UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS

DEPARTAMENTO DE SISTEMAS DE INFORMACIÓN

APLICACIÓN DE LA MINERÍA DE DATOS A LA INFORMACIÓN DEL CENSO DE POBLACIÓN Y VIVIENDA 2000, PARA MEJORAR EL DISEÑO DEL CUESTIONARIO DEL CENSO DE POBLACIÓN Y VIVIENDA 2010

CASO PRÁCTICO
PARA OBTENER EL GRADO DE:

MAESTRIA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES

PRESENTA

L.I. Simón Sánchez Trinidad

DIRECTOR DE CASO PRÁCTICO

Dra. Laura A. Garza González

SINODALES

M.C. Jorge Eduardo Macías Luévano

M.C. César Eduardo Velázquez Amador

Aguascalientes, Aguascalientes, Mayo 2008

AGRADECIMIENTOS

Agradezco principalmente a dios por haberme permitido alcanzar este logro.

A los maestros y doctores por los conocimientos vertidos durante las clases, su guía y apoyo durante todo el trayecto de la maestría.

A cada una de las personas que de alguna u otra manera han contribuido con la maestría y realización de este trabajo.

¡Gracias!

DEDICATORIAS

Le dedico este trabajo a mi familia (mi gran familia) ya que sin ellos no hubiera sido posible la consecución de este logro, a mi madre y a mi segunda madre, una dedicatoria especial para mi esposa Erika y a mis tres princesas Ceci, Sofi y Carolina, por su comprensión y apoyo durante la maestría y la realización de este trabajo.

Por este conducto, autorizamos al tesista:

L.I. Simón Sánchez Trinidad

La impresión de su documento final de Tesis, ya que cumple con los requisitos de contenido y forma exigidos en la Universidad Autónoma de Aguascalientes.

Asesor

Dra. Laura A. Garza González
Asesor principal

M.C. Jorge Eduardo Macías Luévano
Revisor

M.C. César Eduardo Velázquez Amador
Revisor

RESUMEN

La minería de datos es una tecnología de información que involucra los métodos de análisis tradicionales con algoritmos sofisticados para procesar grandes volúmenes de información, en esta tesina se trata el uso de la minería de datos en un campo práctico como lo es el diseño de cuestionarios, se exponen la metodología utilizada para la realización de la minería de datos, así como los resultados obtenidos.

La contribución de este trabajo práctico y de investigación, se centra en la propuesta de la utilización de la minería de datos en apoyo a los procesos de generación de información hacia los usuarios basándose principalmente en la metodología CRISP-DM (CRoss-Industry Standard Process for Data Mining) así como el uso de herramientas tales como WEKA, STATISTICA y RATTLE.

En base a la metodología CRISP-DM se desarrolla el proyecto de minería de datos sobre los datos del XII Censo de Población y Vivienda del año 2000, con el objetivo de mejorar el diseño del cuestionario del Censo 2010.

Se exponen las diferentes fases del desarrollo de un proyecto de minería de datos, así como la aplicación de diversas herramientas de software en cada una de ellas, el desarrollo de la tarea de predicción de la minería de datos, en la cual se han utilizado las redes neuronales las cuales son colecciones de nodos con entradas, salidas y procesamiento en cada nodo, entre la entrada y la salida existen un número de capas ocultas de procesamiento.

Palabras claves:

Diseño de cuestionarios, Minería de Datos, Data Warehouse, Redes neuronales, Ingreso.

CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCIÓN

Cada vez más las políticas públicas se enfocan a la optimización de recursos, así mismo, es cada vez más frecuente que las instituciones del gobierno y privadas se vean afectadas por la restricciones presupuestales en la realización de sus actividades y por lo tanto se ven obligadas a realizar las mismas actividades con un menor presupuesto, lo cual las ha llevado a buscar nuevas alternativas para la consecución de sus objetivos utilizando y optimizando sus recursos.

Los datos que se generan al nivel transaccional, son la principal fuente para la generación de información, dependiendo de la calidad con la cual se recaben los datos a nivel transaccional será la calidad de la información que se obtenga.

Para generar información a partir de los datos es necesario someterlos a diversos análisis para encontrar patrones ocultos en los datos, de forma que la información contenida en los datos sea utilizada para llevar a cabo una toma de decisiones basada en información, al conjunto análisis utilizado para encontrar estas relaciones en los datos se le denomina minería de datos, la cual es una nueva disciplina que trata sobre estadística, tecnología de bases de datos, reconocimiento de patrones, aprendizaje automatizado y otras áreas, se relaciona con el análisis secundario de grandes bases de datos, con la intención de encontrar previamente relaciones no contempladas, las cuales son de interés o de valor a los dueños de la información tal y como lo expone David J. Han en su artículo sobre estadística [4].

Sin embargo, emprender una tarea de este tipo requiere de la realización de pasos ordenados que nos permitan generar un proceso

En el capítulo 1 se exponen los antecedentes y el contexto del proyecto de investigación, así como una breve reseña de lo que actualmente es Instituto Nacional de Estadística Geografía e Informática, los objetivos que persigue así como una visión de lo que será la institución a mediano plazo, se define la situación problemática así como la relevancia de la investigación, así como la factibilidad técnica con la que se cuenta, en este mismo capítulo se exponen los objetivos, preguntas y preposiciones de la investigación que focalizan el área problemática.

En el capítulo 2 se presentan las definiciones sobre el tema las cuales tienen como objetivo clarificar al lector sobre el tema y los tópicos tratados, asimismo se presentan tres casos de estudio, los cuales ofrecen visión de cómo ha sido atacado el problema, estos forman el punto de referencia de los resultados y los modelos obtenidos en capítulos subsecuentes.

El capítulo 3 presenta la metodología CRISP-DM (CRoss-Industry Standard Process for Data Mining) la cual es utilizada como base para llevar a cabo el proceso de minería de datos, en el presente trabajo, la metodología es expuesta en forma de actividades y tareas a realizar, las cuales ofrecen una idea clara de los que se debe, lo que no se debe hacer y los puntos de especial cuidado en la realización de la minería de datos.

En capítulo 4, se exponen los resultados obtenidos de la aplicación de la metodología de minería de datos, así como las mediciones estadísticas realizadas a los datos, este capítulo concluye con la explicación de los resultados obtenidos de la aplicación de la minería de datos.

En los capítulos 5, 6 y 7 se describen los resultados de la investigación, conclusiones y recomendaciones que se obtienen de la presente investigación.

1. FORMULACIÓN DEL PROBLEMA

1.1. Contexto y antecedentes

El Instituto Nacional de Estadística, Geografía e Informática (INEGI) es un órgano desconcentrado de la Secretaría de Hacienda y Crédito Público (SHCP), dotado de autonomía técnica y administrativa, los objetivos del Instituto son:

- Generar e integrar información estadística y geográfica sobre el territorio, la población y la economía de México;
- Proporcionar a la sociedad el servicio público de información estadística y geográfica; así como
- Normar, coordinar y promover el desarrollo de los Sistemas Nacionales Estadístico y de Información Geográfica, con el objeto de
- Satisfacer las necesidades de información de los diversos sectores de la sociedad.

El Instituto fue creado por decreto presidencial, el 25 de enero de 1983 la Coordinación General de los Servicios Nacionales de Estadística, Geografía e Informática pasa a ser el Instituto Nacional de Estadística, Geografía e Informática, dependencia subordinada a la entonces Secretaría de Programación y Presupuesto.

Su establecimiento fue la respuesta del Gobierno de la República para garantizar la mejora sustancial en la calidad y homogeneidad de la información, y además, hacer posible unir esfuerzos de las diferentes instancias y niveles de gobierno, en la integración de un sistema nacional que amplió los alcances que puede tener el uso de la información estadística y geográfica en la instrumentación del plan nacional y de los programas sectoriales y regionales de desarrollo. Lo anterior, bajo la perspectiva de que a finales de los años 70, México no contaba con la suficiente información precisa y detallada sobre la estructura y crecimiento de la economía nacional que le permitiera planear su desarrollo.

El instituto ha expresado sus objetivos a través de su misión:

> **Generar, integrar y proporcionar información estadística y geográfica de interés nacional, así como normar, coordinar y promover el desarrollo de los Sistemas Nacionales Estadístico y de Información Geográfica, con objeto de satisfacer las necesidades de información de los diversos sectores de la sociedad.**

Por otro lado, el Instituto tiene una clara visión de lo que será la situación en el plano estadístico a través de la siguiente declaración:

México pertenece al grupo de países que basan su desarrollo en el uso de la información y en el conocimiento organizado y diseminado electrónicamente al contar con un Sistema Nacional de Información Estadística y Geográfica sustentado en una Red Nacional de Información, que facilita la toma de decisiones de todos los sectores de la sociedad con base en información oportuna y confiable, asimismo el INEGI es responsable de coordinar el Sistema Nacional de Información Estadística y Geográfica, así como la Red Nacional de Información.

Para la realización de la misión y la visión antes vertidas, el INEGI, basa sus labores en la siguiente afirmación:

Todo producto o servicio que se genere en el INEGI, debe tender a la plena satisfacción de las necesidades de información estadística y geográfica de la sociedad mexicana, mediante el desarrollo de su personal y la mejora continua, privilegiando la integración de metodologías y tecnologías en sus procesos y proyectos.

Por su parte en materia ambiental, el INEGI, tiene un claro compromiso sobre la ecología, lo cual expone a través de su política ambiental:

El INEGI, coordinador y generador de información por medios electrónicos, ópticos e impresos, comprometido con el cumplimiento de la legislación ambiental vigente y otros requisitos que la organización suscriba, así como la prevención de la contaminación; establece para sus instalaciones los objetivos de promover el consumo responsable de materiales de oficina, energéticos, agua y el adecuado manejo de los residuos, procurando una mejora continua en el desempeño ambiental de sus tareas encomendadas como se describe en la identidad institucional [8].

Un objetivo dentro de la declaración de la misión es la producción de la información estadística y geográfica, objetivo que se logra mediante diversos proyectos entre los que se incluyen los censos de población y vivienda, censos económicos, censos agropecuarios y a través de encuestas ya sean tradicionales o especiales, las cuales se caracterizan por aplicarse a una muestra de la población.

Para la generación de información se realizan diversos eventos, uno de los más conocido y difundido es el Censo de Población y Vivienda, el cual es un operativo a nivel nacional, en el cual se visitan todas las viviendas del país para obtener información en cada una de ellas.

Los datos recabados en los censos tienen características especiales, entre las que se encuentra la posibilidad de la realización estudios con un niveles pequeños de desagregación geográfica, permitiendo realizar una descripción de la población y sus diversas características, sin embargo este tipo de eventos son muy costosos, en nuestro país este evento se realiza cada 10 años, en el periodo intercensal se llevan a cabo encuestas simultaneas.

1.2. Situación Problemática

El proceso de generación de estadística básica e independientemente del método de captación de datos (censal, de encuesta por muestreo o por registros administrativos), se realizan una serie de actividades de naturaleza técnica, las cuales son propias de un proyecto estadístico, considerando el método para generar la estadística y su realización única o periódica, estas actividades se dividen en:

- Diseño conceptual.
- Diseño de la muestra (para encuestas por muestreo).
- Diseño de la captación y el procesamiento.
- Captación.
- Procesamiento.
- Presentación de resultados.

Dentro de la actividad de diseño conceptual un aspecto importante a considerar es el diseño del cuestionario ó instrumento de captación, dado que es la herramienta básica para obtener y registrar los datos de interés, conforme al objetivo del proyecto, con el fin atender por un lado el mejoramiento de la calidad de los datos y por el otro la satisfacción a la demanda social de información, la cual está compuesta por las necesidades de información de diversos niveles, a saber, gobierno, la academia y el sector privado.

El método de levantamiento de datos por excelencia dentro del Instituto es a través de cuestionarios ya sean en medios físicos o por medio de Internet, este método para la obtención de datos supone establecer un proceso de comunicación en el que el investigador, de acuerdo las necesidades y objetivos de una investigación establece en el marco conceptual qué se pregunta, de qué forma se pregunta, a quién y dónde, según lo expone Pedro Antonio García López et al. en su artículo sobre los problemas en el diseño y validación de los cuestionarios [24], indican que a este conjunto de elementos se denominan las cuatro cuestiones “W” (What, hoW, Who, Where), este proceso de comunicación se establece de forma oral o escrita, y está sujeto a “interferencias” que pueden alterar o

dañar la calidad de los datos obtenidos en el proceso de generación de información estadística.

El mal diseño y la "explotación" del cuestionario suponen ser las causas principales de los errores y puede propiciar que en fases posteriores de explotación de los datos, los resultados sean meramente "creencias" o "intenciones" [24].

De esta forma el diseño de los cuestionarios constituye la operacionalización del Marco Conceptual, donde se identifican y justifican cada uno de los conceptos involucrados en la captación. El diseño del cuestionario ha de ser apropiado a las características del ámbito y circunstancias en que se han de aplicar, a nivel de los elementos individuales de la población de estudio, de tal forma que se facilite la fase de captación y los datos correspondan efectivamente al significado de cada concepto.

Actualmente una vez definido el diseño de cuestionario, se realiza la prueba de campo, en la cual es aplicado a una muestra para validar si cumple con los objetivos propuesto, si no es el caso, se realizan adecuaciones al instrumento de levantamiento, sin embargo, debido a que solo es aplicado a una muestra, estas adecuaciones pueden no ser representativas de la población bajo estudio, ya en campo, pueden identificarse problemas, a través de la experiencia de los encuestadores, los cuales aplican esas experiencias en los levantamientos subsecuentes, así al finalizar la captura de los instrumentos de captación, se hace necesaria la realización de análisis detallados de los datos, lo cual permita identificar problemas en el instrumento de captura y mejoramiento del mismo como se expone en el documento d diseño de cuestionarios [13], este análisis puede ser realizado de la forma tradicional o a través de los métodos estadísticos avanzados de la minería de datos.

Aunado a la necesidad de mejora en el diseño de los cuestionarios, actualmente los recursos económicos para realizar levantamientos de información son cada vez más reducidos, los informantes son sobreexplotados en aras de la obtención de información, mientras que, las necesidades de información estadística continúan creciendo en varios sentidos, los usuarios de la información estadística solicitan que la información tenga las siguientes características: mayor frecuencia, mayor cobertura temática, mayor capacidad para el estudio de relaciones entre temas, mayor cobertura geográfica, calidad y precisión tal y como lo expone en el artículo sobre la estimación de agregados municipales utilizando información muestral censal [5], Esta presión económica y técnica hace que la alternativa de cuestionarios que cubran una amplia gama de fenómenos y por consiguiente con un número mayor de preguntas sea inviable. Dado que los costos aumentan conforme crece el número de preguntas con lo cual se hace necesario la mejora de los cuestionarios tanto en su diseño como en el tamaño de los mismos buscando mejorar la calidad los datos con el objetivo de que sean:

- Útiles para la realización de inferencias;
- Permitan ser útiles para futuros levantamientos;
- Estar libres de inconsistencias.

En resumen el diseño del cuestionario es una actividad importante dentro del proceso de generación de estadística básica y por consiguiente importante en la generación de la estadística oficial, la información que se recaba en los instrumentos de es utilizada principalmente para propósitos de difusión de la información y en menor medida utilizada para la evaluación de los cuestionarios; los datos e información generada de los diversos eventos se encuentra disponible en el data warehouse estadístico posibilitando la aplicación técnicas avanzadas de análisis de datos, que permitan ofrecer un valor agregado a la actividad del diseño de cuestionarios.

1.3. Relevancia del Caso o Proyecto

En la actualidad la información es un activo para cualquier tipo de empresas, en la iniciativa privada los datos que se recaban en los sistemas transaccionales son sometidos a diversos análisis para encontrar patrones ocultos en los datos, de forma que la información contenida en los datos sea utilizada para llevar a cabo una toma de decisiones basada en información, al conjunto análisis utilizado para encontrar estas relaciones en los datos se le denomina minería de datos, la cual es una nueva disciplina que trata sobre estadística, tecnología de bases de datos, reconocimiento de patrones, aprendizaje automatizado y otras áreas, se relaciona con el análisis secundario de grandes bases de datos, con la intención de encontrar previamente relaciones no contempladas, las cuales son de interés o de valor a los dueños de la información tal y como es expuesto en el artículo sobre estadística en [4], las bases de datos de gran tamaño, por lo general son almacenadas en bases de datos sintéticas de información, a los cuales previamente se les ha aplicado el proceso de data warehouse, los cuales generaran un acervo de información que puede ser explotado para propósitos diferentes para los cuales fueron pensados inicialmente.

La realización de esta investigación apoya al instituto en la mejora de sus procesos de trabajo aportando una mejora en la información que se le ofrece a los usuarios, para el logro de la misión de la institución, con especial énfasis en la fase de diseño de cuestionarios a través del análisis de la información estadística utilizando minería de datos, por otra parte, no se limita solamente a esta institución, ya que puede ser utilizado por otras instituciones generadoras de información estadísticas tanto nacionales como internacionales.

Se considera que este proyecto es factible, debido a que en el instituto ya se cuenta con un almacén de datos estadístico generado mediante la utilización del proceso de data warehouse, este acervo informativo está integrado por información proveniente de diversos eventos entre ellos el censo de población vivienda del año 2000, la información almacenada en el almacén de datos estadístico, es utilizada como fuente para divulgación en el sitio, sirve como base para análisis de la situación poblacional de México, así como ser la base para el marco muestral maestro y marcos muestrales especiales, lo cual lo convierte en un campo fértil para la experimentación y aplicación de nuevas técnicas de análisis de los datos, que permitan generar mejoras en los procesos institucionales.

1.4. Objetivos, Preguntas y Proposiciones del Proyecto

1.4.1. Objetivo general

Mejorar el diseño del cuestionario del censo de población y Vivienda 2010 mediante la aplicación de la minería de datos al acervo informativo del censo de población y vivienda del año 2000.

1.4.2. Objetivos específicos

Mejorar el diseño del cuestionario del censo de población y vivienda 2010.

Mejorar la calidad de los datos recabados por el cuestionario del censo de población y vivienda 2010.

Aplicar la minería de datos en la mejora del diseño del cuestionario del censo de población y vivienda 2010.

Analizar la información del Censo de Población y Vivienda 2000 contenida en el data warehouse mediante la minería de datos.

1.4.3. Preguntas de investigación

Las preguntas a las cuales responde esta investigación son:

¿Es posible mejorar del diseño del cuestionario del Censo de Población y Vivienda 2010 analizando la información del censo anterior?

¿Es posible mejorar la calidad de los datos recabados a partir de un diseño de cuestionario mejorado?

¿La minería de datos puede ser utilizada en la actividad de mejora de cuestionarios?

¿La información del Censo de Población y Vivienda del año 2000 en el data warehouse puede ser utilizada para realizar minería de datos?

1.4.4. Preposiciones específicas

La mejora del diseño del cuestionario del CPV 2010 puede realizarse a través de la explotación de los datos del CPV anterior.

Al utilizar un diseño de cuestionario mejorado se incrementa la calidad de los datos captados.

La minería de datos puede ser aplicada a la actividad de diseño de cuestionarios

La información CPV contenida en el DWH puede ser utilizada para realizar minería de datos.

2. MARCO TEÓRICO

2.1. Marco teórico

2.1.1. Diseño de cuestionario

El diseño de cuestionarios es la macroactividad del Diseño Conceptual, en la que se resuelve la redacción de las preguntas, su secuencia, instrucciones, distribución de contenidos y edición del formato utilizado para obtener la información que es de interés para el proyecto.

Como fase de un proyecto de estadística básica, el Diseño Conceptual es la serie de actividades mediante la cual se identifican las necesidades de información, con base en las cuales se determinan:

- El marco conceptual (temas, categorías, variables y clasificaciones) a que serán referidos los datos.
- Los esquemas para la presentación de resultados.
- Los instrumentos para su captación (cuestionarios u otro tipo de formatos).
- Los criterios de validación.

El diseño de los cuestionarios constituye la operacionalización del Marco Conceptual, donde se identifican y justifican cada uno de los conceptos involucrados en la captación. El diseño ha de ser apropiado a las características del ámbito y circunstancias en que se han de aplicar, a nivel de los elementos individuales de la población de estudio, de tal forma que se facilite la fase de captación y los datos correspondan efectivamente al significado de cada concepto.

Por ello, las preguntas deben ser comprensibles y facilitar el desarrollo de la entrevista, ya que de esto depende la calidad de los datos a obtener.

2.1.2. Instrumento de captación

Es el formato que se utiliza para el registro de los datos, en un proyecto estadístico; tal información se ha definido previamente y organizado en el marco conceptual.

2.1.3. Cuestionario

Es un tipo de instrumento de captación que presenta preguntas y/o enunciados dirigidos a los informantes, para obtener datos específicos acerca de las variables que serán objeto de captación.

2.1.4. Marco conceptual

De un proyecto estadístico es el ordenamiento de temas, categorías, variables y clasificaciones al cual se referirán los datos objeto de captación, incluido el glosario con las definiciones formales de cada uno de los conceptos a utilizar en el cuestionario.

Tema

Enunciado genérico referente a un campo de conocimiento. Su estudio constituye la justificación del proyecto estadístico.

Categoría

Conjunto objeto de cuantificación y caracterización.

En el diseño del cuestionario, los temas y las categorías se toman en cuenta para marcar las diferentes secciones en la distribución de las preguntas.

Variable

Concepto que admite distintos valores para la caracterización o clasificación de un elemento o un conjunto. En el diseño del cuestionario, las variables se traducen generalmente en preguntas aplicables a cada elemento de una categoría específica.

Clasificación

Ordenamiento de todas las modalidades nominales o intervalos numéricos admitidos por una variable.

Clase

Cada una de las modalidades nominales o intervalos numéricos admitidos por una variable.

En el instrumento de captación, las clases se relacionan con las opciones de respuesta a una pregunta.

2.1.5. Fuentes de error de la generación de estadística básica

El investigador se enfoca en el estudio de estas fuentes de error dado que pueden afectar las fases posteriores de la explotación de datos, haciendo que los resultados sean solo "creencias" o "intenciones". Fuentes de error en el proceso de generación de estadísticas básica, como lo expone Pedro Antonio García López et al. en su artículo sobre diseño y validación de cuestionarios [24]; la mayoría de los autores listan como fuente de errores los siguientes:

Imputables al investigador

Marco muestral deficiente.

Muestra no representativa, debido a un mal esquema de muestreo.

Entrevistadores mal seleccionados ó deficientemente capacitados.

Diseño de cuestionario deficiente.

Imputables al entrevistador

Conducción deficiente de la entrevista.

Presentación errónea por parte del entrevistador.

Conducción errónea de la entrevista.

Mal formulación de preguntas

Deficiencias en el control y registro de preguntas.

Seguimiento erróneo de las instrucciones.

Fraude o falsificación de cuestionarios.

Imputables al entrevistado

Falta de respuesta debido al temor del no anonimato, debido a preguntas de contenido intimo.

Falta de comprensión de las preguntas.

Falta de sinceridad y veracidad.

Respuestas sesgadas o forzadas por complacer el sentido del cuestionario.

2.1.6. Pruebas y ajuste de cuestionarios

En el documento Diseño de Cuestionario [13], Expone que toda medición efectuada mediante cualquier instrumento de captación está expuesto a errores, los cuales pueden ser sistemáticos o aleatorios, por lo cual a través de la pruebas se busca obtener elementos a través de los cuales se pueda verificar la validez y confiabilidad del cuestionario, indicando que para que un cuestionario sea valido debe de cumplir con la premisa de obtener la información para la cual fue diseñado, esta validez representa la relación existente entre lo que se mide y lo que realmente se deseaba medir, con lo cual se refleja hasta que punto existe una desviación sistemática (sesgo) en las respuestas captadas.

Se identifican dos pruebas para verificar la validez de un instrumento de captación, a saber: De contenido, la cual permite conocer el grado en el cual un instrumento refleja un campo específico de conocimiento o características de un fenómeno, de criterio, la cual se basa en la comparación de los resultados obtenidos por el cuestionario, con respecto a un criterio externo el cual puede ser un valor fijado ya sea por los especialistas o mediante evidencia empírica.

Con respecto a las pruebas de confiabilidad, las cuales se orientan a la valoración de la consistencia del instrumento, verificando si este es capaz de obtener valores iguales en el mismo entrevistado y en diversos momentos, evaluando hasta qué punto la información obtenida de las respuestas corresponde con una variación por azar, indican que la medición de la confiabilidad se basa por lo general en métodos cuantitativos aplicables a la obtención de coeficientes de correlación, los cuales miden el grado de cambio de una variable con respecto a otra.

Para medir la calidad debido al uso de los cuestionarios como instrumentos de captación Mercedes Rodríguez et al. en su artículo sobre el uso de la estadística [17], indican que pueden ser medidos a través de coeficiente de fiabilidad y validez, para la prueba de fiabilidad la cual se refiere a la consistencia entre diversas preguntas que miden la misma dimensión, si se obtienen respuestas homogénea o lo que es igual puntuaciones semejantes se utiliza el alfa de Cronbach.

En lo que respecta a la validez Mercedes Rodríguez et al. [17] indican que existen procedimientos empíricos para medir la validez de la prueba, como el uso de criterios internos cuando no se tiene un criterio externo, o a través del análisis discriminante, el cual trata de diferenciar grupos a partir de un elemento criterio.

Indica además que se deben de tener en cuenta los criterios de utilización del cuestionario, los cuales son:

Transparencia

Definida como el grado en que resulta significativa, tanto para el evaluador como para el evaluado, la actitud del evaluado con respecto a los resultados obtenidos (validez aparente).

Aceptabilidad

Está relacionada con la anterior, como resultado de la aplicación de los cuestionarios (como se lleva a cabo, grado de participación, etc.).

Valor de la información

Grado de la información que aporta para conseguir los objetivos de mejora, en otras palabras, saber que se ha realizado mal, ó bien y como mejorarlo.

2.1.7. Censos de población y vivienda

En México el primer censo de población en tiempos modernos fue llevado a cabo en el año de 1895, cinco años después se levanto el segundo (1900), desde entonces se ha realizado cada década (en los años terminados en cero), con excepción de 1920, cuando por razones políticas y sociales tuvo que levantarse en 1921, hasta la fecha se han realizado doce censos de población [2], más dos recuentos intercensales, el primero entre 1990 y el 2000, denominado Conteo de Población y Vivienda 1995 y el segundo realizado en el año 2005, denominado II Conteo de Población y Vivienda 2005.

Se debe de hacer hincapié en que a partir del año de 1950 se realizan censos de población y vivienda de forma simultánea; en censos anteriores, solamente se captaban algunas características de la vivienda.

2.1.8. XII Censo General de Población y Vivienda 2000

Es un proyecto de generación de estadísticas que realizo el INEGI, en el cual se capto información sobre las características sociodemográficas de la población en México plamada en la síntesis metodológica del XXI Censo General de Población y Vivienda [10], los preparativos para llevar a cabo este censo comenzaron en 1997 con la evaluación de dos proyectos estadísticos previos: El XI Censo General de Población y Vivienda 1990 y el Conteo de Población y Vivienda de 1995, así como el análisis de las recomendaciones internacionales sobre población y vivienda para la ronda censal descritas en las características metodológicas del XII Censo general de Población y Vivienda [2], esto dio como resultado la base para determinar el contenido temático, los conceptos, el diseño conceptual y las estrategias generales para el operativo de campo así como el procesamiento de la información.

Los objetivos y las metas que se fijaron para el XII Censo General de Población y Vivienda 2000 fueron:

- Generar información demográfica y socioeconómica sobre el país
- Asegurar la máxima desagregación geográfica de la información
- Enriquecer la serie histórica de datos estadísticos, manteniendo en lo posible la comparabilidad nacional e internacional.
- Construir marcos de muestreo para encuestas

Por otro lado, las metas planteadas para este proyecto estadístico fueron las siguientes:

- Lograr la máxima cobertura de la población y las viviendas
- Obtener información de óptima calidad
- Alcanzar la mayor oportunidad en la publicación de los resultados
- Ampliar la oferta de información y diversificar los productos censales

Las principales unidades de análisis de este evento estadístico fueron los residentes habituales y las viviendas, considerando al residente habitual de la vivienda a toda persona que habita normalmente en la vivienda, esto es que en ella, duerme, prepara sus alimentos, como y se protege del ambiente y por ello la reconoce como su lugar de residencia.

Como vivienda se consideró a todo espacio delimitado normalmente por paredes y techos, de cualquier material, con entrada independiente; que se utiliza para vivir, esto es dormir, preparar alimentos, comer y protegerse del ambiente.

Las características metodológicas expuestas en las síntesis metodológica del XII Censo [10] fueron:

- Un periodo de dos semanas para el levantamiento de la información
- Fue un censo de hecho o de jure, lo que significa censar a la población en su lugar residencia habitual.
- Se aplicó un solo tipo de cuestionario por vivienda
- Se captó información a partir de entrevista directa a un informante adecuado, definido como una persona de 15 años y más cumplidos, que habitara en la vivienda y que conociera los datos de los residentes.
- Se utilizaron dos tipos de cuestionario: uno básico y otro ampliado, este último se aplico a una muestra probabilística de viviendas, en cambio el básico se aplico a todas las viviendas de manera exhaustiva.

2.1.9. Data warehouse

Proceso de integración de información a partir de bases de datos transaccionales, la cual es organizada mediante modelos multidimensionales de información cuyo propósito es ofrecer información para su análisis por los niveles directivos.

2.1.10. Data warehouse estadístico

Está compuesto de bases de datos relacionales que contienen información fuente (a nivel cuestionario) de los proyectos estadísticos, actualmente cuenta con la información de más de

29 proyectos estadísticos, algunos de ellos con más de cien millones de registros como se describe en la presentación del Data Warehouse del servicio público de información estadística [7].

2.1.11. Descubrimiento de conocimiento en bases de datos (Knowledge discovery in databases[KDD])

En su libro introductorio a la minería de datos Pang-Ning Tan et al. [21], exponen que la minería de datos es una parte integral de KDD, el cual es el proceso completo de conversión de datos sin procesar en información útil, este proceso consiste en una serie de pasos de trasformación, desde el procesamiento de datos hasta los resultados de la minería de datos, este proceso puede ser observado en la figura 1:

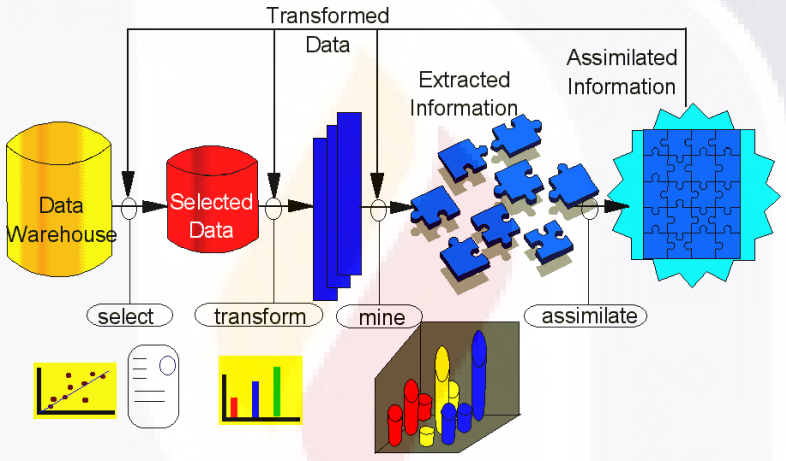


Figura 1: Proceso de descubrimiento de conocimiento

El descubrimiento de conocimiento en bases de datos, también es llamado proceso de descubrimiento de conocimiento, cuyo objetivo es el de buscar nuevo conocimiento en un dominio de aplicación descrito por Krzyztof et al. en su libro sobre minería de datos [16].

