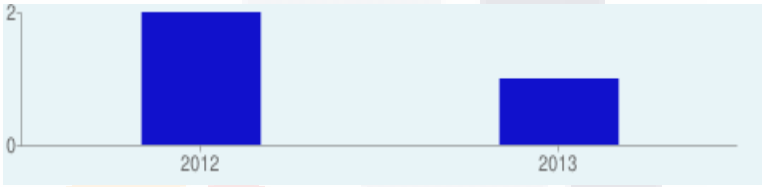
Dirección de correo verificada de correo.uaa.mx

Mi perfil es privado. Editar

Índices de citas

	Total	Desde 2008
Citas	3	3
Índice h	1	1
Índice i10	0	0

Citas sobre mis artículos



1-4

Título / Autor	Citado por	Año
Outlier analysis for plastic card fraud detection a hybridized and multi-objective approach A Elías, A Ochoa-Zezzatti, A Padilla, J Ponce Hybrid Artificial Intelligent Systems, 1-9	2	2011
New Implementations of Data Mining in a Plethora of Human Activities A Ochoa, J Ponce, F Ornelas, R Jaramillo, R Zataraín, M Barrón, C Gómez, J ... Knowledge-Oriented Applications in Data Mining, ISBN, 978-953	1	
Analysis of Cyber-bullying in a virtual social networking A Ochoa, J Ponce, R Jaramillo, F Ornelas, A Hernandez, D Azpeitia, A Elias Hybrid Intelligent Systems (HIS), 2011 11th International Conference on, 229-234		2011
Credit Scoring for Microfinance Institutions in México an Ensemble and Hybridized Approach A Elías, A Padilla, F Padilla		

1-4