Krzyztof et al.. En [16] describe que desde los años 1990, diferentes procesos de descubrimiento de información se han desarrollados, enfatiza que los primeros esfuerzos fueron desarrollados por la investigación académica, pero rápidamente fueron seguidos por la industria, entre los modelos académicos destacan los realizados por Fayad con su modelo de nueve pasos, el modelo de nueve pasos de Fayad se desglosa en:

- Desarrollo y Entendimiento del dominio de aplicación, incluye el aprendizaje de conocimiento anterior relevante y los objetivos del usuario del conocimiento descubierto.
- Creación de un grupo de datos objetivo, en este paso el minero de datos selecciona un subgrupo de variables (atributos) e instancias (ejemplos), el cual será usado para

realizar las tareas de descubrimiento, este paso usualmente incluye la consulta a la base de datos y la selección del subgrupo seleccionado.

- Limpieza de los datos y preprocesamiento, este paso consiste en la eliminación de observaciones atípicas, tratando con el ruido y valores perdidos dentro de los datos, tomando en cuenta del tiempo de secuencia de la información y cambios conocidos.

- Reducción de los datos y proyección, este paso consiste en encontrar atributos útiles por medio de métodos de reducción y la trasformación, encontrando representaciones invariante de los datos.

- Elección de la tarea de minería de datos, el minero de datos alinea los objetivos definidos en el paso 1 con un método particular de la minería de datos (predicción, clasificación, regresión, agrupamiento).

- Elección del algoritmo de minería de datos, el minero de datos selecciona los métodos para la búsqueda de patrones y decide que modelos y parámetros de los métodos usados pueden ser apropiados.

- Minería de datos, este paso genera los patrones enana forma representativa particular, ya sean reglas de clasificación, árboles de decisión, modelos de regresión, tendencias, etc.

- Interpretación de los patrones encontrados, en este paso el analista realiza la visualización de los patrones extraídos y los modelos y la visualización de los datos en base a los modelos obtenidos.

- Consolidación del conocimiento encontrado, el paso final consiste en la incorporación del conocimiento descubierto en el sistema de desempeño, documentando y reportándolo hacia las partes interesadas, este paso puede incluir la verificación y la resolución de conflictos potenciales con el conocimiento previo aceptado.

Este proceso es iterativo, en su libro sobre minería de datos Krzyztof et al. [16] explican que el autor dice que un número de ciclos entre dos pasos son generalmente realizados, pero no da más explicaciones.

Los autores del libro minería de datos: un enfoque de descubrimiento de conocimiento [16] mencionan que entre los modelos industriales más difundidos se encuentran los modelos de Cabena et al.. El cual es apoyado por IBM y el proceso estándar industrial interrelacionado para la minería de datos (CRISP-DM), esta metodología fue concebida a finales de 1996 por tres "veteranos" del mercado joven e inmaduro de minería de datos: DaimlerChrysler (Daimler-Benz), SPSS (entonces ISL) y NCR, los cuales establecieron grupos de consultores de minería de datos y especialistas para atender los requerimientos de los clientes.

2.1.12. Proceso estándar interrelacionado para la minería de datos (CRISP-DM)

Es un modelo de proceso estándar el cual no es propietario y disponible libremente, es descrito en términos de un modelo jerárquico de procesos, consiste en un grupo de tareas detalladas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea

genérica, tarea especializada e instancia de proceso descrita en la guía sobre la metodología CRISP-DM versión 1.0 [22].

El proceso de descubrimiento de conocimiento CRISP-DM consiste en seis pasos los cuales se describen a continuación:

Entendimiento del negocio, este paso se enfoca en el entendimiento de los objetivos y requerimientos desde la perspectiva del negocio. También lo convierte en una definición de problema de minería de datos y diseña un plan preliminar para alcanzar los objetivos, se desglosa en los siguientes subpasos:

- Determinación de los objetivos del negocio.
- Análisis de la situación
- Determinación de los objetivos del negocio
- Generación del plan de minería de datos

Entendimiento de los datos, este paso comienza con la recolección de datos y familiarización con ellos, el propósito específico es la identificación de problemas de calidad en los datos. Incursiones iníciales en los datos y detección de grupos de datos interesantes, se desglosa en las siguientes subtareas:

- Recolección de los datos iníciales
- Descripción de los datos
- Exploración de los datos
- Verificación de la calidad de los datos

Preparación de los datos, cubre las actividades requeridas para integrar la base de datos final, la cual será la que alimentará las herramientas de minería de datos del siguiente paso. Este incluye la selección de tablas, registros y atributos; la limpieza de los datos; la construcción de nuevos atributos; y la trasformación de los datos, este paso incluye las siguientes subtareas:

- Selección de los datos
- Limpieza de los datos
- Construcción de los datos
- Integración de los datos
- Formateo de datos.

Modelado, en este paso varias técnicas de modelado son seleccionadas y aplicadas, el modelado generalmente involucra el uso de algunos métodos para el mismo tipo de problema de minería de datos y la calibración de sus parámetros a los valores óptimos, dado que algunos métodos requieren un formato especifico en los datos de entrada, regularmente la reiteración a los pasos anteriores es necesaria, este paso se subdivide en:

- Selección de la técnica de modelado
- Generación de la prueba de diseño
- Creación del modelo
- Análisis de los modelos generados
- Evaluación de los modelos

Evaluación, Después de que uno o más modelos se han construido que tienen calidad alta desde la perspectiva del análisis de los datos, el modelo se evalúa desde una perspectiva objetiva de negocios. Se realiza una revisión de los pasos realizados para la construcción del modelo. Un objetivo clave es determinar si algunas cuestiones importante de la empresa no han sido suficientemente consideradas. Al final de esta fase, se debe de tener una decisión sobre el uso de los resultados de la minería de datos. Las sub fases que incluye esta fase son:

- Evaluación de los resultados,
- Proceso de revisión, y
- Determinación de la siguiente etapa.

Desarrollo, el conocimiento obtenido debe ser organizado y presentado de una manera que el cliente puede utilizar. Dependiendo de los requisitos, este paso puede ser tan simple como la generación de un informe o tan compleja como la aplicación de un proceso repetible de minería de datos, esta fase se divide en:

- Plan de desarrollo,
- Plan de vigilancia y mantenimiento,
- Generación de informe final, y
- Revisión del las fases del proceso.

CRISP-DM hace de la minería de datos un proceso de negocio al enfocar la tecnología a la resolución de problemas de negocio específicos, lo cual es descrito en el artículo de la herramienta Clementine descrito en el artículo promocional de esta herramienta [3].

Según José Ramón Cano et al.. En su artículo sobre el uso de algoritmos evolutivos aplicados a la selección de instancias para la reducción de datos [15], las tareas de la comprensión del problema y los datos y el preprocesamiento de los datos juegan un papel central en una minería de datos exitosa.

2.1.13. SEMMA

Es el acrónimo de Sample, Explore, Modify, Model, Assess, se refiere a la esencia del proceso de realización de minería de datos a partir de una muestra estadísticamente representativa de datos, SEMMA hace de fácil aplicación: la exploración estadística y las técnicas de visualización, las selección y transformación de las variables predictoras importantes, modelado las variables para predecir los resultados, y confirmación de la exactitud de un modelo.

Antes de examinar cada una de las etapas del SEMMA, un malentendido común es referirse a SEMMA como una metodología de la minería de datos. SEMMA no es una metodología de la minería de datos, sino más bien una lógica de la organización funcional del conjunto de herramientas de SAS Enterprise Miner para llevar a cabo las tareas básicas de la minería de datos. Enterprise Miner se puede utilizar como parte de cualquier metodología iterativa de minería de datos adoptada por el cliente. Naturalmente los pasos de la definición problema del negocio o problema de investigación y la generación fuentes de datos representativas son críticos para el éxito general de cualquier proyecto de minería de datos. SEMMA se centra en el modelo de desarrollo de los aspectos de la minería de datos:

- Muestra (opcional) de los datos mediante la extracción de una parte de un gran conjunto de datos lo suficientemente grande como para contener la información importante, pero lo suficientemente pequeña como para manipular rápidamente. Para optimizar costes y rendimiento, SAS Institute recomienda una estrategia de muestreo, que usa muestra confiable estadísticamente representativa de las grandes fuentes detalladas de datos. La realización de la minería de datos de una muestra representativa en lugar de todo el volumen reduce el tiempo de procesamiento necesario para obtener la información esencial de negocios. Si aparecen patrones generales en los datos en su conjunto, estos serán rastreables en una muestra representativa. Si un nicho es diminuto que no es representado en una muestra y, sin embargo, es tan importante que influye en el panorama general, puede ser descubierto usando los métodos de síntesis. También se recomienda la creación de la partición de conjuntos de datos con el nodo de partición de datos:
    - Entrenamiento, Utilizado para la prueba del modelo
    - Validación, Utilizado para la evaluación y para evitar sobre entrenamiento.
    - Prueba, Utilizado para obtener una evaluación honesta de lo bien que un modelo generaliza.
- Explorar los datos mediante la búsqueda de tendencias y anomalías imprevistas para contribuir al entendimiento y las ideas. La Exploración ayuda a perfeccionar el proceso de detección. Si la exploración visual no revela tendencias claras, puede explorar los datos a través de técnicas estadísticas incluido: análisis factorial, análisis de correspondencia, y las agrupaciones.
- Modificar sus datos mediante la creación, selección, y la transformación de las variables para centrar el modelo de proceso de selección. A partir de sus descubrimientos en la fase de exploración, es posible que tenga que manipular los datos para incluir información como por ejemplo la agrupación de clientes importantes y subgrupos, o para introducir nuevas variables. También puede ser necesario buscar valores atípicos y reducir el número de variables, reduciendo las

más significativas. También puede ser necesario modificar los datos cuando los datos de minería cambien. Debido a la minería de datos es un proceso dinámico e iterativo, se pueden actualizar los métodos de minería de datos o incluir nuevos modelos cuando se dispone de información.

- Modelado de datos, permite al software para la búsqueda automática de combinaciones de datos que permita la predicción del resultado deseado con fiabilidad. Las técnicas de modelado en la minería de datos incluyen redes neuronales, árboles basados en modelos, modelos logísticos, y otros modelos estadísticos (análisis de series de tiempo, la memoria basada en el razonamiento, y de componentes principales). Cada tipo de modelo tiene fortalezas particulares, y es apropiado dentro de las situaciones específicas de la minería de datos en función de los datos. Por ejemplo, las redes neuronales son muy buenos en el montaje muy complejas relaciones no lineales.

- Evaluar los datos mediante la evaluación de la utilidad y fiabilidad de los resultados de la minería de datos y el proceso de estimación de conocer su rendimiento. Una forma de evaluar un modelo es aplicarlo a una parte del conjunto de datos durante la etapa de muestreo. Si el modelo es válido, debe trabajar para esta muestra reservada, así como para la muestra utilizada para construir el modelo. Del mismo modo, puede probar el modelo en contra de los datos conocidos.

Al evaluar los resultados de cada etapa del proceso de SEMMA, se puede determinar cómo modelar las nuevas preguntas formuladas en los resultados anteriores, y, por tanto, proceder a la fase de exploración para una mejora adicional de los datos.

El desarrollo del modelo es el resultado final de minería de datos (en la fase final en la que el retorno de la inversión del proceso de la minería se realiza).

2.1.14. Minería de datos

Pang et al. en su libro introductorio sobre minería de datos [21], describe la minería de datos como el proceso de descubrir automáticamente información útil en grandes repositorios de información. Así mismo describen que las técnicas de minería de datos son desarrolladas para buscar en grandes bases de datos para encontrar patrones nuevos y útiles que de otra forma permanecen ocultos. También puede ofrece capacidades para predecir la salida de una observación futura, tal como predecir cuando un nuevo cliente gastará más de $1000 en un departamento de una tienda.

Por su parte Alberto Ochoa en el artículo sobre el efecto pigmaleón [1], explica que la minería de datos es la extracción de información oculta y predecible de grandes bases de datos, es una nueva tecnología que tiene un gran potencial para ayudar a las compañías o a las organizaciones a enfocarse en la información más importante dentro de sus bases de información (data warehouse), las herramientas de minería de datos predicen futuras tendencias y comportamientos, permitiendo a las organizaciones realizar una toma de decisiones proactiva y basada en información basada en conocimiento.

Tareas de la minería de datos, tomando como referencia el libro introductorio a la minería de datos de Pang et al. [21], en el cual se explica como la minería de datos suele ser utilizadas en tareas de predicción ó tareas de descripción, la primera se refiere a predecir el valor de un atributo en particular basando en el valor de otros atributos, el atributo a predecir es comúnmente conocido como variable objetivo o dependiente, mientras que los atributos usados para realizar la predicción son conocidos como variables explicatorias o independientes, mientras que la segunda se enfoca en la derivación de patrones (correlaciones, tendencias, agrupamientos, trayectorias y anomalías) que agrupan las relaciones esenciales. Las tareas descriptivas son generalmente de naturaleza exploratoria y frecuentemente requieren de técnicas de post procesamiento para validar y explicar los resultados [21].

Según lo expone José Ramón Cano et al. en su artículo sobre extracción de modelos predictivos [14], En minería de datos la extracción de modelos representativos es un proceso básico, estos puede ser:

- Modelos predictivos, cuyo objetivo perseguido es lograr una mayor precisión.
- Modelos descriptivos, los cuales intentan encontrar relaciones y patrones de comportamiento en el conjunto de datos para ofrecer conocimiento sobre un problema concreto.

En su artículo, los autores se centran en los modelos predictivos basados en reglas de clasificación, los modelos son en concreto árboles de decisión los cuales son extraídos mediante el algoritmo C45 de Quinlan, José Ramón Cano et al. [15]. Exponen que los modelos son generados empleando conjuntos de datos grandes, dando como resultado que los modelos extraídos presentan tamaños elevados, lo cual disminuye su interpretabilidad.

2.1.15. Herramientas para el desarrollo de Minería de datos.

WEKA, Es una herramienta de libre distribución, desarrollada por la Universidad de Wikato, permite la exploración de los datos de una forma visual, es una suite de algoritmos de maquinas de conocimiento implementados en JAVA, contiene además, las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización descrito en el manual de WEKA [18].

Rattle, Herramienta de minería de datos de libre distribución, basada en el software estadístico R, esta herramienta es útil en las primeras fases de la minería de datos, permitiendo al usuario familiarizarse con los datos con los cuales trabajará,

Clementine, según se describe en el artículo sobre este producto [3], clementine es una herramienta de minería de datos que permite desarrollar de forma rápida modelos predictivos y desplegarlos para mejorar la toma de decisiones. Clementine es conocida mundialmente como la herramienta líder de minería de datos, ya que entrega el máximo retorno de inversión de sus datos en poco tiempo. Clementine apoya el ciclo completo de minería de datos para reducir el tiempo hasta la solución final. Clementine está diseñada

considerando los estándares de la industria del minería de datos – CRISP-DM (CRoss Industry Standard Process for Data Mining).

Statistica, es un sistema completo para el análisis de datos con miles de pantallas personalizables y gráficos de alta calidad totalmente integrados con todos los procedimientos. Este software es desarrollado por STATSOFT, en las últimas versiones hace uso de la minería de datos como una parte integral del procesamiento de grandes cantidades de información.

2.1.16. Atributo

Un atributo es una propiedad o característica de un objeto, por ejemplo, el color y el tamaño, también se le conoce como: variable, campo o rasgo.

2.1.17. Registro

Es el grupo de atributos que describen un objeto, también se le denomina: registro, punto, muestra, entidad, instancia u observación.

2.1.18. Tipos de datos

La identificación de los tipos de datos es una tarea importante, ya que de esto depende que se realicen las operaciones adecuadas a los atributos.

Cualitativos (categóricos), los atributos representan distintas categorías en lugar de números, las operaciones matemáticas tales como suma o resta no tienen sentido.

Cualitativos Nominales, en este tipo de atributos categóricos no existe un orden predefinido.

Cualitativos Ordinales, en este tipo de atributos categóricos cuentan con un ordenamiento el cual da una idea del nivel de cada categoría.

Cuantitativos, son atributos numéricos y pueden ser tratados como tales, estos pueden ser:

Cuantitativos de intervalo, atributos numéricos cuya principal característica es la inexistencia de un cero absoluto, la división no tiene sentido.

Cuantitativos de razón, atributos numéricos en los cuales existe un cero absoluto, en este tipo de atributos la división si tiene sentido.

Discretos, son atributos que tienen grupo de valores finito o un grupo contable de valores infinitos, por lo general se representan con valores enteros, un caso especial de los atributos discretos son los atributos binarios, los cuales solo tienen dos valores posibles.

Continuos, este tipo de atributos tienen números reales como valores del atributo, pueden realizarse operaciones según lo permita el instrumento, en la práctica los valores reales solo pueden ser medidos y representados con un número finito de dígitos, los atributos continuos por lo general se representan utilizando variables de punto flotante.

Los atributos cualitativos o categóricos son siempre discretos, por su lado los atributos cuantitativos pueden ser discretos o continuos.

2.1.19. Propiedades de los atributos

Del tipo de un atributo depende cual de las siguientes propiedades posee:

Distinción (igual no igual)

Orden (<>)

Adición (+ -)

Multiplicación (* /)

Atributos nominales: Distinción

Atributos Ordinales: Distinción y orden

Atributos de intervalo; Distinción, orden y adición

Atributos de razón: Distinción, orden, adición y multiplicación.

2.1.20. Redes neuronales

Son colecciones de nodos con entradas, salidas y procesamiento en cada nodo, entre la entrada y la salida existen un número de capas ocultas de procesamiento. La red neuronal debe de ser entrenada con un conjunto de datos de entrenamiento, una vez entrenada se utiliza para hacer prediciones.

Las redes neuronales se utilizan para clasificación y reconocimiento de patrones, aún así las redes neuronales no son muy utilizadas en la minería de datos, debido principalmente a dos razones:

Los métodos de clasificación de las redes neuronales ya entrenadas no son explicables

Los tiempos de entrenamiento son lentos e imprácticos cuando se trabaja con grandes volúmenes de datos.

Existen una gran variedad de arquitecturas y métodos de aprendizaje, los métodos de aprendizaje pueden ser supervisados y no supervisados.

2.2. Estudio de casos similares

En su artículo sobre la utilización de las redes bayesianas para la predicción del ingreso Hai Lu et al.. En su artículo sobre predicción del ingreso [6], describe como por lo general se tiene la creencia de que diversos factores pueden influir en la cantidad de ingreso al integrarse a la vida laboral, tales como la inteligencia, el género, la raza y el nivel educativo, realizan una referencia al trabajo realizado en el año de 1959 por Terman & Oden , en el cual se encontró que los adultos “talentosos” del sexo masculino obtienen un ingreso mayor que los del sexo femenino, igualmente hacen referencia al trabajo realizado por Karp & Morgan del año 1989 en el cual se validan los resultados de Terman & Oden, explican además que el estudio de Terman revelo que a diferentes niveles de educación y género pueden dar como resultado diferentes niveles de ingreso, el estudio realizado se

enfoca al grupo de adultos "talentosos", sin embargo puntualizan que el género y el nivel educativo pueden ser factores que influyen en el ingreso de los adultos.

Para propósito del estudio los autores definen "talentoso" a aquellas personas que han estado vinculados con algún programa de talentos.

Hai Lu et al.[6] expone que los datos utilizados son el resultado del estudio denominado "National Education Longitudinal Study" el cual fue realizado por el centro nacional para las estadísticas educativas, indican que la información fue recolectada en cinco ocasiones : 1988, 1990, 1992, 1994 y 2000, los sujetos de estudio fueron una muestra representativa a nivel nacional de estudiantes de octavo grado, los cuales fueron encuestados en el año inicial y en los levantamientos subsecuentes, las temáticas del cuestionario fueron las experiencias en la escuela, trabajo y hogar; recursos y apoyo educacional; el rol en la educación de sus padres y compañeros; características del vecindario; aspiraciones educativas y ocupacionales; y otras percepciones del estudiante. Los autores indican que solo usaron algunas variables del total de los datos disponibles para realizar su estudio.

Continúan exponiendo el modelo en la cual el ingreso es la variable respuesta, la cual fue tratada como una variable aleatoria continua, las variables predictivas utilizadas fueron:

- Talentoso: estuvo en algún programa de talentos.
- Sexo: Hombre/Mujer
- Raza: Raza del informante
- Educación: Nivel máximo de estudios obtenido hasta el año 2000.
- Horas trabajadas: número de horas trabajadas por el informante en el año de 1999.

Hai Liu et al. [6] exponen que las horas trabajadas pueden ser tratadas como una variable continua, en el caso de las otras variables son categóricas por lo cual se utilizaron variables falsas, indican que en su caso la variable de ingreso presenta una distribución sesgada a la izquierda, debido a lo cual se realizo una trasformación logarítmica a la variable ingreso, de tal forma que el modelo resultante es:

$$\text{Log(ingreso)} \sim 1+\text{Talentoso}+\text{Sexo}+\text{Raza}+\text{Educación}+\text{Horas trabajadas}$$

Dado que una total intercepción se incluye en el modelo, el número de variables dentro de cada predictor debe ser uno menos que el nivel de ese predictor, esto es corroborado en por la Universidad de Princeton mediante el servicio de datos y estadísticas [23], en la cual se indica que el número de valores será k-1, donde k es el número de categorías dentro de la variable, Por ejemplo, si existieran 6 niveles educación, se deben de crear 5 variables

falsas asociadas Con el predictor de la educación, asimismo en [23] se otra alternativa para el trabajo con variables falsas, la cual es crear tantas variables falsas como categorías existan en la variable y bajo un proceso iterativo eliminar una diferente en cada análisis realizado, si se opta por la primer alternativa, es decir obtener k-1 variables falsas, se debe de decidir que categoría no será convertida a variable falsa, esta decisión de cual nivel no se codificara tal como se expone en [23] es a menudo arbitraria. Indicando que el nivel que no está codificado es la categoría a la que todas las demás categorías se compararán. Una regla no arbitraría es la elección del grupo con mayor tamaño el cual será la categoría no codificada.

Los análisis realizados por Hai Liu et al. [6], indican que el modelo utilizado es consistente con lo estipulado por los autores, asimismo los autores realizan otro modelo de regresión bayesiano con información sobre los ingresos de la población en el año de 1998, sin embargo según lo exponen los autores, los resultados no difieren mucho del primer modelo desarrollado. Los autores continúan su estudio con la predicción de valores a través del modelo de regresión, indican que utilizaron R para generar una simulación de datos con algunas replicas, se crearon cuartiles para realizar la comparación de los resultados de la predicción, obteniendo resultados cercanos a los reales. Para finalizar los autores realizan una comparación del modelo utilizado con el modelo clásico, lo cual comparado con el modelo propuesto es muy similar, indican que los resultados son bastantes consistentes dados que en ambos modelos se utilizaron los mismos datos y solo tratan los parámetros desde diversos puntos de vista.

En su artículo sobre la estimación de los agregados municipales utilizando información muestral, Juan Hernández [5], expone que Algunos países con la infraestructura adecuada han basado su producción de información estadística a partir de poblaciones y subpoblaciones de registros administrativos, para otros países, dadas sus condiciones políticas, económicas y sociales hacen inviable este enfoque, para los estudiosos, consideran que el primer enfoque es el camino viable en un futuro y están realizando esfuerzos para ponerlo en marcha, otros en cambio, consideran impensable disponer algún día, de por lo menos un registro de población en el cual basar el enfoque [5].

Lo anterior ha conducido a algunas oficinas nacionales de estadística a explorar, y en algunos casos, iniciar la instrumentación del uso combinado o alternativo de fuentes para la generación de información estadística que puede ser usada de la misma forma que la censal, tales como:

- Censos y encuestas sociodemográficas (Estados Unidos)
- Registros administrativos y encuestas (Países bajos)
- Registros administrativos y censos (España)
- Encuestas rotatorias (Francia)

De estas experiencias de los países antes mencionados, surgió la decisión de utilizar para el censo de población un enfoque combinado, a través de la utilización de un cuestionario

básico, el cual fue aplicado al total de la población, y la aplicación de un cuestionario extendido a una muestra del operativo censal [5].

El cuestionario básico obtiene información sobre: el equipamiento de la vivienda, número de residentes en el hogar, características de las personas residentes del hogar, para las personas de 12 años y más: estado conyugal, condición de actividad, ocupación u oficio, situación en el trabajo, horas trabajadas, ingresos por productos del trabajo, actividad económica y para las Mujeres de 12 años cumplidos: número de hijos nacidos vivos, mortalidad infantil, mortalidad del último hijo nacido vivo [11], adicionalmente, en el mismo operativo censal se levanto una muestra de la población, a esta muestra se le aplico un cuestionario ampliado, el cual contiene las preguntas del cuestionario básico más la cobertura de otras temáticas.

Juan Hernández [5], Afirma que el INEGI considera que existe en México la capacidad de evaluar localmente las propuestas planteadas, así como proponer nuevas opciones, evaluarlas y determinar la viabilidad en el contexto de producción estadística [5].

Por regla, para obtener niveles de representatividad geográfica a detalle, supone el uso de grandes tamaños de muestra para conseguir la precisión adecuada, con lo cual se reducen los costos de levantamiento.

Como lo indica Juan Hernández en su artículo [5], la reducción de costos puede ser atendido al relacionar los resultados de las encuestas con otras fuentes: la censal y los registros administrativos; tomando en cuenta las experiencias internacionales así como los ejemplos de aplicación basados en información mexicana que pueden ser mejorados.

Juan Hernández en su artículo [5], continúa describiendo la necesidad de metodologías que permitan lo anterior, con lo cual se puede concebir un esquema de generación de información estadística en la cual se integran el Censo y las encuestas simultaneas que son levantadas durante el periodo intercensal para lograr:

- La reducción en los tamaños de muestra
- Reducir el número de preguntas en los cuestionarios
- Contar con información desagregada con una mayor frecuencia
- Mantener la precisión alcanzada por en los ejercicios muestrales

El logro de estos objetivos conllevaría ahorros sustanciales, tanto en el tiempo de levantamiento como en el costo del mismo.

En el modelo generado, para estimar los ingresos por trabajo a nivel de agregado municipal, utilizaron las siguientes variables:

- Escolaridad promedio de los jefes del hogar (variable discriminante)

- Número de personas que no usan el servicio médico público
- Número de personas que trabajaban
- Número de personas que trabajan como empleados
- Número de personas que trabajan por cuenta propia

Las preguntas a las cuales hacen referencia estos agregados son: Escolaridad, la cual se encuentra en el cuestionario básico y ampliado, Parentesco, la cual se encuentra tanto en el cuestionario básico como el ampliado, Condición de actividad, la cual se encuentra en ambos cuestionarios, Usos de servicios de salud, solo en cuestionario ampliado, Situación en el trabajo, la cual se encuentra en ambos cuestionarios.

Nick Winder y Yinghui Zhou en su artículo sobre predicción del ingreso de los individuos suizos [19], exponen que el ingreso anual es un factor importante en la ciencia social y política, debido a que muchos procesos y eventos están fuertemente correlacionados con el ingreso de los individuos y las familias, exponen que las ecuaciones de regresión por el modelo de microsimulación de Sverige, están basadas en un la regresión logística y necesitan un ajuste continuo para asegurar que la distribución simulada de ingreso corresponda con la observada en la realidad, ellos exponen, el desarrollo de un enfoque conceptual simple para la predicción del ingreso el cual permite la especificación de un modelo que puede seguir la secuencia histórica sin ajustes, tal modelo, si posible, puede ser usado para generar predicciones razonables para aplicaciones futuras para otros modelos de simulación.

Los autores indican que utilizaron información de cinco años referente a individuos de 11 años y más, lo cual les permite realizar un seguimiento del mercado de trabajo, lo cual les permite observar a las personas que se integran al mercado laboral, indican que las variables relevantes contenidas en las bases de datos son:

- Año de levantamiento
- Sexo
- Edad
- País de nacimiento
- Estado civil
- Año de inmigración
- Estatus ocupacional
- Estudia
- Nivel de educación
- Ingreso anual

En [19] continúan exponiendo el proceso para unir las diversas bases de datos, mediante el campo de identificación, aquellos registros que no tenían continuidad se eliminaron del estudio, los autores explican como utilizaron la regresión lineal simple con grupos de variables independientes de acuerdo al tipo de población y la variable dependiente: el salario anual, los resultados de las regresiones son alentadores, ya que los valores R indicaron valores entre el .89, lo cual indica que entre un 89% de la variación del ingreso es explicado por la ecuación de regresión.

Del texto es posible extraer el modelo utilizado, el cual en su forma más general se representa como:

Ingreso anual= F(ingresot-1,sexo, edad, escolaridad, nacionalidad, condición de actividad)

Los autores concluyen con observaciones sobre los distintos grupos de estudio y como se ven afectados por las variables utilizadas, además indican que la variable de educación para cada individuo es significativa.

Giuseppe Iarrossi en el libro "The power of survey design" [25], describe en su libro una guía práctica para la administración de cuestionarios, esta administración debe de ser llevada a cabo por una persona que tiene las habilidades para anticiparse a posibles fuentes de error, indica que en las etapas tempranas pueden incluir la revisión de la literatura, así como, platicas con expertos, lo cual permite conceptualizar problemas potenciales, así mismo, indica que una revisión del o los cuestionarios utilizados anteriormente así como las discusiones con los encargados del diseño, ayudará a determinar cuál es el mejor enfoque, cuales hipótesis serán probadas y que preguntas se adaptan mejor a un cuestionario especifico, una vez que los objetivos han sido identificados, el reto principal es la traducción de estos en un cuestionario metodológica y bien conceptualizado cuestionario, el desarrollo del cuestionario empieza después de que los planes generales han sido escritos y termia solo unos días antes de que comience el trabajo de campo, el cuestionario inicial es revisado muchas veces, la prueba piloto es un componente crítico del diseño del cuestionario, así como las sesiones de capacitación a los enumeradores debe de considerarse en el diseño del cuestionario, dado que en general, permite identificar problemas de fraseado y traducción., la recolección de datos de alta calidad en tiempo depende de cómo el operativo este organizado, la coordinación de todos los involucrados es una tarea vital y compleja.

Que tan bien este diseñado un cuestionario impactará en el tiempo de la entrevista, el uso apropiado de patrones de salto y claridad en las definiciones y los enunciados no solo reducirá el tiempo de la entrevista, si no también asegura datos de calidad.

Un factor importante es la determinación de la tasa de no respuesta ya que esta influye en el tiempo de la entrevista, un cuestionario con una tasa no respuesta se lleva mucho menos tiempo que un cuestionario con una tasa de no respuesta del 10 por ciento.

Giuseppe Larossi en su libro sobre mejores prácticas del diseño de cuestionarios [25], indica un principio general para el mejoramiento del cuestionario, al realizar una pregunta se debe de considerar la relevancia (que tan familiar le resulta al entrevistado la pregunta) y la validez (obtiene la información para la que fue creada).

Así mismo Giuseppe Larossi [25], expone una serie de guías prácticas para el diseño de cuestionarios:

Palabras de la pregunta, el cambio de una palabra en la pregunta puede alterar significativamente la distribución de la respuesta.

Brevedad, solo que la pregunta sea relevante para la investigación, debe de ser incluida en el cuestionario, a un nivel de detalle mayor, al menos que una palabra sea relevante para la pregunta debe de ser incluida.

Como regla general, una pregunta no debe de exceder de 20 palabras y no debe de tener más de tres comas.

La brevedad puede ser alcanzada mediante la realización de una pregunta a la vez, se debe de evitar el uso de preguntas ocultas, es decir preguntas que implícitamente determinan su relevancia con otras preguntas, en caso que se desee captar esa información la recomendación es la realización las preguntas de forma separada.

Objetivas, una pregunta no objetiva tiene la característica principal que sugieren una respuesta, se debe de evitar el uso de preguntas capciosas, las cuales por su contenido, estructura y fraseado, llevan al informante hacia una dirección de respuesta o sugestión.

3. METODOLOGÍA PARA EL DESARROLLO DEL PROYECTO

3.1. Proceso estándar interrelacionado para la minería de datos (CRISP-DM)

La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos (Ver la figura 2.).

En el nivel superior, el proceso de minería de datos es organizado en un número de fases; cada fase consiste de varias tareas genéricas de segundo nivel.

Este segundo nivel lo llaman genérico porque está destinado a ser bastante general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible. Completo significa que cubre tanto al proceso entero de minería de datos y todas las aplicaciones de minería de datos posibles. Estable significa que el modelo debería ser válido para acontecimientos normales y aún para desarrollos imprevistos como técnicas de modelado nuevo.

El tercer nivel, el nivel de tarea especializado, es el lugar para describir como las acciones en las tareas genéricas deberían ser realizadas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe como esta tarea se diferencia en situaciones diferentes, como la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo.

La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos.

En la práctica, muchas de las tareas pueden ser realizadas en una orden diferente, y esto a menudo será necesario volver a hacer tareas anteriores repetidamente y repetir ciertas acciones. El modelo de proceso no intenta capturar todas estas posibles rutas del proceso de la minería de datos porque se requeriría un modelo de proceso demasiado complejo.

El cuarto nivel, la instancia de proceso, es un registro de las acciones, decisiones, y de los resultados de una minería de datos real contratada.

Una instancia de proceso está organizada según las tareas definidas en los niveles más altos, pero representa lo que en realidad pasó en un contrato particular más bien que lo que pasa en general.

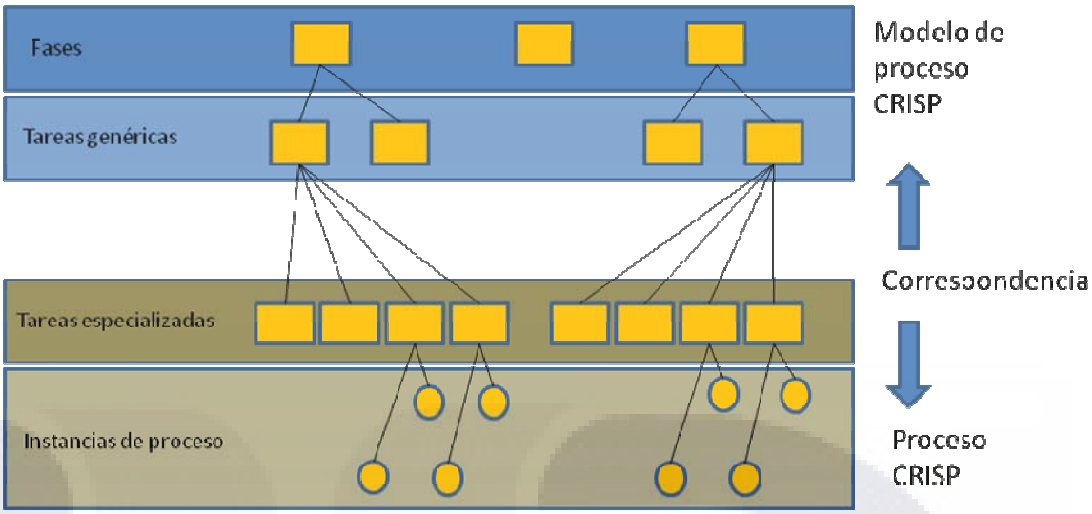


Figura 2: Cuatro niveles de interrupción de la metodología CRISP-DM

Para la solución de este caso de estudio se propone la utilización del proceso industrial estándar interrelacionado para la minería de datos (CRISP-DM) por sus siglas en inglés, la cual es una metodología compuesta por 6 fases principales.

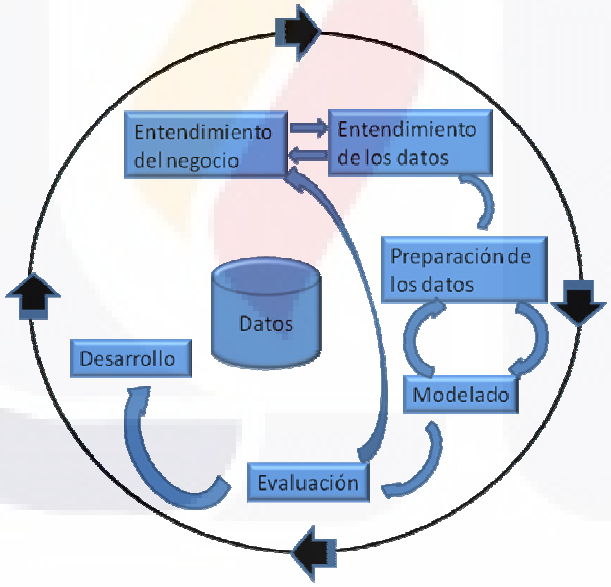


Figura 3: Fases del modelo de referencia CRISP-DM

A continuación se describen cada una de sus fases de forma detalla.

3.1.1. Comprendiendo el negocio

Determinación de los objetivos del negocio

Tarea **Determinar los objetivos de negocio**

El primer objetivo del analista es comprender a fondo, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo el cliente tiene muchos objetivos y restricciones que compiten, las cuales deben ser correctamente balanceadas. El objetivo del analista es encontrar factores importantes en el principio del proyecto dado que esto puede influir en el resultado final. Una consecuencia probable de descuidar este paso es la realización de un gran esfuerzo de producir las respuestas correctas a las preguntas incorrectas.

Salida **Antecedentes**

Coteje la información que conoce sobre la situación de negocio de la organización al principio del proyecto. Estos detalles no sólo sirven para identificar más estrechamente los objetivos de negocio a ser alcanzados, pero también sirven para identificar los recursos, tanto humano como material, que puede ser usado o sea necesario durante el curso del proyecto.

Actividades **Organización**

- Desarrollar organigramas que identifiquen divisiones, departamentos, y grupos de proyectos. El organigrama debería también:
- Identificar los nombres de los gerentes y sus responsabilidades;
- Identificar a personas claves en el negocio y sus roles;
- Identificar a un patrocinador interno (el patrocinador financiero y el experto primario del dominio de usuario).
- Indicar si hay un comité de dirección y lista de miembros.
- Identificar las unidades de negocio que son afectadas por el proyecto de minería de datos (por ejemplo, el Control de comercialización, Ventas, Finanzas).

Área del problema

- Identificar el área del problema (por ejemplo, el control de comercialización, el cuidado de cliente, el desarrollo comercial, etc.)
- Describir el problema en términos generales.
- Comprobar el estado actual del proyecto (por ejemplo, Comprobar si ya está claro que dentro de la unidad de negocio un proyecto de minería de datos debe ser realizado, o si la minería de datos necesita ser promovida como una tecnología clave en el negocio).
- Clarificar los requisitos previos del proyecto (por ejemplo, ¿Cuál es la motivación del proyecto? ¿La minería de datos ya está siendo usada en el negocio?).

- Si es necesario, preparar presentaciones y demostraciones de minería de datos para el negocio.
- Identificar grupos de objetivos para el resultado de proyecto (por ejemplo, ¿Esperamos entregar un informe para la dirección superior o un sistema operacional para ser usado por usuarios finales inexpertos?).
- Identificar las necesidades de los usuarios y sus expectativas.

Solución actual

- Describir cualquier solución usada actualmente para dirigir el problema.
- Describir las ventajas y las desventajas de la solución corriente y el nivel al que esto es aceptado por los usuarios.

Salida **Objetivos de negocio**

Describir el objetivo principal del cliente, desde una perspectiva de negocio. Además del objetivo de negocio primario, hay típicamente un gran número de preguntas relacionadas al negocio a las que al cliente le gustaría dirigir. Por ejemplo, el objetivo primario de negocio podría ser mantener a clientes actuales por predicción cuando ellos son propensos a moverse a un competidor, mientras un objetivo secundario de negocio podría ser de determinar si precios (comisiones) inferiores afectan sólo un segmento particular de clientes.

Actividades

- De manera informal describir el problema a ser solucionado.
- Especificar todas las preguntas de negocio tan precisas como sea posible.
- Especificar cualquier otras exigencias de negocio (por ejemplo, el negocio no quiere perder a ningún cliente).
- Especificar las ventajas esperadas en términos de negocio.

¡Cuidado! Se debe de tener cuidado de establecer objetivos inalcanzables hechos por ellos tan realistas como sea posible.

Salida **Criterios de éxito del negocio**

Describir los criterios para un resultado exitoso o útil al proyecto desde el punto de vista del negocio.

Esto podría ser bastante específico y fácilmente medible, como una reducción de cliente a un cierto grado, o general y subjetivo, como "dar ideas útiles en las relaciones". En el caso último, esté seguro de indicar quien haría el juicio subjetivo.

Actividades

- Especificar criterios de éxito de negocio (por ejemplo, Mejorar la tasa de respuesta en una campaña de correo en el 10 por ciento y marcar la tasa en el 20 por ciento).
- Identificar quien evalúa los criterios de éxito.

¡Recuerde! Cada uno de los criterios de éxito debería relacionarse con al menos uno de los objetivos especificados de negocio.

¡Buena Idea! Antes del comienzo de la evaluación de situación, usted podría analizar las experiencias anteriores de este problema- Internamente, usando CRISP-DM, o externamente, usando soluciones pre-empaquetadas.

Evaluación de la situación

Tarea **Evaluar la situación**

Esta tarea implica una investigación más detallada sobre todos los recursos, restricciones, presunciones, y otros factores que deberían ser considerados en la determinación del objetivo de análisis de datos y en el desarrollo del plan de proyecto.

Salida **Inventario de recursos**

Listar los recursos disponibles para el proyecto, incluyendo el personal (expertos de datos y de negocios, soportes técnicos, expertos en minería de datos), datos (extracciones fijas, acceso a datos existentes en almacenes de datos u operacionales), recursos computacionales (plataformas de hardware), y software (instrumentos de minería de datos, otros software relevantes).

Actividades **Recursos de Hardware**

- Identificar el hardware básico.
- Establecer la disponibilidad del hardware básico para el proyecto de minería de datos.
- Comprobar si la planificación del mantenimiento de hardware se opone a la disponibilidad del hardware para el proyecto de minería de datos.
- Identificar el hardware disponible para ser usado por la herramienta de minería de datos (si el instrumento es conocido en esta etapa).

Fuentes de datos y conocimientos

- Identificar las fuentes de datos.
- Identificar el tipo de fuentes de datos (fuentes en línea, expertos, documentación escrita, etc.).
- Identificar fuentes de conocimiento.
- Identificar el tipo de fuentes de conocimientos (fuentes en línea, expertos, documentación escrita, etc.).
- Comprobar herramientas disponibles y técnicas.
- Describir el conocimiento de generalidades relevante (de manera informal o formalmente).

Fuentes de personal

- Identificar al patrocinador de proyecto (si difiere del patrocinador interno).
- Identificar al administrador de sistema, el administrador de base de

datos, y el personal de soporte técnico para futuras preguntas.

- Identificar a analistas de mercado, los expertos en minería de datos, y estadísticos, y comprobar su disponibilidad.
- Comprobar la disponibilidad de expertos de dominio para fases posteriores.

¡Recuerde! Recuerde que el proyecto puede necesitar personal técnico en cualquier momento en todas partes del proyecto, por ejemplo durante la transformación de datos.

Salidas **Requerimientos, presunciones, y restricciones**

Listar todos los requerimientos del proyecto, incluyendo la planeación de la terminación, la comprensibilidad, y la calidad y seguridad de los resultados, así como cuestiones legales. Como la parte de esta salida, asegúrese de que le permiten usar los datos.

Listar las presunciones hechas por el proyecto. Estos pueden ser presunciones sobre los datos, que pueden ser verificados durante la minería de datos, pero también puede incluir presunciones no comprobables relacionadas con el proyecto. Esto es en particular Importante de ponerlos en una lista si ellos afectarán la validez de los resultados.

Listar las restricciones hechas en el proyecto. Estas restricciones podrían implicar la carencia de recursos para terminar algunas tareas en el proyecto en el tiempo requerido, o allí pueden ser restricciones legales o éticas sobre el uso de los datos o la solución necesita terminar la tarea de minería de datos.

Actividades **Requerimientos**

- Especificar el perfil del grupo objetivo.
- Capturar todos los requerimientos en la planificación.
- Capturar los requerimientos de comprensibilidad, exactitud, desarrollar habilidades, mantenimiento, y repetitividad del proyecto de minería de datos y los modelos resultantes.
- Capturar los requerimientos de seguridad, restricciones legales, de privacidad, información, y planificación de proyecto.

Presunciones

- Aclarar todas las presunciones (incluyendo las implícitas) y las hechas por ellos explícitamente (por ejemplo, dirigir las cuestiones de negocio, a un número mínimo de clientes con la edad por encima de 50 es necesaria).
- Listar las presunciones sobre calidad de datos (por ejemplo, exactitud, disponibilidad).
- Listar las presunciones sobre factores externos (por ejemplo, cuestiones económicas, productos competitivos, avances técnicos).
- Aclarar presunciones que conducen a cualquiera de las estimaciones (por ejemplo, el precio de un instrumento específico es asumido para ser menor que $1,000 dólares).
- Listar todas las presunciones en cuanto a si es necesario entender y describir o explicar el modelo (Por ejemplo, como el modelo y los resultados son presentados a la dirección / patrocinador).

Restricciones

- Comprobar restricciones generales (por ejemplo, cuestiones legales, presupuesto, escalas de tiempo, y recursos).
- Comprobar el correcto acceso a fuentes de datos (por ejemplo, restricciones de acceso, la contraseña requerida).
- Comprobar la accesibilidad técnica de datos (los sistemas de operaciones, el sistema de administración de datos, el formato de archivo y de base de datos).
- Comprobar si el conocimiento relevante es accesible.
- Comprobar restricciones de presupuesto (gastos fijos, gastos de implementación, etc.).

¡Recuerde! La lista de presunciones también incluye presunciones al principio del proyecto, esto es, lo que el punto de inicio del proyecto se ha realizado.

Salidas **Riesgos y contingencias**

Listar los riesgos, es decir los acontecimientos que podrían ocurrir, impactando en la planificación, el costo, o el resultado. Listar los planes de contingencias respectivos: que acción será tomada para evitar o reducir al mínimo el impacto o recuperar de la ocurrencia de los riesgos previstos.

Actividades **Identificar riesgos**

- Identificar riesgos de negocio (por ejemplo, el competidor aparece primero con mejores resultados).
- Identificar riesgos de organización (por ejemplo, el departamento que solicita el proyecto no tiene financiamiento para el proyecto).
- Identificar riesgos financieros (por ejemplo, aumentar el financiamiento depende de los resultados iniciales de minería de datos).
- Identificar riesgos técnicos.
- Identificar los riesgos que dependen de datos y de las fuentes de datos (por ejemplo, la mala calidad y cobertura).

Desarrollo de planes de contingencia.

- Determinar condiciones en las que cada riesgo puede ocurrir.
- Desarrollar planes de contingencia.

Salida **Terminología**

Compilar un glosario de terminología relevante al proyecto. Esto debería incluir al menos dos componentes:

- Un glosario de terminología relevante de negocio, que forma parte de la comprensión de negocio disponible al proyecto.
- Un glosario de terminología de minería de datos, ilustrada con ejemplos relevantes al problema de negocio en cuestión.

Actividades Comprobar la disponibilidad previa de glosarios; si no comience a bosquejar glosarios.

- Hablar a expertos de dominio para entender su terminología.
- Familiarizarse con la terminología de negocio.

Salida **Costos y beneficios**

Preparar un análisis de costo-beneficio para el proyecto, comparando los gastos del proyecto con el beneficio potencial para el negocio si esto es exitoso.

Actividades **Estimar el costo para la colección de datos**

- Estimar el costo de desarrollo y realización de una solución.
- Identificar beneficios (por ejemplo, mejorar la satisfacción del cliente, ROI, y el aumento de las ganancias).
- Estimar gastos de operación.

¡Buena Idea! La comparación debería ser tan específica como sea posible, como esto permite un mejor caso de negocio para ser realizado.

¡Cuidado! Acuérdese de identificar costos ocultos, como la extracción y preparación repetida de datos, cambios en los procesos laborales, y tiempo requerido para el entrenamiento.

Determinar objetivos de minería de datos

Tarea **Determinar objetivos de minería de datos**

Un objetivo de negocio declara objetivos en la terminología de negocio; un objetivo de minería de datos declara objetivos de proyecto en términos técnicos. Por ejemplo, el objetivo de negocio podría ser, "Aumentar la venta por catalogo a clientes existentes", mientras un objetivo de minería de datos podría ser, "Predecir cuantas baratijas comprará un cliente, considerando sus compras durante los tres años pasados, información demográfica relevante, y el precio del artículo."

Salidas **Objetivos de minería de datos**

Describir las salidas planeadas del proyecto que permiten el logro de los objetivos de negocio.

Note que estas son salidas normalmente técnicas.

Actividades Traducir las preguntas de negocio a objetivos de minería de datos (por ejemplo, una campaña de control de comercialización requiere la segmentación de clientes para decidir a quién acercarse en esta campaña; el nivel/tamaño de los segmentos debería ser especificado).

Especificar datos tipo de problema de minería de datos (por ejemplo, la clasificación, la descripción, la predicción, y agrupamiento).

¡Buena idea! Puede ser sabio redefinir el problema. Por ejemplo, modelar la retención de producto más que la retención del cliente cuando la retención del cliente entrega resultados muy tarde para afectar la salida.

Salida **Criterios de éxitos de minería de datos**

Definir los criterios para un resultado acertado para el proyecto en términos técnicos, por ejemplo un cierto grado de exactitud predictiva o un perfil de propensión-a-comprar con un nivel dado "elevación".

Como con los criterios de éxitos del negocio, puede ser necesario describir estos en términos subjetivos, en el caso de que la persona o las personas que hacen el juicio subjetivo debieran ser identificadas.

Actividades Especificar los criterios para evaluar el modelo (por ejemplo, la exactitud del modelo, el funcionamiento y la complejidad).

- Definir el patrón de pruebas para los criterios de evaluación.
- Especificar las reglas que dirigen criterios de evaluación subjetivos (por ejemplo, el habilidad de explicar del modelo y de los datos y la comprensión de mercadeo proporcionada por el modelo).

¡Tenga cuidado! Recuerde que los datos que extraen criterios de éxito son diferentes a los criterios de éxito de negocio definidos antes.

Recuerde es sabio planear para el desarrollo desde el principio del proyecto.

Realizar la planeación del proyecto

Tarea **Producir el plan del proyecto**

Describir el plan propuesto para alcanzar los objetivos de minería de datos y así alcanzar de los objetivos de negocio.

Salida **Plan del Proyecto**

Listar las etapas para ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias. En cualquier parte donde posible, haga explícito las iteraciones en gran escala en el proceso de minería de datos- Por ejemplo, las repeticiones del modelado y fases de evaluación. Como parte del plan de proyecto, esto es también importante analizar dependencias entre el planeamiento de los tiempos y los riesgos. Marcar los resultados de estos análisis explícitamente en el plan de proyecto, idealmente con acciones y recomendaciones para actuar si los riesgos son manifestados.

Aunque esto sea la única tarea en la que el plan de proyecto directamente es llamado, sin embargo debería ser consultado continuamente y repasado en todas partes del proyecto. Deberían consultar el plan de proyecto como mínimo siempre que una tarea nueva sea comenzada o una iteración futura de una tarea o una actividad está comenzando.

Actividades Definir el plan de proceso inicial y hablar de la viabilidad con todo el personal incluido.

- Combinar todos los objetivos identificados y técnicas seleccionadas en un procedimiento coherente que solucione las cuestiones del negocio y encuentre los criterios de éxito de negocio.
- Estimar el esfuerzo y los recursos necesarios para alcanzar y desarrollar la solución. (Es útil considerar la experiencia de otras personas estimando escalas de tiempo para proyectos de minería de datos. Por ejemplo, es a menudo presumido que el 50-70 por ciento del tiempo y el esfuerzo en un proyecto de minería de es usado en la Fase de Preparación de Datos, mientras que solo un 20-30 por ciento es usado en la Fase de Comprensión de Datos, mientras que solo un 10-20 por ciento es gastado en cada uno de las Fase de Modelado, Evaluación, y Comprensión del Negocio Entendiendo y el 5-10 por ciento en la Fase de Desarrollo.).
- Identificar pasos críticos.
- Marcar los puntos de decisión.
- Marcar los puntos de revisión.
- Identificar las principales iteraciones.

Salida **Evaluación de Inicial de herramientas y técnicas**

Al final de la primera fase, el equipo de proyecto realiza una evaluación inicial de herramientas y técnicas. Aquí, es importante seleccionar una herramienta de minería de datos que soporte varios métodos para las diferentes etapas del proceso, ya que la selección de herramientas y técnicas puede influir en el proyecto entero.

Actividades Crear una lista de criterios de selección para herramientas y técnicas (o usar uno existente si está disponible).

- Escoger herramientas y técnicas posibles.
- Evaluar la adecuación de técnicas.
- Revisar y priorizar técnicas aplicables según la evaluación de soluciones alternativas.

3.1.2. Comprensión de los datos

Recolección de los datos iníciales

Tarea **Recoger datos iníciales**

Obtener los datos (o el acceso a los datos) listados en los recursos de proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de datos. Por ejemplo, si usted tiene la intención de usar una herramienta específica para comprender los datos, es lógico cargar sus datos en esta herramienta.

Salida **Informe de la recolección de datos inicial**

Describir toda la variedad de datos usados para el proyecto, e incluya cualquier requerimiento de selección para datos más detallados. El informe de colección de datos también debería definir si algunos atributos son relativamente más importantes que otros.

Recuerde que cualquier evaluación de calidad de datos debería ser hecha no solamente de las fuentes de datos individuales, pero también de algunos datos que son resultado de fuentes de datos que se combinan. Por inconsistencias entre las fuentes, los datos combinados pueden presentar los problemas que no existen en las fuentes de datos individuales.

Actividades **Planificación de requerimientos de datos**

Planee que información es necesaria (por ejemplo, sólo para atributos determinados, o la información adicional específica).

Comprobar si toda la información necesaria (para resolver los objetivos de la minería de datos) está en realidad disponible.

Criterios de selección

Especificar los criterios de selección (por ejemplo, ¿Qué atributos son necesarios para los objetivos específicos de minería de datos? ¿Qué atributos han sido identificados como no pertinentes? ¿Cuántos atributos podemos manejar con las técnicas escogidas?)

- Elegir tablas/archivos de interés.
- Elegir datos dentro de una tabla/archivo.
- Pensar cuanto tiempo de una historial habría que usar (por ejemplo, si 18 meses de datos están disponibles, sólo 12 meses pueden ser necesarios para el ejercicio).

¡Tenga cuidado! Estar consciente de que los datos recolectados de diferentes fuentes pueden dar lugar a problemas de calidad cuando sean combinados (Por ejemplo, los archivos de dirección combinados con una base de datos de cliente pueden mostrar inconsistencias de formato, invalidez de datos, etc.).

Inserción de datos Si los datos contienen libre entradas de texto, ¿tenemos que codificarlos para modelar o necesitamos agruparlos en entradas específicas?

- ¿Cómo podemos encontrar atributos omitidos?

- ¿Cómo podemos mejorar la extracción los datos?

¡Buena Idea! Recordar que algún conocimiento sobre los datos puede estar disponible de fuentes no-electrónicas (Por ejemplo, de gente, de texto impreso, etc.).

Recordar que puede ser necesario a preproceso de los datos (datos de serie tiempo, promedios ponderados, etc.).

Descripción de los datos

Tarea **Describir datos**

Examine las propiedades "gruesas" de los datos obtenidos y el informe sobre los resultados.

Salida **Informe de descripción de datos**

Descripción de los datos que han sido obtenidos, incluyendo el formato de los datos, la cantidad de los datos (Por ejemplo, el número de registros y campos internos de cada tabla), las identidades de los campos, y cualquier otro rasgo superficial que haya sido descubierto.

Actividades **Análisis Volumétrico de datos**

- Identificar datos y métodos de captura.
- Acceder a las fuentes de datos.
- Usar análisis estadísticos si es apropiado.
- Reportar las tablas y sus relaciones.
- Compruebe el volumen de datos, el número de múltiplos, la complejidad.
- Notar si los datos contienen entradas de texto libres.

Atributo tipos y valores

- Comprobar la accesibilidad y disponibilidad de atributos.
- Comprobar los tipos de atributos (numérico, simbólico, la taxonomía, etc.)
- Comprobar el rango de valores de los atributos.
- Analizar los atributos correlativos (correlaciones de atributo).
- Comprender el significado de cada atributo y clasificar (describir) el valor en términos de negocio.
- Para cada atributo, calcular la estadística básica (por ejemplo, calcule la distribución, el promedio, el máximo, el mínimo, la desviación estándar, la varianza, la moda, la inclinación, etc.)
- Analizar la estadística básica y relacionan los resultados con su significado en términos de negocio.
- Decidir si el atributo es relevante para los objetivos específicos de la minería de datos.
- Determinar si el significado del atributo es usado coherentemente (conscientemente).
- Entrevistar a expertos de dominio para obtener su opinión sobre la importancia de los atributos.
- Decidir si es necesario equilibrar los datos (basado en las técnicas que

modelan a ser usado).

Claves

- Analizar relaciones claves.
- Comprobar la cantidad de coincidencias entre valores de atributos claves a través de tablas.

Revisión de Objetivos/Presunciones

- Actualizar la lista de presunciones, si es necesario.

Exploración de datos

Tarea **Explorar datos**

Esta tarea aborda las preguntas de minería de datos que pueden ser dirigidas usando la interrogación, la visualización, y técnicas de informe. Estos análisis pueden directamente dirigir los objetivos de minería de datos. Sin embargo, ellos pueden también contribuir a refinar la descripción de datos e informes de calidad, y alimentar internamente la transformación y otros pasos de preparación de datos necesario antes de que pueda ocurrir un futuro análisis.

Salida **Informe de exploración de datos**

Describir los resultados de esta tarea, incluyendo las primeras conclusiones o las hipótesis iníciales y su impacto sobre el resto del proyecto. El informe también puede incluir gráficos y diseños (plots) que indican las características de los datos o los puntos de interés de subconjuntos de datos dignos de una futura investigación.

Actividades **Actividades Exploración de Datos**

- Analizar en detalles las propiedades de atributos interesantes (por ejemplo, la estadística básica, las sub-poblaciones interesantes).
- Identificar las características de las sub-poblaciones.

Formar suposiciones para análisis futuro

- Considerar y evalúan la información y conclusiones en el informe de descripciones de datos.
- Formar una hipótesis e identifican acciones.
- Transforman la hipótesis en un objetivo de minería de datos, si es posible.
- Aclarar objetivos de minería de datos o hacerlos más exactos. Una búsqueda "ciega" no es necesariamente inútil, pero una búsqueda más dirigida hacia objetivos de negocio es preferible.
- Realizar un análisis básico para verificar la hipótesis.

Verificación de la calidad de los datos

Tarea **Verificar la calidad de datos**

Examine la calidad de los datos, dirigiendo preguntas como: Es los datos completos (¿esto cubre todos los casos requeridos?) ¿Hay en ellos errores o ellos contienen errores? ¿Si hay errores, como son ellos? ¿Hay valores omitidos en los datos? Si es así, ¿cómo son representados, donde ocurren, y como son ellos?

Salida **Informe de calidad de datos**

Listar los resultados de la verificación de calidad de datos; si hay problemas de calidad, Listar las posibles soluciones.

Actividades **Identificar valores especiales y catalogar su significado**

Revisión de atributos claves.

Comprobar la cobertura (por ejemplo, si todos los valores posibles son representados).

Comprobar las claves.

Verificar que los significados de los atributos y valores contenidos se satisfacen simultáneamente.

Identificar atributos omitidos y campos en blanco.

Establecer el significado de datos que faltan o fallan.

Comprobar los atributos con los valores diferentes que tienen significados similares (por ejemplo, la grasa baja, la dieta).

Comprobar la ortografía y el formato de valores (por ejemplo, mismo valor pero a veces comienza con una letra minúscula, a veces con una letra mayúscula).

Comprobar las desviaciones, y deciden si una desviación es "ruido" o puede indicar un fenómeno interesante.

Comprobar la plausibilidad de valores, (por ejemplo, todos los campos que tienen el mismo o casi los mismos valores).

¡Buena idea! Repasar cualquiera de los atributos que dan respuestas que están en desacuerdo con el sentido común (por ejemplo, adolescentes con altos niveles de ingreso).

Use gráficos de visualización, histogramas, etc. para revelar inconsistencias en los datos.

Calidad de datos en archivos planos

Si los datos son almacenados en archivos planos, comprobar que delimitador es usado y si esto es usado coherentemente en todos los atributos.

- Si los datos son almacenados en archivos planos, comprobar el número de campos en cada registro para ver si ellos coinciden.

Ruido e inconsistencias entre fuentes

- Comprobar consistencia y superabundancia entre fuentes diferentes.
- Planear para tratar el ruido.
- Descubrir el tipo de ruido y que atributos son afectados.

¡Buena idea! Recuerde que puede ser necesario excluir algunos datos ya que ellos no exponen comportamiento positivo o negativo (por ejemplo, al comprobar en el comportamiento del préstamo de clientes, excluye a todo los que nunca han tomado prestado, aquellos que no financian una hipoteca de casa, aquellos cuya hipoteca se acerca a la madurez, etc.).

Revisar si las presunciones son válidas o no, considerando la información real o actual en los datos y el conocimiento de negocio.

Preparación de los datos

Salida **Conjunto de datos**

Estos son los conjuntos de dato(s) producidos por la fase de preparación de datos, usada para modelar o para el trabajo de análisis principal del proyecto.

Salida **Descripción del conjunto de datos**

Esto es la descripción del conjunto de datos(s) usado para el modelado o para el trabajo de análisis principal del proyecto.

Datos seleccionados

Tarea **Seleccionar datos**

Decidir los datos a ser usados para el análisis. Los criterios incluyen la importancia a los objetivos de minería de datos, la calidad, y las restricciones técnicas como los límites en el volumen de datos o en los tipos de datos.

Salida **Razonamiento para inclusión/exclusión**

Listar los datos a ser usados / excluidos y los motivos para estas decisiones.

Actividades

- Recoger datos adicionales apropiados (de diferentes fuentes - internos así como externos).
- Realizar las pruebas de importancia y correlación para decidir si los campos son incluidos.
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de calidad de los datos y en la exploración de datos (esto es, puede desear incluir/excluir otros juegos de datos).
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de experiencia de modelado (esto es, la evaluación del modelo puede mostrar que otros conjuntos de datos son necesarios).
- Seleccionar diferentes subconjuntos de datos (por ejemplo, atributos diferentes, sólo los datos que encuentran ciertas condiciones).
- Considerar el uso de técnicas de muestreo (por ejemplo, una solución rápida puede implicar la prueba dura y el entrenamiento del conjunto de datos o la reducción del tamaño de la conjunto de datos de prueba, si la herramienta no puede manejar conjunto de datos llenos. Esto puede también ser útil para tener muestras ponderadas para dar la distinta importancia a atributos diferentes o valores diferentes del mismo atributo.)
- Documentar el razonamiento para la inclusión/exclusión.
- Comprobar técnicas disponibles para el muestreo de datos.

¡Buena idea! Basado en Criterios de Selección de Datos, decidir si uno o más atributos son más importantes que otros de acuerdo al correspondiente peso de los atributos. Decidir, basado en el contexto (esto es, el uso, la herramienta, etc.), como debe manejarse con el peso.

Limpieza de Datos

Tarea **Limpiar datos**

Elevar la calidad de datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de los datos, la inserción de faltas apropiadas, o técnicas más ambiciosas como la estimación de datos omitidos por modelado.

Salida **Informe de la limpieza de datos**

Describir las decisiones y las acciones que fueron tomados para dirigir los problemas de calidad de datos informados durante la Tarea de Verificación de Calidad de Datos. Si los datos están para ser usados en el ejercicio de minería de datos, el informe debería dirigir cuestiones de calidad de datos excepcionales y el efecto posible que esto podría tener sobre los resultados.

Actividades Reconsiderar como tratar con cualquier tipo de ruido observado

- Corregir, remover, o ignorar el ruido.
- Decidir cómo tratar con valores especiales y su significado. El área de valores especiales puede dar lugar a muchos resultados extraños y con cuidado deberían ser examinados. Los ejemplos de valores especiales podrían surgir por los resultados tomados de una revisión donde algunas cuestiones no fueron preguntadas o no

fueron contestadas. Esto podría terminar en un valor de99 para datos desconocidos. Por ejemplo, 99 para estado civil o afiliación política. Los valores especiales también podría surgir cuando los datos son truncados por ejemplo., 00 para gente de100 años o para todos los coches con 100,000 kilómetros en el odómetro.

- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias del os datos limpiados (esto es, usted puede desea incluir/excluir otros conjuntos de datos).

¡Buena idea! Recuerde que algunos campos pueden ser irrelevantes a los objetivos de minería de datos y, por lo tanto, el ruido en aquellos campos no tiene ninguna importancia. Sin embargo, si el ruido es ignorado por estos motivos, esto debería ser totalmente documentado como circunstancias que pueden cambiarse más tarde.

Construcción de Datos

Tarea **Construir datos**

Esta tarea incluye la construir de operaciones de preparación de datos tales como la producción de atributos derivados, completar registros nuevos, o transformar valores para atributos existentes.

Actividades Comprobar los mecanismos de construcción disponibles con la lista de herramientas sugeridas para el proyecto.

- Decidir sí es lo mejor para realizar la construcción dentro de la herramienta o fuera de ella (esto es, que es más eficiente, exacto, repetible).
- Reconsiderar Criterios de Selección de Datos en la luz de las experiencias de construcción de datos (esto es, usted puede desear incluir/excluir otros conjuntos de datos).

Salida **Atributos derivados**

Los atributos derivados son los atributos nuevos que son construidos de uno o atributos más existentes en el mismo registro. Un ejemplo podría ser: área = largo * ancho.

¿Por qué deberíamos tener que construir atributos derivados durante el curso de una investigación de minería de datos? No debería pensarse que sólo los datos de bases de datos u otras fuentes deberían ser usados en la construcción de un modelo. Los atributos derivados podrían ser construidos porque:

- El conocimiento del contexto nos convence que algún hecho es importante y debería ser representado aunque no tengamos ningún atributo actualmente para representarlo.
- El algoritmo de modelado en uso maneja los sólo ciertos tipos de datos -por ejemplo estamos usando regresión lineal y sospechamos que hay ciertas no-linealidades que serán incluidos en el modelo.
- El resultado de la fase de modelado sugiere que ciertos hechos no sean cubiertos.

Actividades **Derivar atributos**

- Decidir si cualquier atributo puede ser normalizado (por ejemplo, usando un algoritmo de agrupamiento (clustering) con el periodo y el ingreso, en ciertas divisas, el ingreso se controlará).
- Considerar agregar nueva información sobre la importancia relevante de los atributos para agregar de nuevos atributos (Por ejemplo, atributos con peso, normalización ponderada).
- ¿Cómo se puede construir o imputar atributos faltantes? [Decidir el tipo de construcción (por ejemplo, la combinación, el promedio, la inducción).]
- Agregar atributos nuevos a los datos acceso de acceso.

¡Buena idea! Antes de agregar Atributos Derivados, intente determinar si y como facilitan el proceso de modelado o facilitan el algoritmo de modelado. Quizás "el ingreso por persona" es un mejor/más fácil atributo para usar que "el ingreso por hogar." No elimine atributos simplemente para reducir el número de atributos de entrada.

Otro tipo de atributo derivado es la transformación de un atributo individual, por lo general realizado para cubrir las necesidades de las herramientas de modelado.

Actividades **Transformaciones de atributo individual**

- Especificar los pasos de transformaciones necesarias en los términos de facilitar las transformación disponibles (por ejemplo, cambiar un valor almacenado de un atributo numérico).
- Realizar pasos de transformación.

¡Buena idea! Las transformaciones pueden ser necesarias para cambiar rangos a campos simbólicos (por ejemplo, años a rangos de edad) o campos simbólicos ("definitivamente sí", "sí", "no se sabe," "no") a valores numéricos. Las herramientas de modelado o los algoritmos a menudo los requieren.

Salida **Registros generados**

Los registros generados son registros completamente nuevos, que agregan nuevo conocimiento o representan nuevos datos que de otro modo no son representado (por ejemplo, habiendo segmentado los datos, puede ser útil generar un registro para represente al miembro prototípico de cada segmento para un tratamiento futuro).

Actividades Comprobar por técnicas disponibles si es necesario (por ejemplo, mecanismos para construir prototipos para cada segmento de datos segmentados).

Integración de Datos

Tarea **Integrar datos**

Estos son métodos para combinar la información de múltiples tablas u otras fuentes de información para crear nuevos registros o valores.

Salida **Datos combinados**

La combinación de tablas se refiere a la unión de dos o más tablas que tienen diferente información sobre los mismos objetos. En esta etapa, también puede ser aconsejable generar registros nuevos.

También puede ser recomendado para generar valores agregados.

La agregación se refiere a operaciones donde los nuevos valores son calculados por información resumida de múltiples registros y/o tablas.

Actividades Comprobar si las aplicaciones de integración son capaces de integrar las fuentes de entrada como se requiere:

- Integrar fuentes y resultados almacenados.
- Reconsiderar Criterios de Selección de Datos (Vea la Tarea 2.1) en la luz de las experiencias de integración de datos (esto es, usted puede desear incluir/excluir otros conjuntos de datos).

¡Buena idea! Recordar que algún conocimiento puede estar contenido en el formato no-electrónico.

Formateo de Datos

Tarea **Formatear datos**

Transformar formateando se refiere principalmente a modificaciones sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.

Salida **Datos reformateados**

Algunas herramientas tienen requerimientos sobre la orden de los atributos, tal que el primer campo sea un único identificador para cada registro o el campo último de el resultado que el modelo debe predecir.

Actividades **Atributos reorganizados**

Algunas herramientas tienen requerimientos sobre el orden de los atributos, tal que el primer campo sea un único identificador para cada registro o el campo último ser el juego de resultados que el modelo debe predecir.

Reordenando registros

Podría ser importante cambiar el orden de los registros en el conjunto de datos. Quizás el instrumento de modelado requiere que los registros sean clasificados según el valor del atributo de resultado.

Reformateado de valores internos

- Estos son cambios puramente sintácticos hechos para satisfacer las exigencias de la herramienta específica de modelado.
- Reconsiderar Criterios de Selección de Datos en la luz de las experiencias de limpieza de datos (esto es, usted puede desear incluir/excluir otros conjuntos de datos).

3.1.3. Modelado

Selección de la técnica de modelado

Tarea **Seleccionar técnicas de modelado**

Como el primero paso en modelado, seleccionar la técnica de modelado inicial actual. Si múltiples esta para ser aplicados, realizar separadamente esta tarea para cada técnica.

Recuerde que no todos los instrumentos y técnicas son aplicables a toda y cada tarea. Para ciertos problemas, sólo algunas técnicas son apropiadas (Vea el Apéndice 2, donde las técnicas asignan para ciertos tipos de problemas de minería de datos es hablada más detalladamente).

"Requerimientos políticos" y otras restricciones adicionales limitan las opciones disponibles para el ingeniero de minería de datos. Puede ser solo una herramienta o técnica están disponibles para solucionar el problema a mano y que el instrumento no pueda ser absolutamente lo mejor, desde un punto de vista técnico.

Salida **Técnicas de modelado**

Registrar las técnicas de modelado real que se usan.

Actividades Decidir las técnicas apropiadas para el ejercicio, teniendo en cuenta la herramienta seleccionada.

Salida **Presunciones de modelado**

Muchas técnicas de modelado realizan presunciones específicas sobre los datos.

Actividades

- Definir cualquier presunción construida realizada por la técnica sobre los datos (por ejemplo, la calidad, el formato, la distribución).

- Comparar estas presunciones con aquellas del Informe de Descripción de Datos.
- Asegurarse que estas presunciones se cumplen y volver a la Fase de Preparación de Datos, si es necesario.

Generar el diseño de prueba

Tarea **Generar el diseño de prueba**

Antes de construir un modelo, es necesario definir un procedimiento para probar la calidad del modelo y la validez. Por ejemplo, en tareas de minería de datos supervisadas como la clasificación, es común usar tasas de error como medidas de calidad para modelos de minería de datos. Por lo tanto, el diseño de prueba especifica que el conjunto de datos debería ser separado en el entrenamiento y en el conjunto de prueba. El modelo está construido sobre el conjunto de entrenamiento y su calidad estimada sobre el conjunto de prueba.

Salida **Diseño de Prueba**

Describir el plan deliberado para el entrenamiento, las pruebas, y la evaluación de los modelos. Un componente primario del plan es para decidir cómo dividir el conjunto de datos disponible sobre datos que se entrenan, datos de prueba, y conjunto de pruebas de validación.

Actividades Comprobar que existen diseños de prueba separadamente para cada objetivo de minería de datos.

Decidir los pasos necesarios (el número de iteraciones, el número de desviaciones o curvas, etc.).

Preparar los datos requeridos para la prueba.

Construcción del modelo

Tarea **Construir el modelo**

Correr la herramienta de modelado sobre el conjunto de datos preparados para crear uno o más modelos.

Salida **Parámetros de ajuste**

Con cualquier herramienta de modelado, hay a menudo un gran número de parámetros que pueden ser ajustados. Listar los parámetros y sus valores seleccionados, con la explicación (el razonamiento) para la elección.

Actividades Determinar los parámetros iníciales

Documentar las razones para elegir aquellos valores

Salida **Modelos**

Controle la herramienta de modelado en el conjunto de datos listos para crear uno o más modelos.

Actividades Ejecutar la técnica seleccionada sobre el conjunto de datos de entrada para producir el modelo

Post-procesar los resultados de minería de datos (por ejemplo, editar reglas, mostrar árboles)

Salida **Descripción del modelo**

Describir el resultado del modelado y evaluar su exactitud esperada, la robustez, y defectos posibles.

Informar sobre la interpretación de los modelos y encontrar cualquier de las dificultades.

Actividades Describir cualquier característica del modelo actual que puede ser útil para el futuro.

Ajustar parámetro de entorno (de registro) usado para producir el modelo.

Dar una descripción detallada del modelo y cualquier rasgo especial.

Para modelos basados por regla, listar las reglas producidas, más cualquier evaluación de cada regla.

La exactitud y alcance total del modelo.

Para modelos no transparentes, listar cualquier información técnica sobre el modelo (como la topología de las redes neuronales) y cualquier descripción de comportamiento producido por el proceso de modelado (como la exactitud o la sensibilidad).

Describir el comportamiento del modelo y la interpretación.

Expresar conclusiones respecto a los patrones en los datos (si hay alguno); a veces el modelo.

revelar hechos importantes sobre los datos sin un proceso de evaluación separado (por ejemplo, que la salida o la conclusión son duplicadas en una de las entradas).

Evaluación del modelo

Tarea **Evaluar el modelo**

El modelo ahora debería ser evaluado para asegurar que se cumplieron los criterios de éxito de la minería de datos y aprobar los criterios de prueba deseados. Esto es una evaluación puramente técnica basada en el resultado de las tareas modelado.

Salida **Evaluación del modelo**

Resumir los resultados de esta tarea, listar las calidades de los modelos generados (por ejemplo, en términos de exactitud), y el nivel de su calidad en relación a cada otro.

Actividades Evaluar los resultados en lo que concierne a criterios de evaluación.

Probar los resultados según una estrategia de prueba (por ejemplo: Prueba y error, validación cruzada, bootstrapping, etc.)

Comparar los resultados de la evaluación y la interpretación.

Crear la clasificación de resultados en lo que concierne a criterios de éxito y evaluación.

Seleccionar los mejores modelos

Interpretar los resultados en términos de negocio (tanto como sea posible en esta etapa).

Conseguir comentarios de los modelos por expertos en datos o en el dominio.

Verificar la credibilidad del modelo

Comprobar los efectos sobre los objetivos de minería de datos.

Comprobar los modelan contra una base de conocimiento determinada para ver si la información descubierta es nueva y útil.

Comprobar la fiabilidad de los resultados.

Analizar el potencial para el desarrollo de cada resultado.

Si hay una descripción verbal del modelo generado (por ejemplo, en forma de reglas), evaluar las reglas: ¿son lógicas, o son factibles, hay demasiadas reglas o hay muy pocas, violan el sentido común?

Evaluar resultados

Conseguir ideas específicas de cada técnica de modelado y ciertos parámetros de ajustes que conduzcan a resultados buenos/malos.

¡Buena idea! "Tablas de Evaluación" y "Tablas de Beneficio" pueden ser construidas para determinar los bien que el modelo predice.

Salida **Revisión de parámetros de ajuste**

Según la evaluación del modelo, revise parámetros de ajuste y témplelos para la siguiente corrida en la tarea de Construcción del Modelo. Itere (repita) la construcción del modelo y evalúe hasta que usted encuentre el mejor modelo.

Actividades Ajustar parámetros para producir mejores modelos.

3.1.4. Evaluación

Los pasos de evaluación previa tratan con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado que el modelo encuentra los objetivos de negocio, y procura determinar si hay alguna razón de negocio por qué este modelo sea deficiente. Esto compara resultados con los criterios de evaluación definidos en el principio del proyecto.

Un modo bueno de definir las salidas totales de un proyecto de minería de datos es usar la ecuación:

RESULTADOS = MODELOS + CONCLUSIONES

En esta ecuación, definimos que la salida total del proyecto de minería de datos no es solamente los modelos (aunque sean, desde luego, importantes) pero también las conclusiones, las que definimos como algo (aparte del modelo) que es importante en:

La búsqueda de los objetivos de negocio o importante para arribar a nuevas preguntas,

Las líneas de aproximación, o los efectos negativos (por ejemplo, los problemas de calidad de datos descubierto por el uso de la minería de datos).

Nota: Aunque el modelo esté directamente conectado a las preguntas de negocio, las conclusiones no necesariamente están relacionadas con cualquiera de las preguntas u objetivos, mientras ellos son importantes para el promotor del proyecto.

Evaluación de los resultados

Tarea	**Evaluar los resultados** Este paso evalúa el grado al que el modelo encuentra los objetivos de negocio, y procura determinar si hay alguna razón de negocio por el cual este modelo es deficiente. Otra opción es probar el (los) modelo(s) sobre la aplicación de prueba en el sistema verdadero, si permiten las restricciones de tiempo y de presupuesto. Además, la evaluación también evalúa otros resultados generados por la minería de datos. Los resultados de minería de datos cubren los modelos que están relacionados con los objetivos originales de negocio y todas las demás conclusiones. Unos son relacionados con los objetivos de negocios originales mientras que otros podrían revelar desafíos adicionales, información, o ideas para futuras administraciones (direcciones).
Salida	Evaluación de los resultados de minería de datos en lo que respecta a criterios de éxito de negocio. Resumir resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final relacionada a si el proyecto ya encuentra los objetivos iníciales de negocio.

Actividades Comprender los resultados de la minería de datos.

- Interpretar los resultados en términos de la aplicación (del uso).
- Comprobar efectos sobre los objetivos de minería de datos.
- Comprobar los resultados de minería de datos contra la base de un conocimiento determinado para ver si la información descubierta es nueva y útil.
- Evaluar y estimar los resultados en lo que respecta a criterios de éxito de negocio (esto es, el proyecto ha alcanzado los Objetivos de Negocio originales).
- Comparar los resultados de la evaluación y la interpretación.
- Clasificar los resultados en lo que respecta a criterios de éxito de negocio.
- Comprobar el efecto de los resultados sobre el objetivo (fin) de la aplicación inicial.
- Determinar si hay nuevos objetivos de negocio para ser dirigidos más tarde en el proyecto, o en nuevos proyectos.
- Expresar recomendaciones para proyectos futuros de minería de datos.

Salida **Modelos aprobados**

Después de evaluar los modelos con respecto a los criterios de éxito de negocio, seleccionar y aprobar los modelos generados que encontraron los criterios seleccionados.

Proceso de revisión

Tarea **Revisar el proceso**

En este punto, el modelo resultante parece ser satisfactorio y parece satisfacer necesidades de negocio. Es ahora apropiado hacer una revisión más cuidadosa de las promesas de minería de datos para determinar si hay algún factor importante o tarea que de algún modo ha sido pasada por alto. En esta etapa del ejercicio de minería de datos, el Proceso de Revisión toma la forma de una Revisión de Garantía de Calidad.

Salida **Revisión de procesos**

Resumir el proceso de revisión y poner en una lista las actividades que han sido omitidas y/o deberían ser repetidas.

Actividades Proporcionar una descripción del proceso de minería de datos usado.

Analizar el proceso de minería de datos. Para cada etapa del proceso pregunte:

¿Esto fue necesario?

¿Esto fue ejecutado óptimamente?

¿En qué modo podría ser mejorado?

Identificar fracasos.

Identificar pasos desviados (engañosos).

Identificar acciones alternativas posibles y/o caminos inesperados en el proceso.

Revisar resultados de minería de datos en lo que concierne a criterios de éxito de negocio.

Determinación de los próximos pasos

Tarea	**Determinar los próximos pasos**
	Basado en los resultados de evaluación y la revisión de proceso, el equipo de proyecto decide como proceder.
	Las decisiones a ser hechas incluyen si hay que terminar este proyecto y seguir adelante al desarrollo, para iniciar futuras Iteraciones, o establecer nuevos proyectos de minería de datos.
Salida	**Lista de acciones posibles**
	Lista acciones futuras posibles con los motivos para y contra de cada opción.
Actividades	Analizar e potencial para el desarrollo de cada resultado.
	Estimar el potencial para la mejora de proceso actual.
	Comprobar los recursos restantes para determinar si permiten iteraciones de proceso adicionales (o si recursos adicionales pueden estar siendo disponibles).
	Recomendar continuar con las alternativas.
	Refinar el plan de proceso.
Salida	**Decisión**
	Describir las decisiones hechas, con el razonamiento para ello.
Actividades	**Clasificar las acciones posibles**
	Seleccionar una de las acciones posibles.
	Documentar las razones para la elección.

3.1.5. Desarrollo

Plan de desarrollo

Tarea **Desarrollo del Plan**

Esta tarea comienza con la evaluación de los resultados y concluye con una estrategia para el desarrollo de los resultados de la minería de datos en el negocio.

Salida **Plan de Desarrollo**

Resumir la estrategia de desarrollo, incluyendo los pasos necesarios y como realizarlos.

Actividades Resumir resultados desarrollados

- Construir y evaluar los planes alternativos para el desarrollo.
- Decidir para cada resultado de conocimiento o información distinto
- Determinar como el conocimiento o la información serán propagados (generados) a los usuarios.
- Decidir cómo será supervisado el uso del resultado y medido sus beneficios (donde sea aplicable).
- Decidir por cada resultado de modelo desarrollado o de software
- Establecer como el modelo o el resultado de software serán desplegados dentro de los sistemas de la organización.
- Determinar cómo su empleo será supervisado y medido sus beneficios (donde sea aplicable).
- Identificar posibles problemas durante el desarrollo (peligros a ser evitados).

Supervisión y mantenimiento del plan

Tarea **Supervisar y mantener el plan**

La supervisión y el mantenimiento son cuestiones importantes si los resultados de la minería de datos se hacen parte del negocio cotidiano y de su ambiente. Una preparación cuidadosa de una estrategia de mantenimiento ayuda evitar innecesariamente largos períodos de uso incorrecto de los resultados de minería de datos. Para supervisar el desarrollo de los resultados de minería de datos, el proyecto necesita un plan detallado para supervisar y mantener. Este plan tiene en cuenta el tipo específico de desarrollo.

Salida **Plan de supervisión y mantenimiento**

Resumir la estrategia de supervisión y mantenimiento, la inclusión de pasos necesarios y como realizarlos.

Actividades Comprobar aspectos dinámicos (esto es, ¿qué cosas podrían cambiar en el entorno?).

- Decidir cómo será supervisada la precisión.
- Determinar cuando el resultado de minería de datos o el modelo

no deberían ser usados más.

- Identifique criterios (la validez, el límite de la exactitud, nuevos datos, cambios en el dominio de aplicación, etc.), y que debería pasar si el modelo o el resultado no pueden ser más usados.
- (Actualización del modelo, establecimiento de nuevos proyectos de minería de datos, etc.).
- ¿Cambiarán con el tiempo los objetivos de negocio del uso empleo del modelo? Documentar totalmente el problema inicial que el modelo intentaba solucionar.
- Desarrollar el plan de mantenimiento y la supervisión.

Realizar el informe definitivo

Tarea **Producir Informe definitivo**

En el final del proyecto, el equipo de proyecto sobrescribe un informe definitivo. Según el plan de desarrollo, este informe puede ser sólo un resumen del proyecto y su experiencia, o una presentación final de los resultados de minería de datos.

Salida **Informe definitivo**

En el final del proyecto, habrá al menos un informe definitivo en el que todos los hilos son encontrados. Así como la identificación de los resultados obtenidos, el informe también debería describir el proceso, mostrar los costos que se han encontrados, definir cualquier desviación del plan original, describir proyectos de implementación, y hacer cualquier recomendación para el futuro trabajo.

El contenido real detallado del informe depende muchísimo de la audiencia planeada.

Actividades Identificar cuales informes son necesarios (presentación de diapositiva, conclusiones de administración, detalles encontrados, explicación de los modelos, etc.).

Analizar que tan bien se han encontrado los objetivos de minería de datos iníciales.

Identificar grupos de objetivos para el informe.

Describir en forma general las estructuras y el contenido de informe(s).

Seleccionar conclusiones para ser incluidas en los informes.

Escribir un informe.

Salida **Presentación final**

Así como un informe definitivo, puede ser necesario hacer una presentación final para concluir el proyecto- tal vez al patrocinador de dirección, por ejemplo. La presentación normalmente contiene un subconjunto del contenido de la información en el informe definitivo, estructurado de un modo diferente.

Actividades Decidir el grupo objetivo para la presentación final y determinar si ellos ya habrán recibido el informe definitivo.

Seleccionar cuales de los artículos del informe definitivo deberían ser incluidos en la presentación final.

3.1.6. Revisión del proyecto

Tarea **Revisar el proyecto**

Evaluar que fue lo correcto y que fue lo errado, cual fue el éxito obtenido, y que necesidades serán mejoradas.

Salida **Documentación de experiencia**

Resumir la gran experiencia ganada durante el proyecto. Por ejemplo, trampas, accesos a información incorrecta (enfoques erróneos), o los puntos para seleccionar las mejores técnicas de minería de datos en situaciones similares podrían ser la parte de esta documentación. En proyectos ideales, la documentación de experiencia también cubre cualquier informe que ha sido escrito por miembros individuales del proyecto durante el proyecto.

Actividades Entrevistar a toda la gente significativa involucrada en el proyecto y preguntarles sobre su experiencia durante el proyecto.

Si los usuarios finales trabajan en el negocio con los resultados de minería de datos, entrevistarlos: ¿Están satisfechos? ¿Cómo podría haber sido mejor realizado? ¿Necesitan de apoyo adicional?

Resumir la realimentación y escribir la documentación de experiencia.

Analizar el proceso (las cosas que se trabajaron bien, los errores producidos, las lecciones aprendidas, etc.).

Documentar el proceso de minería de datos específico (¿Cómo puede los resultados y la experiencia de aplicación del modelo ser realimentado en el proceso?)

Generalizar desde los detalles para producir la experiencia útil para proyectos futuros.

4. PROYECTO DE MINERÍA DE DATOS

4.1. Comprendiendo el negocio

Objetivos del negocio

Área del problema

Se ha identificado al área de Diseño de Cuestionarios, sin embargo dada la mecánica de generación del cuestionario, esto es, cada especialista en el tema incluye las preguntas que considera más importantes en el cuestionario.

Problema: Se tiene una reducción en los presupuestos asignados para llevar a cabo los levantamientos de información, debido a lo cual es necesario mejorar el diseño del cuestionario.

Los directivos de esta institución, tienen plena conciencia de que es necesario la utilización de otras metodologías para alcanzar la mejora de los cuestionarios, como por ejemplo la minería de datos, en la actualidad está técnica no se ha utilizado en el Instituto, la motivación principal, es la de poner a prueba esta metodología para evaluar si es factible su utilización en el logro de los objetivos del instituto, basados en la generación de información estadística confiable y a un nivel de detalle que permita dar respuesta a las necesidades de información de los usuarios.

El área interesada en el este proyecto de minería de datos es la Dirección Adjunta de investigación y Normatividad de la Dirección General de Estadística.

Dirección Adjunta de investigación, ha expresado que una forma de mejorar el diseño del cuestionario pudiera ser la generación de un modelo de predicción del ingreso, el cual permita la eliminación de la pregunta sobre ingresos del trabajo del cuestionario.

El resultado final esperado es un reporte el cual contenga una propuesta del mejoramiento del cuestionario para el próximo levantamiento del año 2010.

Solución actual

Se han llevado a cabo propuestas de solución, en las cuales utilizando un análisis de datos tradicional se ha generado un modelo a través de una relación lineal con coeficientes variables, el cual permite predecir el ingreso a un nivel agregado. Con resultados satisfactorios para los usuarios.

Como ventaja de esta solución se menciona que es posible producir información estadística con mayor cobertura temática, mayor capacidad de estudio de relaciones entre tremas y una mayor desagregación geográfica, una desventaja mencionada por sus autores, es la utilización de una única muestra de información (censo 2000).

Objetivo principal

Mejorar el diseño del cuestionario, sin disminuir la calidad de la información proporcionada a los usuarios.

Preguntas de los usuarios

¿Es posible mejorar el diseño del cuestionario del censo del 2010 a través de la minería de datos?

¿Cuáles son las posibles soluciones para la mejora del diseño del cuestionario?

Se busca que con la mejora del diseño del cuestionario, se reduzcan los costos de levantamiento, y por otra parte se mantenga la calidad de la información que se ofrece a los usuarios.

Las ventajas esperadas, un levantamiento más rápido de la información, reducción de los costos de levantamiento, procesamiento y validación, ofrecer a los usuarios información con la calidad esperada manteniendo los estándares de calidad INEGI.

Criterios de éxito

Reducción del número de preguntas del cuestionario.

Reducir el tiempo promedio de levantamiento.

Evaluación de la situación

En esta actividad se evaluaron los recursos disponibles, tanto en tecnología, personas y procesos disponibles para la realización de la minería de datos.

Inventario de recursos

Data Warehouse, se cuenta con acceso a los datos del censo de población y vivienda del año 2000.

Personal experto

Dr. Alfredo Bustos y de la Tijera, coordinador del censo de población y vivienda del año 2010.

Juan Martínez Rodríguez, especialista en marcos maestrales y método de regresión de coeficientes variables.

Hardware

Computadora personal con procesador pentium IV a 3.2 Ghz y 512 MB de memoria.

Computadora personal con procesador pentium IV a 3.2 doble núcleo y 1 GB de Memoria.

Software

Statistica Versión 7, con módulo de data mining activado.

SPSS versión 15.0.

Clementine 11.1.1.

Rattle.

Weka.

Fuentes de información

Página WEB del Instituto Nacional de Estadísticas, Geografía e Informática (www.inegi.gob.mx).

Sitio interno del data warehouse estadístico (datawarehouse.inegi.gob.mx).

Patrocinador

Dr. Alfredo Bustos y de la Tijera.

Administrador de bases de datos

Lizette Traconis Lugo.

Experto en Minería de datos:

Dr. Alberto Ochoa

Personal técnico y desarrollador del proyecto

Simón Sánchez Trinidad

Riesgos

Los riesgos de este proyecto son listados a continuación:

Falta de acceso a los datos, debido al esquema de seguridad implantado en el instituto es necesario, solicitar el acceso a las bases de datos, esto debe de ser aprobado por los administradores del data warehouse estadístico, si no se permite el acceso, se requiere la búsqueda de nuevas alternativas de datos, para realizar el proyecto.

Comprensión de los datos, si no se comprenden los datos correctamente, existe la posibilidad de realizar inferencias erróneas sobre los datos.

Calidad de los datos, este es un factor importante ya que en gran medida depende la calidad de los modelos encontrados, si los datos sonde baja calidad, los modelos compartirán esta característica con los datos, este factor también incide en los tamaños de muestra requeridos para realizar el modelado.

Elección errónea de la tarea de minería de datos, se debe de ser muy crítico al momento de elegir la tarea de minería de datos y el método para realizarlo, ya que de ello dependen los resultados que se obtengan.

Tiempo de procesamiento, se debe de tener en cuenta los tiempos de procesamiento y las cuales pueden retrasar el proyecto si no son considerados.

Objetivos de minería de datos

Como se ha mencionado anteriormente, una forma de mejorar el cuestionario del censo de población y vivienda 2010, es la reducción de preguntas del cuestionario, utilizando para ello, la realización de predicciones de estas variables en base a otras variables contenidas en el censo de población y vivienda 2000.

El objetivo principal de la minería de datos, es generar un modelo un modelo de predicción que permita obtener las variables económicas utilizando la información contenida en la base de datos del censo de población y vivienda 2000 contenido en el data warehouse estadístico.

El criterio de éxito planteado es la diferencia entre el valor real y el predicho por el modelo, es decir la exactitud del modelo.

Planeación del proyecto

Para la realización del proyecto de minería de datos se propone el siguiente cronograma de actividades, las fechas propuestas pueden ser modificadas conforme se avance en el proyecto.

Tabla 1: Planeación del proyecto de minería de datos

Tareas a realizar	Comienzo	Finalización
Obtención de los catálogos del censo y muestra	22/02/2008	28/02/2008
Análisis estadístico	29/02/2008	06/03/2008
Incorporación de las observaciones encontradas	07/03/2008	07/03/2008
Análisis de relaciones	10/03/2008	14/03/2008
Incorporación de las observaciones encontradas	17/03/2008	17/03/2008
Análisis de la información con Statistica	18/03/2008	24/03/2008
Incorporación de las observaciones	25/03/2008	25/03/2008
Aplicación de los modelos de regresión (elegidos)	26/03/2008	01/04/2008
Análisis estadístico	02/04/2008	04/04/2008
Incorporación de las observaciones encontradas	07/04/2008	09/04/2008
Elección del modelo predictivo a utilizar	10/04/2008	10/04/2008
Aplicación de la predicción mediante el modelo predictivo seleccionado	11/04/2008	17/04/2008
Análisis de los resultados	18/04/2008	24/04/2008
Conclusiones	25/04/2008	01/05/2008

4.2. Comprensión de los datos

Recolección de los datos iniciales

Se identificó al data warehouse estadístico como el proveedor principal de datos, ya que cualquier proyecto que sea cargado al mismo cumple con una serie de requerimientos y validaciones por parte de los administradores, lo cual asegura una limpieza de los datos correcta, además de una mejora de la calidad de la información, permitiendo la consulta de metadatos sobre los proyectos y sus componentes.

Descripción de los datos

Como paso inicial, se realizó una consulta global sobre tres tablas: TR_VIVIENDA, TR_HOGAR y TR_POBLADOR, realizando un conteo de los registros en cada tabla, estas consultas nos brindan la siguiente información:

Tabla 2: Número de registros totales en las tablas del censo de población y vivienda 2000

Tabla	Número de registros
TR_VIVIENDA	
TR_HOGAR	
TR_POBLADOR	

Como segundo paso, se analizaron las tablas para identificar los tipos de atributos de cada una de las tablas:

Tabla 3: Tipos de variables de los atributos contenidos en las tablas principales del censo de población y vivienda 2000

Tabla	Nombre	Tipo de dato	Tamaño	Tipo de variable
TR_VIVIENDA	ID_VIVIENDA	Entero		Categórica
	ENTIDAD	Entero		Categórica
	MUNICIPIO	Texto	255	Categórica
	LOCALIDAD	Texto	255	Categórica
	AGEB	Texto	255	Categórica
	MANZANA	Texto	255	Categórica
	SEGMENTO	Texto	255	Categórica
	NUMERO_VIVIENDA	Texto	255	Categórica
	APELLIDO	Texto	255	Categórica
	JEFE_ZONA	Texto	255	Categórica

Tabla	Nombre	Tipo de dato	Tamaño	Tipo de variable
	COORDINACION_MUNICIPIO	Texto	255	Categórica
	RESPONSABLE_AGEB	Texto	255	Categórica
	LLAVE_UNICA	Texto	255	Categórica
	TOTAL_HOGARES_VIVIENDA	Texto	255	Continua
	TOTAL_CUESTIONARIO	Texto	255	Continua
	CLASE_VIVIENDA	Texto	255	Categórica
	MATERIAL_PARED	Texto	255	Categórica
	MATERIAL_TECHO	Texto	255	Categórica
	MATERIAL_PISO	Texto	255	Categórica
	TIENE_COCINA	Texto	255	Categórica
	UTILIZA_COCINA_DORMITORIO	Texto	255	Categórica
	CUARTOS_DORMITORIO	Texto	255	Continua
	NUMERO_CUARTOS	Texto	255	Continua
	DISPONE_AGUA_ENTUBADA	Texto	255	Categórica
	DISPONE_SANITARIO	Texto	255	Categórica
	SANITARIO_EXCLUSIVO	Texto	255	Categórica
	AGUA_SANITARIO	Texto	255	Categórica
	DISPONE_DRENAJE	Texto	255	Categórica
	DISPONE_ELECTRICIDAD	Texto	255	Categórica
	TIPO_COMBUSTIBLE	Texto	255	Categórica
	VIVIENDA_PROPIA	Texto	255	Categórica
	TIPO_TENENCIA	Texto	255	Categórica
	DISPONE_RADIO	Texto	255	Categórica
	DISPONE_TELEVISION	Texto	255	Categórica
	DISPONE_VIDEOCASETERA	Texto	255	Categórica
	DISPONE_LICUADORA	Texto	255	Categórica
	DISPONE_REFRIGERADOR	Texto	255	Categórica
	DISPONE_LAVADORA	Texto	255	Categórica

Tabla	Nombre	Tipo de dato	Tamaño	Tipo de variable
	DISPONE_TELEFONO	Texto	255	Categórica
	DISPONE_BOILER	Texto	255	Categórica
	DISPONE_AUTOMOVIL	Texto	255	Categórica
	DISPONE_COMPUTADORA	Texto	255	Categórica
	TOTAL_RESIDENTES_VIVIENDA	Texto	255	Continua
	GASTO_COMUN	Texto	255	Categórica
	TOTAL_HOGARES	Texto	255	Continua
	TIPO_CUESTIONARIO	Texto	255	Categórica
	NUMERO_LOTE	Texto	255	Categórica
	TAMAÑO_LOCALIDAD	Entero		Categórica
	CLASE_VIV_RECLASIFICADA	Texto	255	Categórica
	ZONA	Texto	255	Categórica
	UNIDAD_PRIMARIA_MUESTREO	Texto	255	Categórica
	TIPO_UNIDAD_PRIM_MUESTREO	Entero		Categórica
TR_HOGAR	ID_HOGAR	Entero		Categórica
	ID_VIVIENDA	Entero		Categórica
	ENTIDAD	Entero		Categórica
	MUNICIPIO	Texto	255	Categórica
	LOCALIDAD	Texto	255	Categórica
	AGEB	Texto	255	Categórica
	MANZANA	Texto	255	Categórica
	SEGMENTO	Texto	255	Categórica
	NUMERO_HOGAR	Texto	255	Categórica
	INFORMANTE	Texto	255	Categórica
	TIPO_HOGAR	Texto	255	Categórica
	TOTAL_INGRESO_HOGAR	Texto	255	Continua
	TOTAL_RESIDENTES_HOGAR	Texto	255	Continua
	INGRESO_PERCAPITA_HOGAR	Texto	255	Continua

Tabla	Nombre	Tipo de dato	Tamaño	Tipo de variable
	DECIL_PERCAPITA	Texto	255	Ordinal
	ZONA	Texto	255	Categórica
	UNIDAD_PRIMARIA_MUESTREO	Texto	255	Categórica
	TIPO_UNIDAD_PRIM_MUESTREO	Entero		Categórica
TR_POBLADOR	ID_POBLADOR	Entero		Categórica
	ID_HOGAR	Entero		Categórica
	ID_VIVIENDA	Entero		Categórica
	ENTIDAD	Entero		Categórica
	MUNICIPIO	Texto	255	Categórica
	LOCALIDAD	Texto	255	Categórica
	AGEB	Texto	255	Categórica
	MANZANA	Texto	255	Categórica
	SEGMENTO	Texto	255	Categórica
	NUMERO_PERSONA	Texto	255	Categórica
	PARENTESCO	Texto	255	Categórica
	SEXO	Texto	255	Categórica
	EDAD	Texto	255	Continua
	LUGAR_NACIMIENTO_RESIDE	Texto	255	Categórica
	CONDICION_DERHABIEN_IMSS	Texto	255	Categórica
	CONDICION_DERHABIEN_ISSSTE	Texto	255	Categórica
	CONDICION_DERHABIEN_PEMEX	Texto	255	Categórica
	COND_DERHABIEN_OTRA_INS	Texto	255	Categórica
	NO_TIENE_DERECHOHABIENCIA	Texto	255	Categórica
	CONDICION_DISCAP_MOTRIZ	Texto	255	Categórica
	CONDICION_DISCAP_BRAZOS	Texto	255	Categórica
	CONDICION_DISCAP_AUDITIVA	Texto	255	Categórica
	CONDICION_DISCAP_LENGUAJE	Texto	255	Categórica
	CONDICION_DISCAP_VISUAL	Texto	255	Categórica

Tabla	Nombre	Tipo de dato	Tamaño	Tipo de variable
	CONDICION_DISCAP_MENTAL	Texto	255	Categórica
	DISCAPACIDAD	Texto	255	Categórica
	OTRAS_DISCAPACIDADES	Texto	255	Categórica
	NO_TIENE_DISCAPACIDAD	Texto	255	Categórica
	LUGAR_RESIDENCIA_1995	Texto	255	Categórica
	MUNICIPIO_NACI_RES_1995	Texto	255	Categórica
	HABLA_LENGUA_INDIGENA	Texto	255	Categórica
	TIPO_LENGUAS	Texto	255	Categórica
	HABLA_ESPAÑOL	Texto	255	Categórica
	ALFABETISMO	Texto	255	Categórica
	ASISTENCIA_ESCOLAR	Texto	255	Categórica
	GRADO_APROBADO	Texto	255	Ordinal
	NIVEL_ESCOLARIDAD	Texto	255	Ordinal
	ANTECEDENTE_ESCOLAR	Texto	255	Ordinal
	NIVEL_ACADEMICO	Texto	255	Ordinal
	ESCOLARIDAD_ACUMULADA	Texto	255	Ordinal
	CARRERA	Texto	255	Categórica
	RELIGION	Texto	255	Categórica
	ESTADO_CONYUGAL	Texto	255	Categórica
	CONDICION_ACTIVA_INACTIVA	Texto	255	Categórica
	OCUPACION	Texto	255	Categórica
	POSICION_TRABAJO	Texto	255	Categórica
	TOTAL_HORAS_TRABAJADAS	Texto	255	Continua
	TOTAL_INGRESO_TRABAJO	Texto	255	Continua
	RAMA_ACTIVIDAD	Texto	255	Categórica
	NUMERO_HIJOS_NACIDOS_VIVOS	Texto	255	Continua
	NUMERO_HIJOS_FALLECIDOS	Texto	255	Continua
	NUMERO_HIJOS_SOBREVIVEN	Texto	255	Continua

Tabla	Nombre	Tipo de dato	Tamaño	Tipo de variable
	MES_NACIMIENTO_ULT_HIJO	Texto	255	Categórica
	AÑO_NACIMIENTO_ULT_HIJO	Texto	255	Categórica
	SOBREVIVENCIA_ULTIMO_HIJO	Texto	255	Categórica
	DIA_EDAD_MORIR_ULT_HIJO	Texto	255	Continua
	MES_EDAD_MORIR_ULT_HIJO	Texto	255	Continua
	AÑO_EDAD_MORIR_ULT_HIJO	Texto	255	Continua
	GRUPO_QUINQUENAL	Entero		Ordinal
	ZONA	Texto	255	Categórica
	UNIDAD_PRIMARIA_MUESTREO	Texto	255	Categórica
	TIPO_UNIDAD_PRIM_MUESTREO	Entero		Categórica

Como se puede observar la mayoría de las atributos son del tipo categórico, y solo unas pocas son del tipo ordinal ó continuo.

A continuación se describe cada uno de los atributos y su definición, según los metadatos encontrados en el sitio del data warehouse estadístico:

Tabla 4: Definición de las variables contenidas en las tablas del censo de población y vivienda 2000

Tabla	Nombre	Definición
TR_VIVIENDA	ID_VIVIENDA	Consecutivo que identifica a la vivienda por entidad
	ENTIDAD	Unidad geográfica mayor de la división político-administrativa del país. El territorio nacional se divide en 32 entidades: 31 estados y un Distrito Federal.
	MUNICIPIO	Es la unidad político administrativa en que se particiona cada entidad federativa. Cada municipio posee una clave compuesta por tres dígitos que no se repite dentro de una entidad. En el Distrito Federal las 16 delegaciones son equivalentes a los municipios.

Tabla	Nombre	Definición
	LOCALIDAD	Es todo lugar ocupado por una vivienda o conjunto de viviendas, de las cuales al menos una está habitada; este lugar es reconocido comúnmente por un nombre dado por la ley o la costumbre. Las localidades son de dos tipos: urbanas y rurales.
	AGEB	
	MANZANA	Grupo de viviendas y/o edificios, predios, lotes o terrenos destinados a uso habitacional, comercial, industrial, entre otros. Están delimitadas por calles, andadores o vías peatonales y en las periferias por brechas, veredas, cercas, arroyos, límites de parcelas o predios y otros rasgos que definen su superficie
	SEGMENTO	Identifica la segmentación que se realizó en la manzana
	NUMERO_VIVIENDA	Número que identifica el espacio delimitado por paredes y techos de cualquier material de construcción, en donde viven, duermen, preparan sus alimentos, comen y se protegen de las inclemencias del tiempo una o más personas, con o sin lazos de parentesco. La entrada a la vivienda debe ser independiente, de tal modo que sus ocupantes puedan entrar o salir de ella sin pasar por el interior de otra vivienda. Las viviendas pueden ser colectivas o particulares
	APELLIDO	
	JEFE_ZONA	Código del jefe de zona
	COORDINACION_MUNICIPIO	Código del coordinador municipal

Tabla	Nombre	Definición
	RESPONSABLE_AGEB	Código del responsable de Ageb
	LLAVE_UNICA	Cadena de caracteres, útil para identificar el registro durante las etapas del procesamiento de la información
	TOTAL_HOGARES_VIVIENDA	Total de hogares
	TOTAL_CUESTIONARIO	Total de cuestionarios en la vivienda
	CLASE_VIVIENDA	Diferenciación de la vivienda particular de acuerdo con las características de la infraestructura, independencia y construcción
	MATERIAL_PARED	Material de las paredes o muros
	MATERIAL_TECHO	Material de los techos
	MATERIAL_PISO	Material de los pisos
	TIENE_COCINA	Tiene cocina
	UTILIZA_COCINA_DORMITORIO	Utiliza cocina dormitorio
	CUARTOS_DORMITORIO	Cuartos dormitorio
	NUMERO_CUARTOS	Número de cuartos
	DISPONE_AGUA_ENTUBADA	Instalación de tuberías que se planea y construye para abastecer de agua a las viviendas, edificios y escuelas, entre otros. Puede ser administrada por la entidad, el municipio, la comunidad o una empresa particular. No necesariamente es una instalación subterránea construida con tubos, puede ser superficial sin importar el tipo de material
	DISPONE_SANITARIO	Especifica si la vivienda cuenta o no con servicio sanitario.
	SANITARIO_EXCLUSIVO	Sanitario exclusivo

Tabla	Nombre	Definición
	AGUA_SANITARIO	
	DISPONE_DRENAJE	De acuerdo con la disponibilidad de drenaje, la vivienda se clasifica considerando si dispone de drenaje, bien sea que éste se conecte a una barranca o grieta, una fosa séptica, la red pública, un río o lago e incluso al mar o bien si no dispone de drenaje
	DISPONE_ELECTRICIDAD	Energía eléctrica para alumbrar la vivienda, sin considerar la fuente de donde provenga, la cual puede ser un acumulador, el servicio público de energía, una planta particular, una planta de energía solar, entre otras
	TIPO_COMBUSTIBLE	Material o energía que se usa con mayor frecuencia para calentar y/o cocinar los alimentos en la vivienda
	VIVIENDA_PROPIA	Identifica si la vivienda es propiedad de alguna persona residente en la vivienda
	TIPO_TENENCIA	Tipo tenencia
	DISPONE_RADIO	Se refiere a la disponibilidad de radio con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_TELEVISION	Se refiere a la disponibilidad de televisión con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_VIDEOCASETERA	Se refiere a la disponibilidad de videocasetera con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus

Tabla	Nombre	Definición
		habitantes.
	DISPONE_LICUADORA	Se refiere a la disponibilidad de lavadora con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_REFRIGERADOR	Se refiere a la disponibilidad de refrigerador con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_LAVADORA	Se refiere a la disponibilidad de lavadora con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_TELEFONO	Se refiere a la disponibilidad de teléfono con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_BOILER	Se refiere a la disponibilidad de boiler con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_AUTOMOVIL	Se refiere a la disponibilidad de automóvil o camioneta propios con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	DISPONE_COMPUTADORA	Se refiere a la disponibilidad de computadora con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
	TOTAL_RESIDENTES_VIVIENDA	Número total de residentes en el hogar
	GASTO_COMUN	Gasto común de los habitantes

Tabla	Nombre	Definición
	TOTAL_HOGARES	Total de hogares en la vivienda
	TIPO_CUESTIONARIO	Identifica el tipo de cuestionario con el que se levantó la información
	NUMERO_LOTE	Indica el número de lote
	TAMAÑO_LOCALIDAD	Es la diferenciación de las localidades habitadas a partir de su número de habitantes, en diferentes grupos de tamaños de localidad.
	CLASE_VIV_RECLASIFICADA	Diferenciación de las clases de viviendas particular reclasificadas de acuerdo a las características de la infraestructura, independencia y construcción
	ZONA	Zona
	UNIDAD_PRIMARIA_MUESTREO	Es el área geográfica compuesta por un AGEB, parte de ella o de varias AGEB colindantes con características homogéneas y pertenecientes a un mismo municipio (delegación)
	TIPO_UNIDAD_PRIM_MUESTREO	Identifica el agrupamiento de unidades primarias de muestreo al que pertenece la vivienda
TR_HOGAR	ID_HOGAR	Consecutivo identificador del hogar dentro de una vivienda
	ID_VIVIENDA	Consecutivo que identifica a la vivienda por entidad
	ENTIDAD	Unidad geográfica mayor de la división político-administrativa del país. El territorio nacional se divide en 32 entidades: 31 estados y un Distrito Federal.
	MUNICIPIO	Es la unidad político administrativa en que se

Tabla	Nombre	Definición
		particiona cada entidad federativa. Cada municipio posee una clave compuesta por tres dígitos que no se repite dentro de una entidad. En el Distrito Federal las 16 delegaciones son equivalentes a los municipios.
	LOCALIDAD	Es todo lugar ocupado por una vivienda o conjunto de viviendas, de las cuales al menos una está habitada; este lugar es reconocido comúnmente por un nombre dado por la ley o la costumbre. Las localidades son de dos tipos: urbanas y rurales.
	AGEB	
	MANZANA	Grupo de viviendas y/o edificios, predios, lotes o terrenos destinados a uso habitacional, comercial, industrial, entre otros. Están delimitadas por calles, andadores o vías peatonales y en las periferias por brechas, veredas, cercas, arroyos, límites de parcelas o predios y otros rasgos que definen su superficie
	SEGMENTO	Identifica la segmentación que se realizó en la manzana
	NUMERO_HOGAR	Número que representa la unidad doméstica formada por una o más personas unidas o no por lazos de parentesco, que residen habitualmente en la misma vivienda y se sostienen de un gasto común para la alimentación, es decir, que comparten un mismo gasto para la comida

Tabla	Nombre	Definición
	INFORMANTE	Es la persona que proporciona la información de la unidad de observación, puede ser el responsable u otra persona que sea destinada para ello
	TIPO_HOGAR	Es la clasificación de los hogares según si son familiares y no familiares y al interior de cada uno se distingue según el tipo de relación de parentesco con el jefe del hogar.
	TOTAL_INGRESO_HOGAR	Es el total de los ingresos de las personas que conforman un hogar
	TOTAL_RESIDENTES_HOGAR	Número total de residentes en el hogar
	INGRESO_PERCAPITA_HOGAR	Identifica el ingreso per cápita por hogar calculado
	DECIL_PERCAPITA	Identifica el decil al que pertenece el hogar a nivel nacional de acuerdo a su ingreso percápita
	ZONA	Zona
	UNIDAD_PRIMARIA_MUESTREO	Es el área geográfica compuesta por un AGEB, parte de ella o de varias AGEB colindantes con características homogéneas y pertenecientes a un mismo municipio (delegación)
	TIPO_UNIDAD_PRIM_MUESTREO	Identifica el agrupamiento de unidades primarias de muestreo al que pertenece la vivienda
TR_POBLADOR	ID_POBLADOR	Consecutivo identificador para enumerar a los pobladores dentro de un hogar
	ID_HOGAR	Consecutivo identificador del hogar dentro de una vivienda

Tabla	Nombre	Definición
	ID_VIVIENDA	Consecutivo que identifica a la vivienda por entidad
	ENTIDAD	Unidad geográfica mayor de la división político-administrativa del país. El territorio nacional se divide en 32 entidades: 31 estados y un Distrito Federal.
	MUNICIPIO	Es la unidad político administrativa en que se particiona cada entidad federativa. Cada municipio posee una clave compuesta por tres dígitos que no se repite dentro de una entidad. En el Distrito Federal las 16 delegaciones son equivalentes a los municipios.
	LOCALIDAD	Es todo lugar ocupado por una vivienda o conjunto de viviendas, de las cuales al menos una está habitada; este lugar es reconocido comúnmente por un nombre dado por la ley o la costumbre. Las localidades son de dos tipos: urbanas y rurales.
	AGEB	
	MANZANA	Grupo de viviendas y/o edificios, predios, lotes o terrenos destinados a uso habitacional, comercial, industrial, entre otros. Están delimitadas por calles, andadores o vías peatonales y en las periferias por brechas, veredas, cercas, arroyos, límites de parcelas o predios y otros rasgos que definen su superficie
	SEGMENTO	Identifica la segmentación que se realizó en la manzana
	NUMERO_PERSONA	Número de registro de la persona en el interior del hogar

Tabla	Nombre	Definición
	PARENTESCO	Vínculo existente entre los integrantes del hogar con el jefe del mismo, ya sea por consanguinidad, matrimonio, adopción, afinidad o costumbre
	SEXO	Condición biológica que distingue a las personas en hombres y mujeres
	EDAD	Número de años cumplidos por la persona, desde la fecha de su Nacimiento hasta el momento del hecho
	LUGAR_NACIMIENTO_RESIDE	Entidad federativa o país donde nació y/o radica el poblador al momento del levantamiento del hecho
	CONDICION_DERHABIEN_IMSS	Identifica si la persona es derechohabiente al IMSS
	CONDICION_DERHABIEN_ISSSTE	Identifica si la persona es derechohabiente al ISSSTE
	CONDICION_DERHABIEN_PEMEX	Identifica si la persona es derechohabiente a PEMEX, Defensa o Marina
	COND_DERHABIEN_OTRA_INS	Identifica si la persona tiene derechohabiencia a otra institución diferente de IMSS, ISSSTE, PEMEX, Defensa o Marina, Seguro Popular o seguro por Institución Privada
	NO_TIENE_DERECHOHABIENCIA	No tiene derecho la persona a recibir atención médica en instituciones de salud públicas
	CONDICION_DISCAP_MOTRIZ	Disfunción en el aparato psicomotor que limita al individuo para

Tabla	Nombre	Definición
		realizar actividades físicas
	CONDICION_DISCAP_BRAZOS	Identifica si la persona presenta o no discapacidad para mover los brazos
	CONDICION_DISCAP_AUDITIVA	Identifica si la persona presenta o no discapacidad al oír
	CONDICION_DISCAP_LENGUAJE	Identifica si la persona presenta o no discapacidad para hablar
	CONDICION_DISCAP_VISUAL	Identifica si la persona presenta o no discapacidad al ver
	CONDICION_DISCAP_MENTAL	Identifica si la persona presenta o no discapacidad mental
	DISCAPACIDAD	Discapacidad
	OTRAS_DISCAPACIDADES	Identifica si la persona tiene dos limitaciones y una no se pudo clasificar
	NO_TIENE_DISCAPACIDAD	Identifica si la persona no presenta alguna discapacidad
	LUGAR_RESIDENCIA_1995	Entidad federativa donde la persona tenía su residencia habitual en el año 1995
	MUNICIPIO_NACI_RES_1995	Especifica el municipio o delegación de residencia en 1995 de las personas que no son migrantes estatales.
	HABLA_LENGUA_INDIGENA	Especifica si el poblador habla o no alguna lengua indígena
	TIPO_LENGUAS	Conjunto de idiomas que históricamente son herencia de diversas etnias del continente americano y que se hablan en México.
	HABLA_ESPAÑOL	Situación que distingue a la población de 5 años y más que habla alguna lengua indígena

Tabla	Nombre	Definición
		respecto a si habla o no la lengua española
	ALFABETISMO	
	ASISTENCIA_ESCOLAR	Situación que distingue a la población de 5 años y más, según su asistencia pasada o actual a cualquier establecimiento de enseñanza del Sistema Educativo Nacional como preescolar, primaria, secundaria, preparatoria, profesional o postgrado independientemente de su modalidad, ya sea pública o privada, escolarizada, abierta, de estudios técnicos o comerciales, educación especial o de educación para adultos
	GRADO_APROBADO	Indica la descripción del grado aprobado según el nivel de estudios de la persona
	NIVEL_ESCOLARIDAD	Especifica hasta qué grado de escolaridad aprobó la persona
	ANTECEDENTE_ESCOLAR	Tipo de antecedente escolar necesario para cursar su carrera
	NIVEL_ACADEMICO	Identifica el nivel académico de la persona, así como el tipo de antecedente escolar necesario para cursar su carrera. (Fusión de nivel académico y antecedente escolar).
	ESCOLARIDAD_ACUMULADA	Número de años que ha aprobado la persona desde que entro a la escuela y hasta el momento del hecho

Tabla	Nombre	Definición
	CARRERA	Tiene como antecedente inmediato la preparatoria o el bachillerato; abarca licenciatura, normal, carreras técnicas y postgrado
	RELIGION	Creencia o preferencia espiritual que declare la población, sin tener en cuenta si está representada o no por un grupo organizado.
	ESTADO_CONYUGAL	Condición de cada individuo de acuerdo a las leyes o costumbres conyugales o matrimoniales del país, las cuales son: soltero, casado, divorciado, en unión libre, separado y viudo
	CONDICION_ACTIVA_INACTIVA	Situación que distingue a la población de 12 años y más, según haya realizado o no alguna actividad económica en la semana de referencia. Se clasifica en población económicamente activa y población económicamente inactiva
	OCUPACION	Tipo de trabajo, oficio o tarea específica que desarrolló la persona ocupada en su trabajo principal.
	POSICION_TRABAJO	Relación que estableció la población ocupada con su empleo o lugar de trabajo en la semana de referencia. Su clasificación incluye: empleados u obreros, jornaleros o peones, patrones; trabajadores por su cuenta y trabajadores familiares sin pago

Tabla	Nombre	Definición
	TOTAL_HORAS_TRABAJADAS	Total de horas trabajadas
	TOTAL_INGRESO_TRABAJO	Total de ingresos por trabajo (mensual)
	RAMA_ACTIVIDAD	Rama de actividad
	NUMERO_HIJOS_NACIDOS_VIVOS	Número de hijos nacidos vivos
	NUMERO_HIJOS_FALLECIDOS	Número de hijos fallecidos
	NUMERO_HIJOS_SOBREVIVEN	Es todo producto de la concepción (embarazo) que nació vivo y al momento de la entrevista aún vive, independientemente del lugar donde resida
	MES_NACIMIENTO_ULT_HIJO	Mes de nacimiento del último hijo nacido vivo de la mujer de 12 años o más de edad
	AÑO_NACIMIENTO_ULT_HIJO	Año de nacimiento del último hijo nacido vivo de la mujer de 12 años o más de edad
	SOBREVIVENCIA_ULTIMO_HIJO	Identifica la condición de sobrevivencia del último hijo nacido vivo de la mujer de 12 años o más de edad, al momento del hecho
	DIA_EDAD_MORIR_ULT_HIJO	Identifica la edad en días que tenía el último hijo nacido vivo al momento de morir
	MES_EDAD_MORIR_ULT_HIJO	Identifica la edad en meses que tenía el último hijo nacido vivo al momento de morir
	AÑO_EDAD_MORIR_ULT_HIJO	Identifica la edad en años que tenía el último hijo nacido vivo al momento de morir
	GRUPO_QUINQUENAL	Clasificación según los grupos quinquenales de edad donde cae la edad de la persona
	ZONA	Identifica el dominio de estudio, definidos para el marco, al que pertenece la vivienda

Tabla	Nombre	Definición
	UNIDAD_PRIMARIA_MUESTREO	Es el área geográfica compuesta por un AGEB, parte de ella o de varias AGEB colindantes con características homogéneas y pertenecientes a un mismo municipio (delegación)
	TIPO_UNIDAD_PRIM_MUESTREO	Identifica el agrupamiento de unidades primarias de muestreo al que pertenece la vivienda

4.1.5.3. **Exploración de datos**

Al realizar el análisis del contenido de los campos, se obtuvieron las siguientes observaciones:

El 77.29% de las viviendas son de tabique, ladrillo, block, piedra, cemento o concreto, el restante 22.64 son de otro tipo de material.

Un 62.57% de las viviendas tienen techo de losa o concreto, el restante 37.43% están realizados con otro tipo de material.

Un 54.05% de las viviendas a nivel nacional tienen piso de cemento o firme, un 30.41% tiene piso recubierto de madera, mosaico u otros recubrimientos, el restante 15.54% no tiene recubrimiento o no saben si lo tienen.

Un 89.88% de las viviendas tiene un cuarto el cual utilizan como cocina, y un 10% de estos lo utilizan también como dormitorio.

En cuestión de infraestructura de las viviendas esto es lo observado:

Un 82.62% de las viviendas tienen agua entubada en el predio, el resto de las viviendas no cuentan con este servicio y obtienen el agua por otros medios.

Un 87.87% de las viviendas cuentan con sanitario u otro medio séptico, de estos el 84.15% son de uso exclusivo de la vivienda, y el 76.5% del total de las viviendas cuentan con servicio de drenaje.

Un 93.12% de las viviendas tienen energía eléctrica y un 2.8% la utilizan para cocinar.

El combustible más utilizado por las viviendas de México es el gas con un 79.97%.

Un 76.69% de las viviendas son propiedad de algún miembro del hogar que reside en el.

Para las personas de doce y más años, se observa lo siguiente:

El 44.5 % presenta el estado conyugal "casado", un 37.07% se encontraba soltero, un 10.25% vive en unión libre y el restante 8.17% presenta un estado conyugal diferente a los listados anteriormente.

Un 47.55% de la población de doce años y más indico que trabajo la semana anterior al levantamiento,

De la población ocupada de 12 años y más el 49.4% presenta el estado conyugal "casado"

Un 68.40% de las personas ocupadas de 12 años y más, son trabajadores subordinados con pago, mientras que un 2.53% son empleadores o patrones, resalta el número de trabajadores por su cuenta, con un 21.86%, el restante 4.06% son trabajadores sin pago.

El 66.29% de las personas de doce años y más ocupadas trabajan entre 31 y 60 horas semanales.

El promedio nacional de ingreso por producto del trabajo de las personas de 12 años y más es de 3553.19 pesos mensuales, con una desviación estándar de 135,277.55, lo cual nos indica una dispersión pronunciada de los ingresos a nivel nacional.

Debido a que el número de registros de pobladores, y las características de la variable ingreso por trabajo, resulta muy costoso en tiempo y recursos el análisis de todos los registros en las bases de datos, por lo cual se procedió a obtener una muestra, utilizando para ello, una consulta al servidor de bases de datos del data warehouse mediante la utilización de la instrucción SAMPLE, la cual no permite obtener una muestra aleatoria simple.

Se obtuvo una muestra del 1% del total de la población, lo cual representa un total de 973,684 personas, analizando cada uno de los campos, contabilizando el porcentaje de campos en blanco, enfocándose en las variables de horas trabajadas e ingresos del trabajo, la primer variable en esta muestra tiene un 65.29% de registros en blanco, mientras que los ingresos por producto del trabajo tienen un 65.26% con este tipo de registros.

Se analizó la estructura de edad por horas trabajadas, con lo cual se identifico que solo personas de doce años y más tienen los campos TOTAL_HORAS_TRABAJADAS Y TOTAL_INGRESO_TRABAJO tenían esta información.

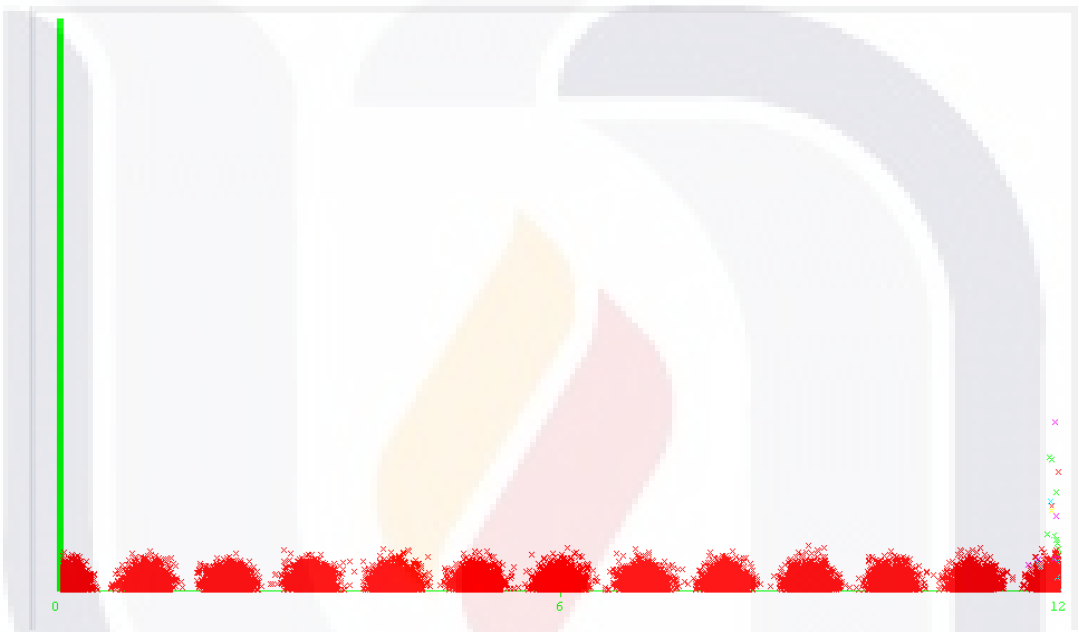


Figura 4: Visualización de los ingresos de productos del trabajo para personas con edades menores o iguales a 12 años

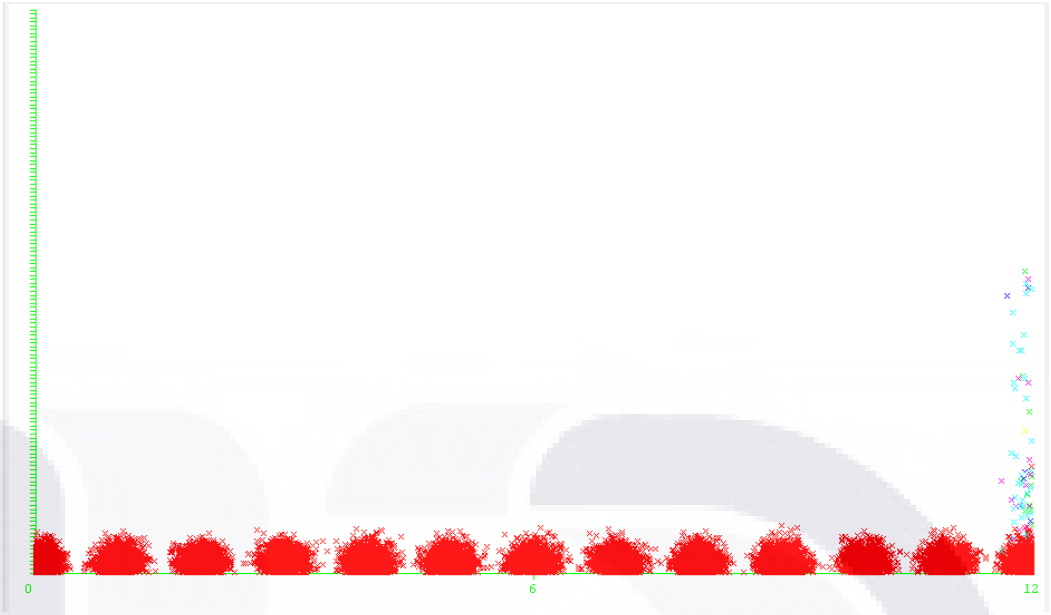


Figura 5: Visualización de las horas trabajadas para edades menores o iguales de 12 años.

Este comportamiento se debe a que en el cuestionario se levanto la información solo para los mayores de 12 años, debido a lo cual es necesario reducir el universo de análisis a las personas de 12 años y más ocupadas (personas que trabajaron en la semana de referencia).

Para la reducción del universo se integraron dos condiciones a la consulta de datos:

EDAD >=12

CONDICION_ACTIVA_INACTIVA >= 10 y <= 20

Como un segundo análisis de la información se generaron las tablas de correlación de las variables continuas, de los cuales se obtuvieron los siguientes resultados:

La variable objetivo TOTAL_INGRESO_TRABAJO, presenta las siguientes correlaciones con respecto otras variables continuas.

Tabla 5: Correlaciones de la variable de ingreso por productos del trabajo con respecto a otras variables continúas

(Muestra_Inicial) Las correlaciones marcadas son significantes a p < .05000 incluye la condición: edad>=12 y (condicion_activa_inactiva >=10 and condicion_activa_inactiva <=20)	
	TOTAL_INGRESO_TRABAJO117
TOTAL_HOGARES_VIVIENDA	-0.01
TOTAL_CUESTIONARIO	0.04
TOTAL_RESIDENTES_VIVIENDA	0.04
TOTAL_INGRESO_HOGAR	0.37
TOTAL_RESIDENTES_HOGAR	0.04
INGRESO_PERCAPITA_HOGAR	0.14
DECIL_PERCAPITA	-0.07
EDAD	-0.01
GRADO_APROBADO	0.01
NIVEL_ESCOLARIDAD	0.04
NIVEL_ACADEMICO	0.04
ESCOLARIDAD_ACUMULADA	0.04
NUMERO_HIJOS_NACIDOS_VIVOS	-0.03
NUMERO_HIJOS_SOBREVIVEN	-0.01
MES_NACIMIENTO_ULT_HIJO	-0.01
AÑO_NACIMIENTO_ULT_HIJO	-0.01
DIA_EDAD_MORIR_ULT_HIJO	-0.04
MES_EDAD_MORIR_ULT_HIJO	-0.08
AÑO_EDAD_MORIR_ULT_HIJO	-0.04

Las variables INGRESO_PERCAPITA_HOGAR y TOTAL_INGRESO_HOGAR muestran una correlación de 0.14 y 0.37 con la variable TOTAL_INGRESO_TRABAJO, lo cual es explicado por la relación que tienen estas variables con el total del ingreso por trabajo, ya que la primera es el resultado de la suma de todos los ingresos del hogar dividido entre el número de residentes del hogar y las segunda es la suma de todos los ingresos de los residentes del hogar.

Tabla 6: Correlaciones de la variable total_residentes_hogar con respecto a otras variables continúas

(Muestra_Inicial) Las correlaciones marcadas son significantes a p < .05000 incluye la condición: edad>=12 y (condicion_activa_inactiva >=10 and condicion_activa_inactiva <=20)	
	TOTAL_RESIDENTES_VIVIENDA
TOTAL_HOGARES_VIVIENDA	0.27
TOTAL_CUESTIONARIO	0.93
CUARTOS_DORMITORIO	0.12
NUMERO_CUARTOS	0.06
TOTAL_HOGARES	0.30
TOTAL_INGRESO_HOGAR	0.01
TOTAL_RESIDENTES_HOGAR	0.99
DECIL_PERCAPITA	-0.06
NIVEL_ESCOLARIDAD	-0.03
NIVEL_ACADEMICO	-0.03
ESCOLARIDAD_ACUMULADA	-0.02
TOTAL_HORAS_TRABAJADAS	0.02
TOTAL_INGRESO_TRABAJO	0.04
NUMERO_HIJOS_NACIDOS_VIVOS	0.07
NUMERO_HIJOS_FALLECIDOS	0.01
NUMERO_HIJOS_SOBREVIVEN	0.04
AÑO_NACIMIENTO_ULT_HIJO	0.01
DIA_EDAD_MORIR_ULT_HIJO	0.02
MES_EDAD_MORIR_ULT_HIJO	-0.06
AÑO_EDAD_MORIR_ULT_HIJO	0.08

La variable TOTAL_RESIDENTES_HOGAR tiene una alta correlación (0.99) con la variable TOTAL_RESIDENTES_VIVIENDA, lo cual es explicado por el 97.26 por ciento de las viviendas cuentan con un solo hogar, asimismo, se observa una correlación alta con la variable TOTAL_CUESTIONARIO, dado que al aumentar el número de residentes de un hogar se incrementa el número de las personas a quienes se les aplico el cuestionario.

Tabla 7: Tabla de correlaciones de la variable nivel de escolaridad con las variables continúas de la base de datos

Muestra_Inicial) Las correlaciones marcadas son significantes a p < .05000 incluye la condición: edad>=12 y (condicion_activa_inactiva >=10 and condicion_activa_inactiva <=20)

	NIVEL_ESCOLARIDAD
TOTAL_HOGARES_VIVIENDA	-0.03
TOTAL_CUESTIONARIO	-0.03
NUMERO_CUARTOS	0.03
TOTAL_RESIDENTES_VIVIENDA	-0.03
TOTAL_HOGARES	-0.02
TOTAL_INGRESO_HOGAR	0.02
TOTAL_RESIDENTES_HOGAR	-0.02
DECIL_PERCAPITA	-0.03
EDAD	0.06
GRADO_APROBADO	0.44
ANTECEDENTE_ESCOLAR	0.82
NIVEL_ACADEMICO	0.98
ESCOLARIDAD_ACUMULADA	0.39
TOTAL_HORAS_TRABAJADAS	-0.01
TOTAL_INGRESO_TRABAJO	0.04
NUMERO_HIJOS_FALLECIDOS	0.05
NUMERO_HIJOS_SOBREVIVEN	0.04
MES_NACIMIENTO_ULT_HIJO	0.02
AÑO_NACIMIENTO_ULT_HIJO	0.04
DIA_EDAD_MORIR_ULT_HIJO	-0.05
MES_EDAD_MORIR_ULT_HIJO	-0.02
AÑO_EDAD_MORIR_ULT_HIJO	-0.01

La variable NIVEL_ESCOLARIDAD, presenta altas correlaciones con el grado aprobado (0.44), antecedente escolar (0.82) y nivel académico (0.98), esta última casi una correlación perfecta, mientras que la relación con la variable escolaridad acumulada alcanza una valor de 0.39, el antecedente escolar mantiene un valor de 0.82 con la variable de nivel de escolaridad.

Tabla 8: Correlaciones de la variable grupo_quinquenal con otras variables de la base de datos.

Correlations (Muestra_Inicial) Marked correlations are significant at p < .05000 Include condition: edad82>=12 and (condicion_activa_inactiva113 >=10 and condicion_activa_inactiva113 <=20)

	GRUPO_QUINQUENAL
TOTAL_HOGARES_VIVIENDA	-0.02
TOTAL_CUESTIONARIO	-0.02
CUARTOS_DORMITORIO	-0.01
NUMERO_CUARTO	0.02
TOTAL_RESIDENTES_VIVIENDA	-0.02
TOTAL_HOGARES	-0.04
TOTAL_INGRESO_HOGAR	-0.05
TOTAL_RESIDENTES_HOGAR	-0.02
INGRESO_PERCAPITA_HOGAR	-0.04
DECIL_PERCAPITA	0.05
EDAD	0.54
GRADO_APROBADO	0.19
NIVEL_ESCOLARIDAD	0.17
ANTECEDENTE_ESCOLAR	-0.05
NIVEL_ACADEMICO	0.15
ESCOLARIDAD_ACUMULADA	0.05
TOTAL_HORAS_TRABAJADAS	0.02
NUMERO_HIJOS_NACIDOS_VIVOS	0.25
NUMERO_HIJOS_FALLECIDOS	0.20
NUMERO_HIJOS_SOBREVIVEN	0.25
MES_NACIMIENTO_ULT_HIJO	0.04
AÑO_NACIMIENTO_ULT_HIJO	0.37
DIA_EDAD_MORIR_ULT_HIJO	-0.09
MES_EDAD_MORIR_ULT_HIJO	-0.09
AÑO_EDAD_MORIR_ULT_HIJO	0.26

La variable relacionada con el agrupamiento de edades (GRUPO_QUINQUENAL) y la variable edad, muestran una ligera correlación con la variable del nivel de escolaridad con un valor de 0.54 y 0.17 respectivamente.

Verificación de la calidad de los datos

Al realizar una exploración a los datos, se identifica que en las variables categóricas se utiliza el número 9 como no especificado, este se repite según el número de caracteres a utilizar, por ejemplo, en el campo edad el valor para no especificado es 999 y en el caso de la variable de ingreso por producto del trabajo se identifica con el 999999.

Por lo general los valores permitidos en las variables categóricas es un número del 1 al 9 inclusive, en los campos a los que se hace referencia la tenencia de bienes en el hogar se utiliza los números del 1 al 8 para identificar la tenencia, utilizando los nones para el SI y los pares para el NO, se repite la secuencia de números a partir de la quinta opción (En este hogar tienen refrigerador)

Como resultado de la estructura del cuestionario las preguntas se realizaron segmentadas de acuerdo a condiciones sobre edad, ocupación y sexo, así se identifican las siguientes secciones:

- Características de la vivienda
- Características de la persona en general.
- Personas de 5 años y Más
- Personas de 12 años y más
- Mujeres mayores de 12 años

Las pregunta de derechohabiencia esta segmentada en las diferentes opciones, donde los valores de si están representados por el valor definido en el cuestionario.

1) IMSS
2) ISSSTE
3) PEMEX
4) Otra institución
5) No tiene servicio médico

Esta situación también se encuentra en el tipo de discapacidad, en el cual se segmento la variable en sus diversas categorías.

En el caso de la variable "Antecedente escolar", la gran cantidad de nulos en la base de datos se debe a que por estructura del cuestionario solo se pregunto por el antecedente escolar a las personas que cursaron la Normal, una carrera técnica o comercial o cursaron el nivel superior de educación, tal y como se muestra en la figura 6.

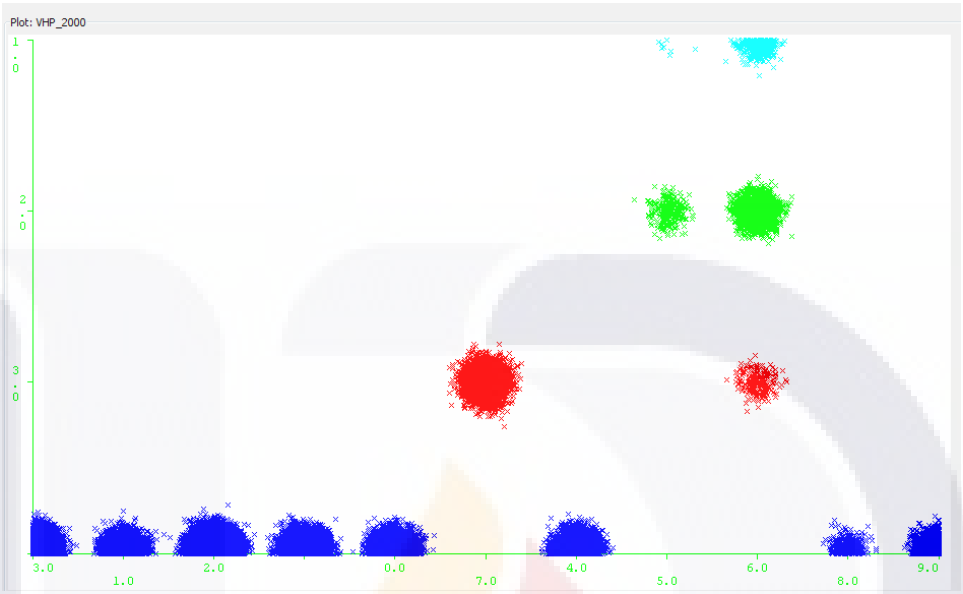


Figura 6: Visualización del antecedente escolar según nivel de escolaridad

La variable TIPO_UNIDAD_PRIM_MUESTREO tiene un 100% de nulos, debido a esto no aporta ningún valor al análisis de la minería de datos, otra variable que presenta baja calidad de datos es el campo DECIL_PERCAPITA, el cual tiene un 90.3% de valores nulos, el restante 9.7% solo tiene valores de 1 ó 99.

La variable TOTAL_HOGARES, tiene un porcentaje de 97.2% de valores nulos, los cuales están relacionados con hogares únicos.

Preparación y Selección de datos

Para continuar con la exploración de los datos fue necesario calcular el tamaño de la muestra, tomando en cuenta la media de ingreso desde los datos poblacionales, contrastando la media poblacional con una media cercana, esto con el fin de evitar introducir sesgo al modelo.

Para calcular el tamaño de la muestra se utilizo el software Statistica a través del cálculo de muestra con los siguientes parámetros:

1 Sample t-Test: Sample Size Calculation

H0: Mu = Mu0

Type I Error Rate (Alpha): 0.05

Power Goal: 0.95

Null Hypothesized Mean (Mu0): 3553.07

True Population Mean (Mu): 3553.19

Population S.D. (Sigma): 10

Standardized Effect (Es): 0.019

Con lo cual se obtuvo como resultado que el tamaño de la muestra requerida es de **72,971** registros, lo cual representa un 0.22 por ciento de la población ocupada de 12 años con un ingreso declarado.

Una vez obtenida la muestra, se procedió a una revisión inicial de los campos, para identificar a aquellos de ser susceptibles de análisis a través de la minería de datos, de este análisis, se descartaron del análisis algunas variables que identifican a los registros.

Limpieza de Datos

De acuerdo a lo analizado anteriormente, se eliminaron del procesamiento aquellas variables con altas correlaciones con otras, del grupo de variables se elimino del procesamiento las siguientes:

Ingresos:

Utilizar

TOTAL_INGRESO_TRABAJO

Descartar

TOTAL_INGRESO_HOGAR

TOTAL_RESIDENTES_HOGAR

INGRESO_PERCAPITA_HOGAR

Escolaridad

Utilizar

NIVEL_ESCOLARIDAD

Descartar

GRADO_APROBADO

ANTECEDENTE_ESCOLAR

NIVEL_ACADEMICO

ESCOLARIDAD_ACUMULADA

Residentes

Utilizar

TOTAL_RESIDENTES_HOGAR

Descartar

TOTAL_CUESTIONARIO

TOTAL_RESIDENTES_VIVIENDA

Edad

Utilizar

EDAD

Descartar

GRUPO_QUINQUENAL

En la búsqueda de una mejor predicción se decidió la eliminación de registros en los cuales no se contara con el valor del ingreso ó tuvieran un ingreso no especificado y no estuvieran ocupados en la semana de referencia.

Formateo de Datos

Para mantener una coherencia entre los tipos de variables detectados en el cuestionario, al cargar los datos en el software STATISTICA, se procedió a cambia r los tipos de variable según los tipos detectados en el punto comprensión de los datos.

4.3. Modelado

Selección de la técnica de modelado

En base al conocimiento de los datos hasta el momento y considerando que la tarea de minería de datos que se desea realizar es la predicción de valores de ingreso en base a algunos atributos independientes, se selecciono la técnica de redes neuronales artificiales como la técnica a utilizar para el desarrollo del modelo de predicción.

Generar el diseño de prueba

Para probar la calidad del modelo se utilizarán las siguientes medidas estadísticas:

Media, Desviación Estándar, porcentaje de diferencia entre lo real y lo observado.

Construcción del modelo

Con el software "STATISTICA", se utilizo el módulo de minería de datos, en la opción de selección de atributos, para identificar los mejores atributos predictores del atributo TOTAL_INGRESO_TRABAJO, al aplicar este procedimiento a los registros contenidos en la muestra para realizar una selección de atributos interesantes para la realización de minería de datos para la tarea de predicción del TOTAL_INGRESO_TRABAJO, de esta tarea, se obtuvieron los resultados mostrados en la tabla 9.

Tabla 9: Mejores predictores para la variable TOTAL_INGRESO_TRABAJO

Atributo	F-Value	P-Value
Ingreso_percapita_Hogar	2875.99	0.000000
Posicion_Trabajo	709.68	0.000000
Sexo	281.42	0.000000
Total_residentes_vivienda	217.34	0.000000
Nivel_Escolaridad	213.66	0.000000
Total_residentes_hogar	210.84	0.000000
Dispone_Refigerador	167.13	0.000000
Nivel_academico	159.51	0.000000
Dispone_Videocasetera	150.38	0.000000

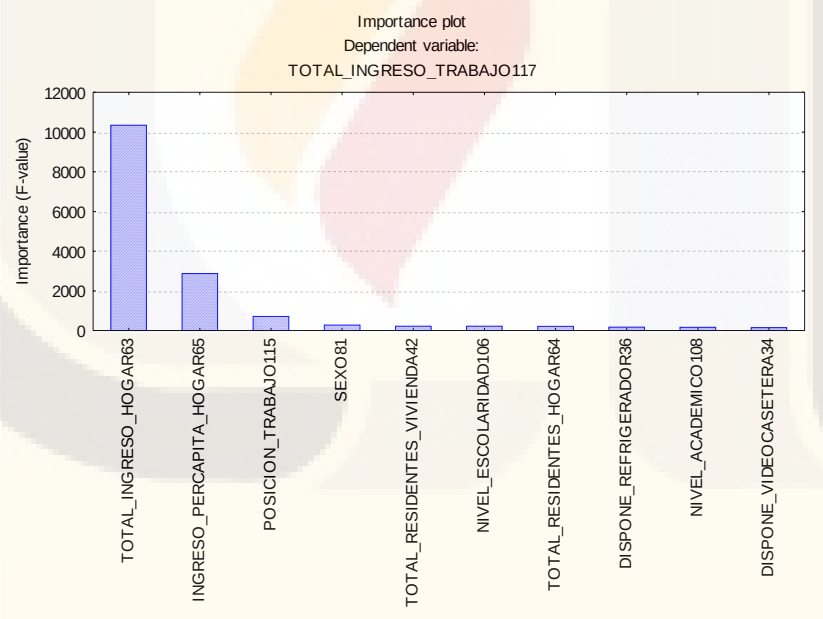


Figura 7: Importancia de los factores utilizados para la creación del modelo

Como se puede observar el atributo Ingreso_percapita_hogar, está ligado directamente a el TOTAL_INGRESO_TRABAJO de las personas, ya que este atributo es resultado de la suma de todos los ingresos por productos del trabajo dividido entre el número total de personas en el hogar.

Como se observo anteriormente el número de viviendas con hogares únicos es alrededor del 96%, por lo cual se decidió la utilización del atributo total_residentes_vivienda como parte de La predicción,

Como se identifico con anterioridad, el atributo nivel_academico es una fusión de los atributos de escolaridad y antecedente escolar necesario para obtener el nivel académico de la persona, debido a que presenta un valor F menor que el atributo Escolaridad, fue eliminado de la tarea de minería de datos.

De esta forma los atributos para la realización de la predicción del ingreso por producto del trabajo de una persona son:

Tabla 10: mejores predictores de la variable TOTAL_INGRESO_TRABAJO después de la eliminación de variables muy correlacionadas

Atributo	F-Value	P-Value
Posicion_Trabajo	709.68	0.000000
Sexo	281.42	0.000000
Total_residentes_vivienda	217.34	0.000000
Nivel_Escolaridad	213.66	0.000000
Dispone_Refigerador	167.13	0.000000
Dispone_Videocasetera	150.38	0.000000

Una vez elegidos los atributos que se usaran para realizar la tarea de la minería de datos, a la muestra obtenida desde el data warehouse se le agrego un campo con el propósito de identificar tres grupos de registros:

- Entrenamiento
- Selección
- Prueba

Estas asignaciones fueron realizadas de forma aleatoria a través de la función aleatoria de un número en este caso 3, de esta forma se asignaron los valores de 1ó 2 ó 3 a cada registro, se decidió que el grupo de entrenamiento sería aquel que tuviera más elementos en este caso el grupo con el valor 2.

Con la ayuda del software STATISITCA, se procedió a generar posibles redes neuronales que permitiese la predicción de los ingresos del trabajo, en base a los atributos seleccionados.

Las redes neuronales obtenidas del proceso anterior son las siguientes:

	Desempeño			Errores			Red Neuronal		
Perfil	Entrena	Selec.	Prueba	Entrena	Selec.	Prueba	Entrenamiento/Miembros	Entradas	Ocultas
1) MLP 1:1-4-1:1	1.000	1.000	1.000	0.110	0.109	0.108	BP100,CG20,CG0b	1	4
2) MLP 1:3-6-1:1	1.000	1.000	1.000	0.109	0.109	0.108	BP100,CG20,CG0b	1	6
3) Linear 6:15-1:1	0.992	0.994	0.994	0.109	0.109	0.107	PI	6	0
4) RBF 7:16-354-1:1	0.982	1.000	1.001	0.008	0.008	0.008	SS,KN,PI	7	354
RBF 7:16-177-1:1	0.986	0.996	0.997	0.008	0.008	0.008	SS,KN,PI	7	177

Tabla 11:Redes neuronales generadas a partir de los mejores predictores de la variable TOTAL_INGRESO_TRABAJO

Al aplicar los modelos obtenidos de estas redes neuronales a un conjunto de datos mayor, fueron descartadas debido a que el porcentaje de diferencia contra lo observado oscila entre -62.90 y 327.10.

Debido a los resultados obtenidos se selecciono un número mayor de variables independientes, en total 20 atributos, los cuales se muestran a continuación:

Tabla 12: Selección de los nuevos mejores predictores de la variable TOTAL_INGRESO_TRABAJO

Atributo	**F-value**	**p-value**
POSICION_TRABAJO	179.432	0.000000
SEXO	160.656	0.000000
TOTAL_RESIDENTES_HOGAR	52.067	0.000000
NIVEL_ESCOLARIDAD	47.155	0.000000
DISPONE_REFRIGERADOR	36.368	0.000000
DISPONE_VIDEOCASETERA	33.128	0.000000
DISPONE_COMPUTADORA	29.687	0.000000
AGUA_SANITARIO	29.085	0.000000
VIVIENDA_PROPIA	28.857	0.000000
DISPONE_TELEFONO	27.316	0.000000
DISPONE_BOILER	23.880	0.000000
DISPONE_LAVADORA	21.615	0.000000
RAMA_ACTIVIDAD	19.869	0.000000
MATERIAL_PISO	17.948	0.000000
ASISTENCIA_ESCOLAR	16.929	0.000000
DISPONE_AUTOMOVIL	15.136	0.000000

Atributo	F-value	p-value
OCUPACION	7.492	0.000000
TOTAL_HORAS_TRABAJADAS	5.666	0.000007
ESTADO_CONYUGAL	5.263	0.000001
PARENTESCO	3.540	0.000000

Al igual que el caso anterior se utilizaron tres grupos de datos para realizar el entrenamiento de las redes neuronales, utilizando para ello el software STATISTICA, de la realización del entrenamiento con estos atributos se obtuvieron las redes neuronales descritas en la tabla 13.

Tabla 13: Redes neuronales generadas a partir de los nuevos mejores predictores de la variable TOTAL_INGRESO_TRABAJO

Referencia		Desempeño			Error				Redes	
		Entrena	Selec.	Prueba	Entrena	Selec-	Prueba	Entrenamiento	Ent.	Ocultas
1	MLP 1:3-1-1:1	1.000009	1.000017	0.999978	0.108455	0.108937	0.109771	BP100, CG20, CG0b	1	1
2	MLP 1:404-10-1:1	0.986640	0.997897	0.995957	0.102432	0.103870	0.104729	BP100, CG20, CG0b	1	10
3	MLP 1:404-7-1:1	0.988611	0.998479	0.997345	0.097398	0.098824	0.099728	BP100, CG20, CG0b	1	7
4	Linear 14:47-1:1	0.996392	0.997155	0.997777	0.097833	0.098386	0.099451	PI	14	0
5	Linear 16:57-1:1	0.995473	0.996821	0.996954	0.097742	0.098353	0.099369	PI	16	0
6	Linear 17:63-1:1	0.991275	0.993022	0.992159	0.097330	0.097978	0.098892	PI	17	0
7	RBF 14:486-163-1:1	0.985573	0.990434	0.988774	0.006273	0.006335	0.006389	SS, KN, PI	14	163
8	RBF 14:486-126-1:1	0.986139	0.990421	0.989537	0.006277	0.006335	0.006394	SS, KN, PI	14	126
9	RBF 14:486-252-1:1	0.982911	0.989139	0.989530	0.006257	0.006327	0.006394	SS,KN,PI	14	252
10	RBF 14:486-189-1:1	0.984555	0.988913	0.988344	0.006267	0.006325	0.006386	SS,KN,PI	14	189

Como se puede observar la red del tipo perceptrón multinivel con el número 3, la red del tipo lineal número 6, así como la red 7 y 10 del tipo RBF son las que tienen mayor

posibilidad de ser utilizadas en la predicción ya que ofrecen los valores más bajos al realizar la predicción en los datos de prueba.

Evaluación del modelo

La evaluación del modelo es una tarea importante, dado que nos ayuda a entender el desempeño del modelo con respecto al objetivo que se desea alcanzar.

A continuación se presentan los estadísticos obtenidos de las predicciones de cada una de las redes neuronales:

Tabla 14: Estadísticos de los ingresos generados a partir de las redes neuronales generadas

	RNA1	RNA2	RNA3	RNA4	RNA5
Media	69.3401	69.3401	69.3401	69.3401	69.3401
Desviación Estándar	157.8882	157.8882	157.8882	157.8882	157.8882
Error Medio	73.7172	-53.2077	-12.7167	0.4170	-0.0712
Error estándar	157.8887	156.6027	156.8348	157.4045	157.2865
Error medio absoluto	125.7335	61.3282	65.0477	71.6137	71.2150
proporción D.S	1.0000	0.9919	0.9933	0.9969	0.9962
Correlación	-0.0013	0.1275	0.1168	0.0783	0.0876

	RNA6	RNA7	RNA8	RNA9	RNA10
Media	69.3401	69.3401	69.3401	69.3401	69.3401
Desviación Estándar	157.8882	157.8882	157.8882	157.8882	157.8882
Error Medio	-0.0846	0.0453	0.1123	-0.9790	0.0738
Error estándar	156.6157	155.9324	156.0067	155.7044	155.7754
Error medio absoluto	70.5612	70.5903	70.6532	70.1398	70.5964
proporción D.S	0.9919	0.9876	0.9881	0.9862	0.9866
Correlación	0.1268	0.1573	0.1543	0.1665	0.1634

Al realizar la exploración de los datos observados y los predichos por las redes neuronales encontramos:

Tabla 15: Comparación de los ingresos observados y los predichos por las redes neuronales

Ingresos total por trabajo

	Observado	RNA1	RNA2	RNA3	RNA4	RNA5
Suma	5,443,890.00	11,231,431.29	1,266,557.68	4,445,503.39	5,476,632.13	5,438,297.95
Promedio	69.34	143.06	16.13	56.62	69.76	69.27
Diferencia	0.00	5,787,541.29	-4,177,332.32	-998,386.61	32,742.13	-5,592.05
% de Diferencia con respecto al observado	0.00	106.31	-76.73	-18.34	0.60	-0.10

Ingresos total por trabajo

	Observado	RNA6	RNA7	RNA8	RNA9	RNA10
Suma	5,443,890.00	5,437,247.89	5,447,448.67	5,452,710.23	5,367,026.18	5,449,686.96
Promedio	69.34	69.26	69.39	69.45	68.36	69.41
Diferencia	0.00	-6,642.11	3,558.67	8,820.23	-76,863.82	5,796.96
% de Diferencia con respecto al observado	0.00	-0.12	0.07	0.16	-1.41	0.11

Como se puede observar, en base a los datos de la muestra, la red neuronal *7* es la que ofrece mejores resultados.

Esta red neuronal es una red tipo Función Radial Básica (RBF, por sus siglas en inglés), la cual tiene una capa de entrada, una capa oculta y una capa de salida, las neuronas en la capa oculta utilizan funciones de trasferencia Gausianas cuyas salidas son inversamente proporcionales a la distancia desde el centro de la neurona.

Las redes RBF son muy similares a las redes PNN/GRNN. La principal diferencia, es que las redes PNN/GRNN tienen una neurona para cada entrada del archive de entrenamiento, mientras que las redes RBF tienen un número variable de neuronas, que por lo general es mucho menor que las entradas de entrenamiento, para problemas con conjuntos de entrenamiento pequeño o mediano, las redes PNN/GRNN son por lo general más precisas que las redes RBF, pero las redes PNN/GRNN no son prácticas para grandes conjuntos de entrenamiento.

Aunque la implementación es muy diferente, las redes neuronales RBF son conceptualmente similares a los modelos del K-Vecino Más Cercano (k-NN), la idea básica es que el valor objetivo a predecir de un elemento se es similar al de otros elementos que tienen valores cercanos de las variables predictoras.

Para el cálculo de las salidas de cada neurona, se calcula la distancia euclidiana desde el punto que se evalúa hasta el centro de cada neurona y una función radial básica es aplicada a la distancia para calcular el peso (influencia) de cada neurona.

$$Peso = \text{RBF}(distancia)$$

Cuanto más lejos se encuentre una neurona del punto evaluado, menor será su influencia.

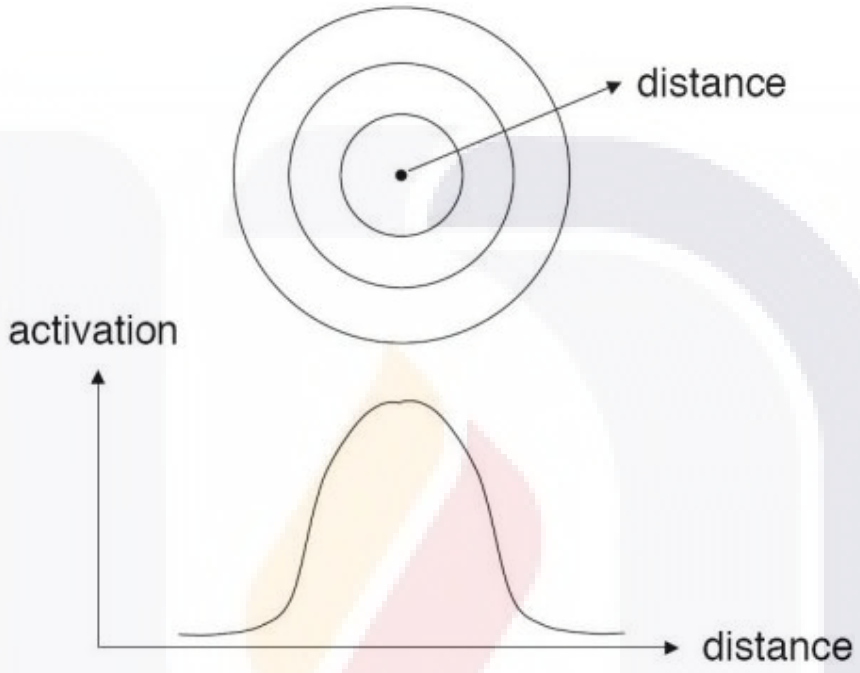


Figura 8: Función radial básica

Como se comento anteriormente la función utilizada en este tipo de redes es la Gausiana:

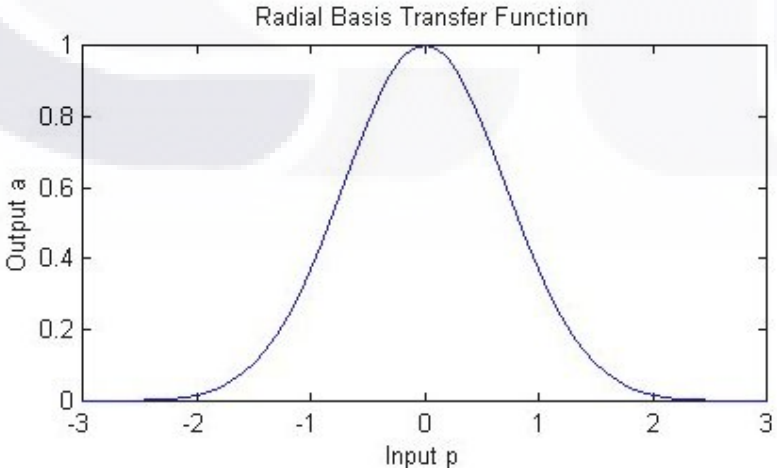


Figura 9: Grafica de la función radial básica de transferencia

Si existe más de una variable predictor, entonces la función RBF tendrá tantas dimensiones como variables existan, el mejor valor predicho para un punto, es el resultado de la suma de valores de la Funciones RBF multiplicadas por sus respectivos pesos.

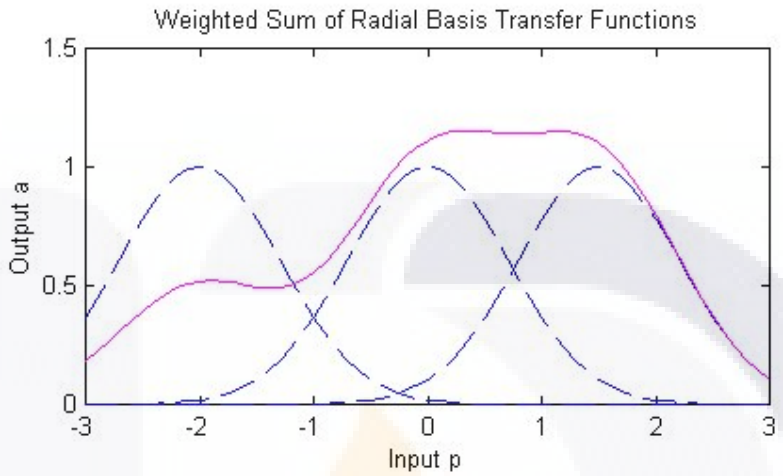


Figura 10: Suma ponderada de las funciones radiales básicas de transferencia

Las función radial de una neurona tienen un centro y un radio, el radio puede ser diferente para cada neurona, con un radio mayor, los puntos alejados del centro tienen más influencia, en otras palabras son menos selectivas.

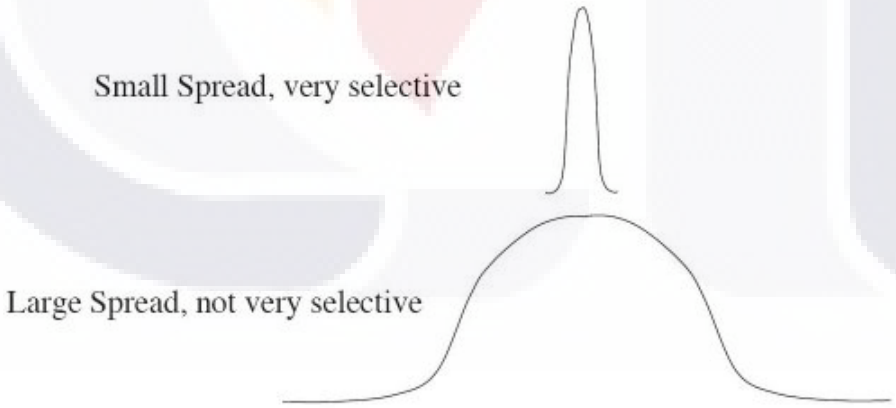


Figura 11: Cobertura de la selección de acuerdo al radio de la función radial

Arquitectura de una red RBF

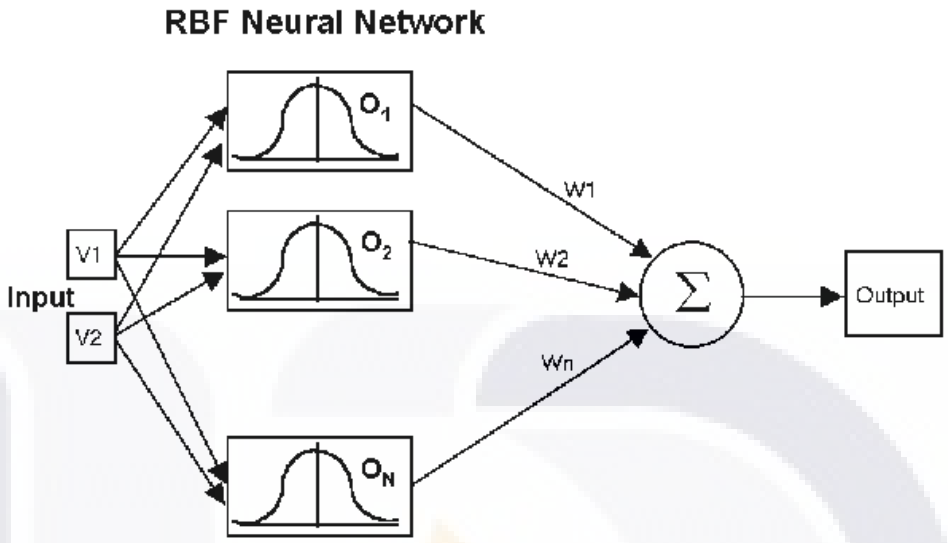


Figura 12: Arquitectura de una red RBF

Una red RBF tiene tres capas:

1. Capa de entrada – Existe una neurona en la capa de entrada para cada variable predictor, en el caso de variables categóricas N-1 neuronas son usadas donde N es el número de categorías. Las neuronas de entrada (o procesamiento antes de la capa de entrada) estandarizan el rango de valores restando la media y dividiendo el resultado por el rango intercuartil. Las neuronas de entrada entonces alimentan estos valores a cada una de las neuronas de la capa oculta.
2. Capa oculta, esta capa tiene un número variable de neuronas, el número optimo de neuronas se determina en el proceso de entrenamiento. Cada neurona consta de una función básica radial con centro en un punto de las dimensiones utilizadas (variables predictoras). El radio de la función RBF puede ser diferente en cada dimensión. El centro y el radio son determinados por el proceso de entrenamiento. Cuando se presenta el vector X de valores de entrada desde la capa de entrada, cada neurona oculta calcula la distancia Euclidiana desde el centro de la neurona y después aplica la función RBF central a la distancia utilizando el radio. El valor resultante es pasado a la capa de sumatoria.
3. Capa de agregación, los valores provenientes de la capa oculta son multiplicados por los pesos asociados a cada neurona y enviados a la agregación, la cual suma todos los valores ponderados y los presenta como la salida de la red. En la imagen, no se muestra el valor de desviación de 1.0 el cual es multiplicado por el valor W0 y que es enviado a la capa de agregación.

La red neuronal del tipo RBF, obtenida a través del software Statistica es la siguiente:

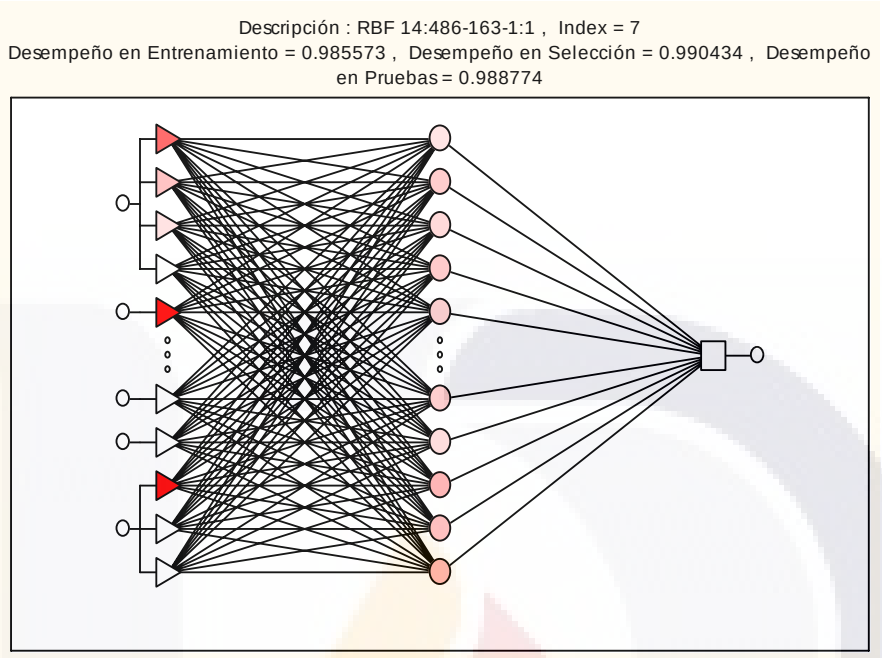


Figura 13: Red neuronal RBF para la predicción del ingreso por trabajo

4.4. Evaluación

Para la evaluación de los resultados de la minería de datos, las diversas redes neuronales fueron aplicadas a un conjunto mayor de datos, realizando agrupamientos por entidad, localidad y ageb, a continuación se muestran los resultados obtenidos:

Las predicciones de los ingresos observados contra los predichos presentan una diferencia del 13.22 en promedio con una desviación estándar de 243.58, lo cual nos indica la variabilidad de predicción, con una subestimación del ingreso.

Al comparar el número de casos en los cuales la predicción es idéntica a lo observado, ninguna predicción fue igual al valor observado.

Tabla 16: Estadísticos de los ingresos observados y predichos

	TOTAL_INGRESO_TRABAJO	TOTAL_INGRESO_TRABAJO RNA7
Media	81.58299743	68.35746324
Mediana	19	67.02567936
Desviación	245.722237	26.28021761
Sumatoria	35176141.000	29473687.4272
Mínimo	1.000	-28.3135
Máximo	3450.000	201.8212
_25%	11.000	49.5507
_75%	63.000	85.9247

Al comparar la media de los valores observados y predichos, se puede observar que la red neuronal esta prediciendo valores por debajo de la media observada, así mismo genera una menor variabilidad de los datos, lo cual puede ser observado por la desviación estándar 245.7 contra 26.28 de la predicción.

La sumatoria de los valores, indica que se tiene una predicción hacia valores inferiores que los observados, sin embargo resalta, que el primer cuartil, los datos observados alcanzan un valor de 11.00, mientras que los valores observados alcanzan un nivel de 49.55.

Otro indicador de que la predicción tiende a la subestimación son los valores del máximo y el mínimo, los cuales se observa una diferencias de 3248.17 y -27.31 respectivamente.

Tabla 17: Evaluación de los resultados obtenidos a partir de la red neuronal seleccionada

	Valores			
Entidad	TOTAL_INGRESO_TRABAJO	TOTAL_INGRESO_TRABAJO RNA7	Diferencia	% con respecto al observado
ZACATECAS	420,705.00	280,064.61	-140,640.39	-33.43
BAJA CALIFORNIA	1,321,485.00	899,237.53	-422,247.47	-31.95
CHIAPAS	1,409,023.00	965,820.38	-443,202.62	-31.45
YUCATAN	784,164.00	545,564.46	-238,599.54	-30.43
BAJA CALIFORNIA SUR	230,744.00	161,584.76	-69,159.24	-29.97
CHIHUAHUA	1,483,114.00	1,087,342.74	-395,771.26	-26.69
TAMAULIPAS	1,289,030.00	945,819.20	-343,210.80	-26.63
NAYARIT	365,937.00	273,236.13	-92,700.87	-25.33
TABASCO	655,219.00	492,835.56	-162,383.44	-24.78
COLIMA	238,528.00	179,862.30	-58,665.70	-24.59
SINALOA	1,036,929.00	787,491.66	-249,437.34	-24.06
AGUASCALIENTES	668,454.00	521,915.52	-146,538.48	-21.92
SAN LUIS POTOSI	760,697.00	601,433.83	-159,263.17	-20.94
SONORA	934,214.00	740,190.15	-194,023.85	-20.77
HIDALGO	758,044.00	601,404.19	-156,639.81	-20.66
TLAXCALA	325,511.00	259,401.45	-66,109.55	-20.31
COAHUILA DE ZARAGOZA	990,523.00	791,879.21	-198,643.79	-20.05
DURANGO	489,209.00	392,181.97	-97,027.03	-19.83
QUINTANA ROO	389,832.00	325,254.29	-64,577.71	-16.57
NUEVO LEON	1,694,854.00	1,424,407.95	-270,446.05	-15.96
GUERRERO	869,085.00	733,643.32	-135,441.68	-15.58
VERACRUZ-LLAVE	2,275,533.00	1,939,128.76	-336,404.24	-14.78
DISTRITO FEDERAL	4,013,611.00	3,421,178.30	-592,432.70	-14.76
QUERETARO DE ARTEAGA	452,367.00	408,526.88	-43,840.12	-9.69

	Valores			
Entidad	TOTAL_INGRESO_TRABAJO	TOTAL_INGRESO_TRABAJO RNA7	Diferencia	% con respecto al observado
ESTADO DE MEXICO	4,140,191.00	3,824,444.54	-315,746.46	-7.63
PUEBLA	1,446,765.00	1,339,300.80	-107,464.20	-7.43
MORELOS	501,110.00	470,787.38	-30,322.62	-6.05
OAXACA	895,224.00	842,521.22	-52,702.78	-5.89
MICHOACAN DE OCAMPO	1,027,160.00	974,443.63	-52,716.37	-5.13
JALISCO	2,121,750.00	2,030,452.96	-91,297.04	-4.30
GUANAJUATO	1,187,129.00	1,212,331.76	25,202.76	2.12
Total general	35,176,141.00	29,473,687.43	-5,702,453.57	-16.21

Como se puede observar a nivel general, se tiene una diferencia con respecto al observado de 16 por ciento menos, lo cual nos indica que la red neuronal generada esta prediciendo valores menores a los observados, cuando se analizan las salidas a nivel de entidad se puede observar que Guanajuato es el estado en el cual la predicción se encuentra por arriba de lo observado con un 2.12% más, la peor predicción la encontramos en el estado de Zacatecas, entidad en la cual se tiene una predicción de un 33.43 por ciento menos.

En general, la red neuronal generada obtiene un 19.25 por ciento menor que el observado, es decir, la red neuronal se aproxima al ingreso observado, un 80.75 por ciento.

4.5. Proceso de revisión

La revisión del proceso de minería de datos, se realizo el reprocesamiento de una nueva muestra, siguiendo los pasos realizados con anterioridad, se revisaron las condiciones de selección de los datos, así como la presencia de valores atípicos.

Debido a que el software Statistica es muy críptico con respecto a los modelos encontrados y en especial con los valores de los pesos y funciones de cada neurona, se utilizo WEKA en la búsqueda de una red neuronal, de tal forma que permita describir cada uno de los componentes del modelo.

Sin embargo, dado que las redes neuronales generadas por cada una de las herramientas, no fue posible obtener los pesos y bias de la red neuronal utilizada para la predicción.

Determinación de los próximos pasos

Para la continuidad de este proyecto de minería de datos, se proponen las siguientes acciones:

Realizar la programación de la red neuronal en otro lenguaje que permita la predicción en cada uno de los registros del Censo de Población y Vivienda, debido a que las herramientas utilizadas solo pueden ser aplicadas a una muestra de la información, con lo cual se obtendrá una visión más amplia de la aplicación de esta tecnología al proceso de generación de información estadística.

Para solventar los problemas en la predicción a niveles desagregados, se sugiere la implementación de modelos al nivel de desagregación deseada, verificando el siguiente nivel de predicción y de ser necesario generar una red neuronal para cada nivel de desagregación.

Aplicar otros métodos emergentes de minería de datos, tal como la optimización de colonia de hormigas.

4.6. Desarrollo

Plan de desarrollo

1. Programación del modelo en un lenguaje de programación
2. Aplicar el modelo a todo el universo de población ocupada.
3. Verificar los diversos niveles de desagregación
4. Rentrenar las redes seleccionadas para incrementar el poder de predicción.
5. Verificación de los valores predichos contra los observados
6. Al obtener un valor de predicción cercano a lo observado en cada nivel de desagregación deseado, proponer la eliminación de la pregunta sobre ingreso.

4.7. Comentarios finales

La minería de datos, es una herramienta la cual es una amalgama de tecnologías y disciplinas diversas entre las que se encuentran la estadística, sistemas inteligentes y bases de datos, estas técnicas usadas en conjunto permite el análisis de los datos y la posterior generación de conocimiento a partir de esos datos.

La utilización de la técnica de predicción utilizando redes neuronales permite la realización de modelos de predicción, los cuales tratan de aproximarse a la realidad, los resultados son por tanto variables de acuerdo al modelo que se utilice así como la técnica de modelado que se utilice.

Se deben de tomar en cuenta los obstáculos, ya que en el estado de arte actual, los realizadores de minería de datos, Los obstáculos encontrados en la realización del presente documento son:

Tiempo de procesamiento excesivo.

Utilización de modelos con un bajo poder de predicción.

Investigación de las herramientas a utilizar.

5. RESULTADOS

En este trabajo se presentó el proceso de minería de datos, el cual forma parte de un proceso mayor denominado Descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés), utilizando la metodología CRISP-DM como guía en la realización de las diversas tareas de minería de datos, así como la utilización principal del software Statistica para realizar las tareas de conocimiento de los datos, las redes neuronales se utilizaron como herramienta para la tarea de predicción del ingreso del trabajo.

Con los resultados obtenidos en el presente trabajo, se observa que es posible mejorar del diseño del cuestionario del Censo de Población y Vivienda 2010, dado que es posible la eliminación de preguntas, ya sea identificando a aquellas preguntas que tienen en mayor o menor medida una relación con otras variables o como identificar a aquellas variables susceptibles de ser eliminadas bajo la premisa de que las preguntas que se eliminen puedan ser inferidas por otras variables, en este trabajo se ha analizado la posibilidad de predicción de la variable relacionada con el ingreso por productos del trabajo, lo cual a la luz de los resultados obtenidos es posible.

Como lo indica Giuseppe Larossi [25], lo anterior se ve reflejado en la calidad de la información captada, ya que se reduce el tiempo de entrevista, obteniéndose como beneficios el menor costo y carga reducida de preguntas hacia el informante.

Por otra parte a través del uso de las tareas de preprocesamiento y limpieza de datos es posible mejorar la calidad de los datos utilizados para la realización de la minería de datos, ya que nos permiten realizar la selección de variables a través de las correlaciones tanto para las variables continuas como categóricas mediante el uso de del coeficiente R de Pearson y el coeficiente Tau de Kendall respectivamente, estos análisis permiten identificar variables duplicadas, lo que permite la reducción de atributos utilizados para la realización de la tarea de minería de datos elegida, así mismo al utilizar un diseño mejorado por medio de la reducción de las preguntas se mejora la calidad de los datos a captar, ya que se reduce el tiempo de entrevista, reduciendo la carga de preguntas al entrevistado.

Del proceso realizado en este trabajo, se deduce que la minería de datos puede ser utilizada en la actividad de mejora de cuestionarios, los análisis realizados se enfocaron a la predicción de la variable de ingreso del trabajo, sin embargo, también puede ser utilizada para la realización predicciones para variables categóricas (agrupamiento o clusterización).

Los datos utilizados para este trabajo, se obtuvieron del almacén de datos estadístico del INEGI, del proyecto Censo de Población y Vivienda 2000, realizando las tareas iníciales de la minería de datos, en lo general es posible la utilización de los datos tal y como se encuentran en el almacén sin embargo, es necesaria una recodificación de algunas variables, para evitar introducir sesgos a los modelos generado.

A partir del análisis de correlaciones de los datos del censo de población y vivienda 2000 contenido en el almacén de datos estadísticos es posible la eliminación de los siguientes reactivos:

De la sección II residentes, hogares y lista de personas, es posible la eliminación de la pregunta 1 (Número de personas), con respecto a la los datos captados en la caratula del cuestionario denominada “Total de cuestionarios en la vivienda”, dado que entre estos dos reactivos existe una correlación positiva perfecta.

De esta misma sección es posible la eliminación del la pregunta 3 (Número de hogares), debido a que existe una correlación casi perfecta (0.96) con el reactivo de la caratula denominado “TOTAL DE CUESTIONARIOS EN LA VIVIENDA”.

A partir de los experimentos sobre la predicción del ingreso por trabajo, los cuales han arrojado una diferencia del -16.21% a nivel nacional, es factible la eliminación de esta pregunta.

A partir de los resultados obtenidos en los experimentos realizados, el cuestionario propuesto puede observarse en el anexo 1.

6. CONCLUSIONES

Conclusiones

Este trabajo es una primera aproximación de la utilización de la minería de datos al rediseño de cuestionario, después de realizar este trabajo se muestran varias posibilidades para la investigación de este tema incluyendo la aplicación de nuevas técnicas de modelado.

Al realizar minería de datos es necesaria la utilización de esquemas de muestreo los cuales permitan obtener muestras representativas de los datos originales, los cuales nos permitan realizar los análisis necesarios con un mínimo de costo en tiempo.

Una vez que se obtiene la muestra, es necesario la selección de los campos, dado que pueden existir campos ya calculados que pueden ser utilizados para el análisis de la información, además existen campos que solo son identificadores de los campos y no aportan valor agregado a los análisis a realizar.

La técnica de predicción a través de redes neuronales ha permitido aproximarse a los valores observados, mediante la generación de un modelo, el cual obtiene una predicción a nivel global aproximada al observado, esto habilita el mejoramiento de los cuestionarios del censo de población y vivienda, dado que es posible realizar una predicción, tal y como se ha realizado en este trabajo a través de la inferencia de la variable de ingreso a partir de otras variables contenidas dentro del censo de población y vivienda del año 2000.

Sin embargo como se ha observado, al realizar la predicción para un nivel de detalle mayor, se muestran mayores variaciones entre lo observado y lo predicho, por lo cual la eliminación de la pregunta sobre ingreso por productos del trabajo debe de realizar con las reservas correspondientes.

Como lo han mostrado los resultados es posible mejorar la calidad de los datos recabados, al reducir preguntas en el cuestionario, enfocándose en aquellas preguntas que son atributos predictores de otros.

Como se ha expuesto en este documento. la minería de datos puede ser utilizada para el mejoramiento de los cuestionarios, ya que a través de la predicción de variables, es posible la reducción de las preguntas en el cuestionario.

Como ha quedado expuesto en el documento, la información contenida en el data warehouse, permite la estructuración de un proceso de minería de datos, en base a objetivos de negocio establecidos, resalta, la incorporación de metadatos sobre la información contenida, la cual permite que la curva de aprendizaje se vea disminuida considerablemente, por su parte, los cubos de información permiten en lapsos cortos de tiempo conocer la información sobre la cual se trabaja, permitiendo una visión global del problema y sus posibles soluciones, por otro lado es necesario realizar un arduo trabajo de recodificación de los valores de los atributos, dado que actualmente el utilizar los valores tal y como están pueden incorporar ruido a los modelos generados.

La minería de datos es un proceso que requiere como principal entrada información sobre los datos, es posible obtener una parte de esta información por medio de análisis estadístico, sin embargo, una cantidad enorme de información está entre los expertos de

cada uno de los temas, la generación de metadatos, es en la práctica una forma de concentrar este conocimiento de los expertos.

La aplicación de la minería de datos, ha demostrado sus bondades en diversas áreas, a estas se suma la aplicación en la generación de información estadística, permitiendo la predicción de los ingresos del trabajo de las personas a través de otras variables predictoras, la minería de datos junto con los avances tecnológicos permitirá en un futuro la realización de análisis de datos de una forma más rápida, cubriendo una mayor cantidad de datos, sin embargo, en el momento es posible la utilización de las técnicas del muestreo junto con la minería de datos para aproximarnos a la realidad que se desea medir.

Áreas del conocimiento aplicadas

Las áreas de conocimiento aplicadas en la realización de este trabajo son:

Sistemas de soporte a la toma de decisiones, utilizada para conocer los diferentes niveles de mando en la empresa y como están relacionados con las diferentes actividades, desde las transaccionales hasta la toma de decisiones.

Estadística, la cual fue utilizada para comprender y analizar los datos, sus relaciones y comportamientos, así como la utilización de las herramientas de software estadístico y la interpretación de los resultados obtenidos.

Sistemas de información inteligentes, utilizada como base para la incorporación de la red neuronal como la herramienta para realizar la predicción de los ingresos por productos del trabajo.

Data warehouse, utilizada como base para conocer las estructuras del data warehouse estadístico.

Bases de datos, utilizada como base para la realización de la estructuración y explotación de los datos para los propósitos de la minería de datos expuesta.

Minería de datos, utilizada como base para la utilización de las áreas de conocimiento anteriores y la estructuración del presente proyecto.

7. RECOMENDACIONES

El uso de metodologías para realizar minería de datos, es un soporte estructural para las personas que realicen este proceso, sin embargo, cada persona tiene la tarea de identificar dentro de esa metodología que se debe realizar en cada fase.

Busqué personal calificado de las diversas áreas involucradas, una sola persona puede obtener resultados, pero tal vez en un periodo de tiempo bastante prolongado, en cambio un conjunto de personas con cada uno de los perfiles, ofrecerá resultados en un periodo de tiempo corto.

Use muestreo, aun y que la metodología no lo indique como tal o lo considere como un paso secundario, el tener una muestra representativa, le permitirá realizar inferencias hacia toda la población.

La tarea de preprocesamiento de datos es una tarea importante, no olvide realizarla, aunque los propietarios de la información aseguren que la calidad de los datos es excepcional, ya que puede encontrarse con sorpresas y realidades en sus fuentes de datos.

Identifique claramente los atributos, en especial su tipo, ya que de esto depende el tiempo de procesamiento para encontrar el modelo.

Analice las relaciones entre los datos, muchas veces encontramos atributos que están directamente correlacionados con otros, y tarde o temprano quedarán fuera del análisis.

En lo general un minero de datos, deberá enfrentarse a tiempos de entrega del producto final, recuerde que la búsqueda de modelos y los análisis de la información consumirán la mayor parte del tiempo.

Todo es perfectible, cuando un modelo no le ofrezca los resultados deseados, no desespere, tarde o temprano encontrará alguno que satisfaga sus necesidades.

Como experiencia, un proyecto de minería de datos, no es una tarea que se deba de realizar por una sola persona, se requiere de un conjunto de personas que obtengan la mayor información de los datos y expertos en diversos temas como son: estadística, sistemas inteligentes y bases de datos.

8. BIBLIOGRAFÍA

[1] Alberto Ochoa, A. Tcherassi, i.Shinggavera, A. Pademéterakiris, J. Gyllenhaleale & Jose Alberto Hernández, Italianitá: Discovering a Pygmalion Effect on Italian Communities using Data Mining.

[2] Características Metodológicas del XII Censo General de Población y Vivienda 2000, Junio de 2002.

[3] Clementine genera el máximo retorno en el menor tiempo, SPSS, consultado de http://www.spss.com/es/clementine/, el día 22 de febrero de 2008.

[4] David ¡ Hand, Data mining: Statistics and More?, The American Statistician, May 1998, Vol. 52, No. 2.

[5] Juan Hernández, Estimación de agregados municipales utilizando información muestral censal, INEGI, diciembre de 2006.

[6] HAI LIU, YI-LUNG KUO, YI HE, JIN LIU, Bayesian Regression Model For Predicting Income, December 2006.

[7] INEGI, Data Warehouse para la presentación del servicio público de información estadística en México.

[8] INEGI, Identidad institucional consultado de http://intranet.inegi.gob.mx/Identidad/default.aspx el día 08 de noviembre de 2007.

[9] INEGI, Proceso Estándar para Realizar Encuestas por Muestreo, consultado de http://proyectos.inegi.gob.mx/de/invnor/normatividad/Documentos%20en%20revisin/PROCESO_ESTANDAR.pdf, el día 08 de noviembre de 2007.

[10] INEGI, Síntesis Metodológica del XII Censo General de Población y Vivienda 2000, Julio de 2003.

[11] INEGI, XII Censo de Población y Vivienda 2000: Cuestionario Básico, 2000.

[12] INEGI, XII Censo General de Población y Vivienda 200, consultado de http://www.inegi.gob.mx/est/contenidos/espanol/proyectos/censos/cpv2000/default.asp?c=701, el día 14 de noviembre de 2007.

[13] Instituto Nacional de Estadística Geográfia e Informática (INEGI), Diseño de Cuestionarios, Diciembre 2006.

[14] Jose Ramón Cano, Francisco Herera y Manuel Lozano, Extracción de modelos predictivos e interpretables en conjuntos de datos de gran tamaño mediante selección de conjuntos de entrenamiento, 2005.

[15] Jose Ramón Cano, Francisco Herera y Manuel Lozano, Using Evolutionary Algoritms as Instance Selection for Data Reduction in KDD: An Experimental Study, 2003.

[16] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A Krugan, Data Mining A Knowledge Discovery Approach, Springer, 2007.

[17] La estadística como instrumento de medida de un programa de intervención relacionadon con el medio ambiente, Mercedes Rodriguez Sánchez, Ma. Teresa Cabero Morán, José Chamoso y Ma. José Rodriguez Conde. ISSN 0214-9915 CODEN PSOTEGEN.

[18] Manual de WEKA, Diego García Morate.

[19] Nick Winder and Yinghui Zhou, Predicting the annual income of Swedish individuals:SMC Internal Discussion Paper, consultado de http://www3.umu.se/soc_econ_geography/smc/eng_publications.asp, el día 22 de febrero de 2008.

[20] OCDE, a Framework for biotechnology statistics, pág. 16, 2005,

[21] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Datamining, Pearson Adidson Wesley, 2006.

[22] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR),Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler),Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), CRISP-DM 1.0 Step-by-step data mining guide, 2000.

[23] Princenton University, Data and Statistical Services, consultado de http://dss.princeton.edu/online_help/analysis/dummy_variables.htm, el día 28 de enero de 2008.

[24] Problemas en el Diseño y Validación de Cuestionarios: tratamiento con QUESTPOT v.1.2, PEDRO ANTONIO GARCÍA LÓPEZ, ANDRÉS GONZÁLEZ CARMONA, JUAN ANTONIO MALDONADO JURADO, Departamento de Estadística e I.O.Universidad de Granada.1999

[25] The power of survey design, Giuseppe Larossi, The Work Bank, 2005.

9. GLOSARIO

Captación. Serie de actividades para obtener los datos de cada elemento de la población de estudio o una muestra de ella, siguiendo las estrategias determinadas en programas y procedimientos de trabajo.

CRISP-DM. Es un modelo de proceso estándar el cual no es propietario y disponible libremente, es descrito en términos de un modelo jerárquico de procesos, consiste en un grupo de tareas detalladas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada e instancia de proceso .

Cuestionario. Tipo de instrumento de captación, que presenta preguntas y/o enunciados dirigidos a los informantes, para obtener datos específicos acerca de las variables que serán objeto de captación.

Data Warehouse. Proceso de integración de información a partir de bases de datos transaccionales, la cual es organizada mediante modelos multidimensionales de información cuyo propósito es ofrecer información para su análisis por los niveles directivos.

Diseño conceptual. Serie de actividades mediante la cual se identifican las necesidades de información, que sirven para determinar: el marco conceptual, los instrumentos para la captación de los datos, los criterios de validación para la revisión y depuración de inconsistencias, así como los esquemas para la presentación de resultados.

Diseño de cuestionarios. Actividad del proceso de generación estadística, en la que deben combinarse de manera adecuada varios aspectos: sintaxis de la redacción, secuencia de las preguntas; formato y edición, con el fin de facilitar la captación y procesamiento de la información.

Diseño de la muestra. Serie de actividades para determinar: el método de muestreo por aplicar, bajo las consideraciones de cobertura y desglose temático y geográficos establecidos en el diseño conceptual, así como los insumos disponibles en cuanto al marco de muestreo de referencia y recursos financieros; el tamaño de muestra, procedimientos de selección, así como el diseño y cálculo de estimadores, con base en el análisis y elección de las mejores alternativas para el proyecto.

Estadística básica. Conjunto de datos obtenidos de un proyecto censal, de encuesta por muestreo o de aprovechamiento de registros administrativos, cuyo cálculo se realiza mediante operaciones matemáticas sin la aplicación de criterios o métodos que involucran conceptualizaciones ajenas al proyecto.

Instrumento de captación, Formato que se utiliza para el registro de los datos, en un proyecto estadístico; tal información se ha definido previamente y organizado en el marco conceptual.

Marco conceptual de un proyecto estadístico. Ordenamiento de temas, categorías, variables y clasificaciones al cual se referirán los datos objeto de captación, incluido el glosario con las definiciones formales de cada uno de los conceptos utilizados.

Metadato. Información referente a la conceptualización, generación o cálculo de uno o varios datos estadísticos.

Minería de datos. Proceso de descubrir automáticamente información útil en grandes repositorios de información

Muestra. Es el grupo de individuos que realmente se estudiarán, es un subconjunto de la población. Para que se puedan generalizar a la población los resultados obtenidos en la muestra, ésta ha de ser representativa de dicha población. Para ello, se han de definir con claridad los criterios de inclusión y exclusión y, sobre todo, se han de utilizar las técnicas de muestreo apropiadas para garantizar dicha representatividad.

Población. Es el conjunto de elementos o individuos que reúnen las características que se pretenden estudiar. Cuando se conoce el número de individuos que la componen, se habla de población finita y, cuando no se conoce su número, de población infinita.

Predicción. Acción de predecir el valor de un atributo en particular basando en el valor de otros atributos, el atributo a predecir es comúnmente conocido como variable objetivo o dependiente, mientras que los atributos usados para realizar la predicción son conocidos como variables explicatorias o independientes

Presentación de resultados. Serie de actividades para la elaboración de productos, definidos en el diseño conceptual y conforme a un Programa de Divulgación.

Procesamiento. Serie de actividades para preparar los archivos de datos, asegurándose que sean congruentes y ordenados para su aprovechamiento.

Proceso de generación de estadística básica. Series de actividades agrupadas con base en sus características similares, las cuales interactúan bajo distintos esquemas de orden y secuencia.

SEMMA. Es el acrónimo de Sample, Explore, Modify, Model, Assess, se refiere a la esencia del proceso de realización de minería de datos a partir de una muestra estadísticamente representativa de datos.

ANEXO 1CUESTIONARIO PROPUESTO

XII CENSO DE POBLACIÓN Y VIVIENDA 2000

Cuestionario básico

1. IDENTIFICACIÓN GEOGRÁFICA

ENTIDAD FEDERATIVA ____ |__|__|

MUNICIPIO O DELEGACIÓN ____ |__|__|__|

CLAVE DE AGEB ..... |__|__|__|__| – |__|

LOCALIDAD ____ |__|__|__|__|

MANZANA ..... |__|__|__|

SEGMENTO ..... |__|

2. CONTROL DE VIVIENDA Y CUESTIONARIOS

CONSECUTIVO DE LA VIVIENDA ..... |__|__|__|

NÚMERO DE HOGAR ..... |__|

TOTAL DE HOGARES EN LA VIVIENDA ..... |__|

TOTAL DE CUESTIONARIOS EN LA VIVIENDA ..... |__|

3. DIRECCIÓN DE LA VIVIENDA

____ CALLE, AVENIDA, CALLEJÓN, CARRETERA, CAMINO

NÚMERO EXTERIOR ____ NÚMERO INTERIOR ____ COLONIA, FRACCIONAMIENTO, BARRIO, UNIDAD HABITACIONAL

4. CONTROL DE PAQUETE

FOLIO DE PAQUETE ..... |__|__|__|

CONSECUTIVO DEL CUESTIONARIO EN EL PAQUETE ..... |__|__|

5. CLASE DE VIVIENDA

CÍRCULE UN SOLO CÓDIGO

CASA INDEPENDIENTE	1
DEPARTAMENTO EN EDIFICIO	2
VIVIENDA O CUARTO EN VECINDAD	3
VIVIENDA O CUARTO EN LA AZOTEA	4
LOCAL NO CONSTRUIDO PARA HABITACIÓN	5
VIVIENDA MÓVIL	6
REFUGIO	7

6. NOMBRE DE LOS RESPONSABLES

ENTREVISTADOR(A) ____ |__|__|

JEFE (A) DE ENTREVISTADORES ____ |__|__|

RESPONSABLE DE AGEB ____ |__|__|

VALIDADOR(A) ____ |__|__|

7. RESULTADO DE LA VALIDACIÓN

VALIDADO	1		
A VERIFICACIÓN POR ERROR EN:			
IDENTIFICACIÓN GEOGRÁFICA	2	GASTO COMÚN, NÚMERO DE HOGARES / CONTROL DE VIVIENDA	5
CONTROL DE VIVIENDA Y CUESTIONARIOS	3	LISTA DE PERSONAS / CARACTERÍSTICAS DE LAS PERSONAS	6
NÚMERO DE PERSONAS / LISTA DE PERSONAS	4	SEXO, EDAD / NÚMERO DE HIJOS	7

I. Características de la vivienda

1. PAREDES

¿De qué material es la mayor parte de las paredes o muros de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Material de desecho	1
Lámina de cartón	2
Lámina de asbesto o metálica	3
Carrizo, bambú o palma	4
Embarro o bajareque	5
Madera	6
Adobe	7
Tabique, ladrillo, block, piedra, cantera, cemento o concreto	8

2. TECHOS

¿De qué material es la mayor parte del techo de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Material de desecho	1
Lámina de cartón	2
Lámina de asbesto o metálica	3
Palma, tejamanil o madera	4
Teja	5
Losa de concreto, tabique, ladrillo o terrado con viguería	6

3. PISOS

¿De qué material es la mayor parte del piso de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Tierra	1
Cemento o firme	2
Madera, mosaico u otros recubrimientos	3

4. COCINA

¿Esta vivienda tiene un cuarto para cocinar?

CIRCULE UN SOLO CÓDIGO

Sí 1

No ..... 2 → PASE A 5

En el cuarto donde cocinan, ¿también duermen?

CIRCULE UN SOLO CÓDIGO

Sí 3

No 4

5. NÚMERO DE CUARTOS

¿Cuántos cuartos se usan para dormir sin contar pasillos?

|__|__|

ANOTE CON NÚMERO

Sin contar pasillos ni baños, ¿cuántos cuartos tiene en total esta vivienda? Cuente la cocina.

|__|__|

ANOTE CON NÚMERO

6. DISPONIBILIDAD DE AGUA

¿En esta vivienda tienen:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

agua entubada dentro de la vivienda?	1
agua entubada fuera de la vivienda, pero dentro del terreno?	2
agua entubada de llave pública (o hidrante)?	3
agua entubada que acarrean de otra vivienda?	4
agua de pipa?	5
agua de un pozo, río, lago, arroyo u otra?	6

7. SERVICIO SANITARIO

¿Esta vivienda tiene:

excusado o sanitario?
retrete o fosa?
letrina?
hoyo negro o pozo ciego?

CIRCULE UN SOLO CÓDIGO

Sí 1

No ..... 2 → PASE A 10

8. USO EXCLUSIVO

¿Este servicio lo usan solamente las personas de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

9. CONEXIÓN DE AGUA

¿Este servicio sanitario:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

tiene conexión de agua?	1
le echan agua con cubeta?	2
¿No se le puede echar agua?	3

Continúe con la pregunta 10 ➡

10. DRENAJE

¿Esta vivienda tiene drenaje o desagüe de aguas sucias:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

a la red pública?	1
a una fosa séptica?	2
a una tubería que va a dar a una barranca o gr...	3
a una tubería que va a dar a un río, lago o mar?	4
¿No tiene drenaje	5

11. ELECTRICIDAD

¿Hay luz eléctrica en esta vivienda?

CIRCULE UN SOLO CÓDIGO

Sí	1
No	2

12. COMBUSTIBLE

¿El combustible que más usan para cocinar es:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

gas?	1
leña?	2
carbón?	3
petróleo?	4
electricidad?	5

13. TENENCIA

¿Esta vivienda es propiedad de alguna persona que vive aquí?

CIRCULE UN SOLO CÓDIGO

Sí ..... 1 → PREGUNTE

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

¿Está pagando?	3
¿Está totalmente pagada?	4
¿Está en otra situación?	5

No ..... 2 → PREGUNTE

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

¿Está rentada?	6
¿Está prestada, la cuidan o en otra situación?	7

14. BIENES EN LA VIVIENDA

¿En esta vivienda tienen:

LEA TODAS LAS OPCIONES Y CIRCULE EL CÓDIGO SEGÚN LA RESPUESTA

	Sí	No
radio o radiograbadora?	1	2
televisión?	3	4
videocasetera?	5	6
licuadora?	7	8
refrigerador?	1	2
lavadora?	3	4
teléfono?	5	6
calentador de agua (boiler)?	7	8
automóvil o camioneta propios?	1	2
computadora?	3	4

Continúe con la siguiente sección

II. Residentes, hogares y lista de personas

2. GASTO COMÚN

¿Todas las personas que viven en esta vivienda comparten un mismo gasto para la comida?

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

CUANDO EN LA VIVIENDA EXISTA MÁS DE UN HOGAR O GRUPO DE PERSONAS, APLIQUE UN CUESTIONARIO PARA CADA HOGAR A PARTIR DE LA LISTA DE PERSONAS

4. LISTA DE PERSONAS EN EL HOGAR

Por favor, dígame el nombre de las personas que viven en su hogar, empezando por el jefe o la jefa; déme también el nombre de los niños/chiquitos y los ancianos (incluya a los sirvientes que duermen aquí):

PERSONA 1

PERSONA 2

PERSONA 3

PERSONA 4

PERSONA 5

PERSONA 6

SI EN EL HOGAR Y MÁS DE 6 PERSONAS, UTILICE OTRO CUESTIONARIO Y CONTINÚE CON LA LISTA

Copie el nombre de todas las personas en los espacios destinados para ello en la Sección III y haga las preguntas usando el nombre de cada una de las personas.

III. Características de las personas

Ahora le voy a preguntar por (NOMBRE):

PERSONA ___ _____ Anote el nombre de la persona

1. PARENTESCO

¿Qué es (NOMBRE) del jefe(a) del hogar?

SI ES EL JEFE(A) SÓLO CONFIRME Y CIRCULE UN SOLO CÓDIGO

Jefe(a) 1

Esposo(a) o compañero(a) 2

Hijo(a) 3

Otro _____ ANOTE EL PARENTESCO

2. SEXO

(NOMBRE) es mujer

(NOMBRE) es hombre

CIRCULE UN SOLO CÓDIGO

Hombre 1

Mujer 2

3. EDAD

¿Cuántos años cumplidos tiene (NOMBRE)?

MENOR DE UN AÑO, ANOTE "000"

/___/___/___/ ANOTE CON NÚMERO

4. LUGAR DE NACIMIENTO

¿En qué estado de la República o en qué país nació (NOMBRE)?

Aquí, en este estado 1

En otro estado _____ ANOTE EL ESTADO

En otro país _____ ANOTE EL PAÍS

5. DERECHOHABIENCIA

¿Tiene (NOMBRE) derecho a servicio médico en:

LEA TODAS LAS OPCIONES Y CIRCULE LAS RESPUESTAS AFIRMATIVAS

el Seguro Social (IMSS)? 1

el ISSSTE? 2

Pemex, Defensa o Marina? 3

otra institución:

_____ ANOTE LA INSTITUCIÓN

Entonces, no tiene derecho a servicio médico 5

6. TIPO DE DISCAPACIDAD

¿(NOMBRE) tiene limitación para:

LEA TODAS LAS OPCIONES Y CIRCULE LAS RESPUESTAS AFIRMATIVAS

moverse, caminar o lo hace con ayuda? 1

usar sus brazos y manos? 2

¿Es sordo(a) o usa un aparato para oír? 3

¿Es mudo(a)? 4

¿Es ciego(a) o sólo ve sombras? 5

¿Tiene algún retraso o deficiencia mental? 6

¿Tiene otra limitación física o mental?

_____ ANOTE LA LIMITACIÓN

Entonces, no tiene limitación física o mental 8

PARA PERSONAS DE 5 AÑOS CUMPLIDOS O MÁS

7. ENTIDAD O PAÍS DE RESIDENCIA EN 1995

Hace 5 años, en enero de 1995, ¿en qué estado de la República o en qué país vivía (NOMBRE)?

Aquí, en este estado 1

En otro estado _____ ANOTE EL ESTADO

En otro país _____ ANOTE EL PAÍS → PASE A 9

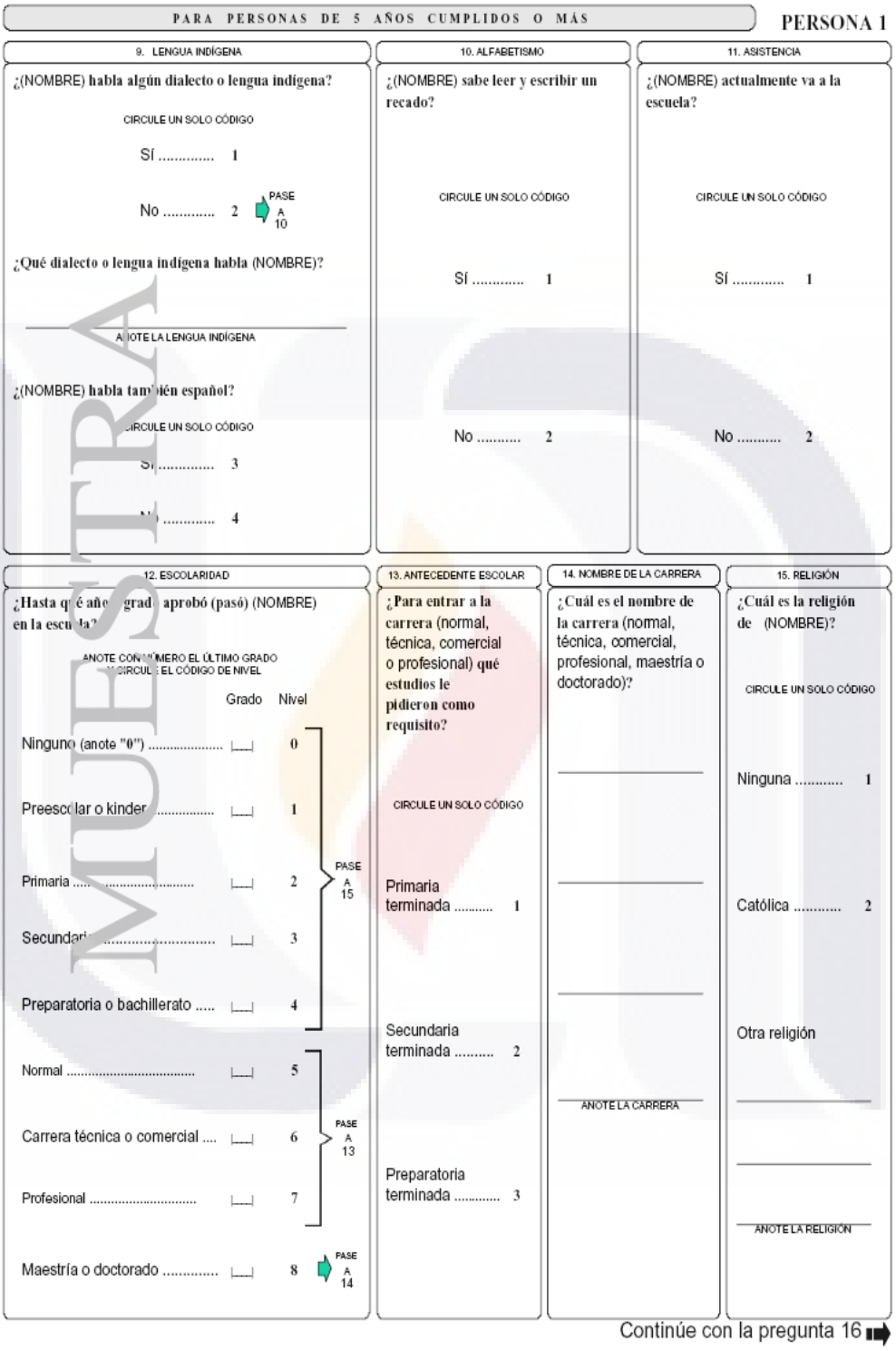
8. MUNICIPIO DE RESIDENCIA EN 1995

¿En qué municipio (delegación) vivía (NOMBRE) en enero de 1995?

Aquí, en este municipio o delegación 2

En otro municipio o delegación _____ ANOTE EL MUNICIPIO O DELEGACIÓN

Continúe con la pregunta 9



PARA PERSONAS DE 5 AÑOS CUMPLIDOS O MÁS

PERSONA 1

9. LENGUA INDÍGENA

¿(NOMBRE) habla algún dialecto o lengua indígena?

CIRCULE UN SOLO CÓDIGO

Sí 1

No ..... 2 ➡ PASE A 10

¿Qué dialecto o lengua indígena habla (NOMBRE)?

ANOTE LA LENGUA INDÍGENA

¿(NOMBRE) habla también español?

CIRCULE UN SOLO CÓDIGO

Sí 3

No 4

10. ALFABETISMO

¿(NOMBRE) sabe leer y escribir un recado?

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

11. ASISTENCIA

¿(NOMBRE) actualmente va a la escuela?

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

12. ESCOLARIDAD

¿Hasta qué año o grado aprobó (pasó) (NOMBRE) en la escuela?

ANOTE CON NÚMERO EL ÚLTIMO GRADO Y CIRCULE EL CÓDIGO DE NIVEL

	Grado	Nivel
Ninguno (anote "0")	⌊__⌋	0
Preescolar o kinder	⌊__⌋	1
Primaria	⌊__⌋	2
Secundaria	⌊__⌋	3
Preparatoria o bachillerato	⌊__⌋	4
Normal	⌊__⌋	5
Carrera técnica o comercial	⌊__⌋	6
Profesional	⌊__⌋	7
Maestría o doctorado	⌊__⌋	8 ➡ PASE A 14

PASE A 15

PASE A 13

13. ANTECEDENTE ESCOLAR

¿Para entrar a la carrera (normal, técnica, comercial o profesional) qué estudios le pidieron como requisito?

CIRCULE UN SOLO CÓDIGO

Primaria terminada 1

Secundaria terminada 2

Preparatoria terminada 3

14. NOMBRE DE LA CARRERA

¿Cuál es el nombre de la carrera (normal, técnica, comercial, profesional, maestría o doctorado)?

ANOTE LA CARRERA

15. RELIGIÓN

¿Cuál es la religión de (NOMBRE)?

CIRCULE UN SOLO CÓDIGO

Ninguna 1

Católica 2

Otra religión

ANOTE LA RELIGIÓN

Continúe con la pregunta 16 ➡

PERSONA 1 — PARA PERSONAS DE 12 AÑOS CUMPLIDOS O MÁS

16. ESTADO CONYUGAL

¿Actualmente (NOMBRE):

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- vive con su pareja en unión libre?... 1
- está separado(a)? 2
- está divorciado(a)? 3
- es viudo(a)? 4
- está casado(a):
  - ¿Sólo p... 5
  - ¿Sólo religiosamente? 6
  - ¿Civil y religiosamente? 7
- está soltero(a)? 8

17. CONDICIÓN DE ACTIVIDAD

¿La semana pasada (NOMBRE):

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- trabajó? ..... 1 → PASE A 19
- tenía trabajo, pero no trabajó? ..... 2 → PASE A 19
- buscó trabajo? 3
- ¿Es estudiante? 4
- ¿Se dedica a los quehaceres de su hogar? 5
- ¿Es jubilado(a) o pensionado(a)? 6
- ¿Está incapacitado(a) permanentemente para trabajar? ..... 7 → PASE A 24
- ¿No trabaja? 8

18. VERIFICACIÓN DE ACTIVIDAD

Además de (RESPUESTA DE 17), ¿la semana pasada (NOMBRE):

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- ayudó en un negocio familiar? 1
- vendió algún producto? 2
- hizo algún producto para vender? 3
- ayudó trabajando en el campo o en la cría de animales? 4
- a cambio de un pago realizó otro tipo de actividad? Por ejemplo: lavó o planchó ajeno, cuidó coches 5
- ¿No trabaja? ..... 6 → PASE A 24

19. OCUPACIÓN U OFICIO

¿Qué hizo (NOMBRE) en su trabajo de la semana pasada?

_____

_____

ANOTE LAS ACTIVIDADES O TAREAS

¿Cuál es el nombre de su ocupación, oficio o puesto?

Por ejemplo: campesino(a), maestro(a) de primaria, vendedor(a) ambulante.

_____

_____

ANOTE LA OCUPACIÓN, OFICIO O PUESTO

20. SITUACIÓN EN EL TRABAJO

¿(NOMBRE) en su trabajo de la semana pasada fue:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- empleado(a) u obrero(a)? 1
- jornalero(a) o peón? 2
- patrón(a)? (contrata trabajadores) 3
- trabajador(a) por su cuenta? 4
- trabajador(a) sin pago en el negocio o predio familiar? 5

Continúe con la pregunta 21

PERSONA 1

21. HORAS TRABAJADAS	22. ACTIVIDAD ECONÓMICA
En total, ¿cuántas horas trabajó (NOMBRE) la semana pasada? |___|___|___| ANOTE CON NÚMERO	**¿En dónde trabajó (NOMBRE) la semana pasada?** Por ejemplo: en el campo, en una fábrica, en un taller mecánico. _____ _____ ANOTE EN DÓNDE TRABAJÓ **El negocio, empresa o lugar donde trabajó ¿a qué se dedica?** Por ejemplo: a cultivar maíz, a hacer muebles, a vender ropa. ANOTE A QUÉ SE DEDICA _____ _____

PARA MUJERES DE 12 AÑOS CUMPLIDOS O MÁS

23. NÚMERO DE HIJOS	24. HIJOS FALLECIDOS	25. HIJOS SOBREVIVIENTES	26. FECHA DE NACIMIENTO	27. SOBREVIVENCIA	28. EDAD AL MORIR
En total, ¿cuántas hijas e hijos que nacieron vivos ha tenido (NOMBRE)? NINGUNO, ANOTE "00" Y PASE A LA SIGUIENTE PERSONA |___|___| ANOTE CON NÚMERO	**De las hijas e hijos que nacieron vivos ¿cuántos han muerto?** NINGUNO, ANOTE "00" |___|___| ANOTE CON NÚMERO	**¿Cuántas de las hijas e hijos de (NOMBRE) viven actualmente?** NINGUNO, ANOTE "00" |___|___| ANOTE CON NÚMERO	**¿En qué mes y año nació la última hija o hijo nacido vivo de (NOMBRE)?** ANOTE EL MES Y EL AÑO Mes ..... |___|___| y Año |___|___|___|___|	**Esta última hija o hijo de (NOMBRE) ¿vive actualmente?** CIRCULE UN SOLO CÓDIGO Sí .... 1 ➡ PASE A LA SIGUIENTE PERSONA No .. 2	**¿Qué edad tenía cuando murió?** ANOTE SÓLO UNA RESPUESTA EN: DÍAS O MESES O AÑOS SI VIVIÓ MENOS DE UN DÍA ANOTE "00" EN DÍAS Días ..... |___|___| o Meses .. |___|___| o Años ..... |___|___|

Pase a la persona 2 ➡

CONFIDENCIALIDAD

Conforme a las disposiciones del **Artículo 38 de la Ley de Información Estadística y Geográfica en vigor**, "Los datos e informes que los particulares proporcionen para fines estadísticos o provengan de registros administrativos o civiles, serán manejados, para efectos de esta Ley, bajo la observancia de los principios de confidencialidad y reserva y no podrán comunicarse, en ningún caso, en forma nominativa o individualizada, ni harán prueba ante autoridad administrativa o fiscal, ni en juicio o fuera de él."

OBLIGATORIEDAD

De acuerdo con el **Artículo 42, párrafo primero, de la Ley de Información Estadística y Geográfica en vigor**, "Los informantes estarán obligados a proporcionar con veracidad y oportunidad los datos e informes que les soliciten las autoridades competentes para fines estadísticos, censales y geográficos, y a prestar el auxilio y cooperación que requieran las mismas."