# BENEMÉRITA

# UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

# “ANÁLISIS DE FACTORES EDUCATIVOS QUE INFLUYEN EN EL DESEMPLEO EN MÉXICO”

## TESIS PROFESIONAL

PARA OBTENER EL TITULO DE:

**LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN**

PRESENTA

**JESÚS TEODORO TEPOXTECATL BAILON**

ASESOR

**DRA MARÍA JOSEFA SOMODEVILLA GARCIA**

**PUEBLA, MÉXICO** **2014**

# Agradecimientos

Agradezco…

…Antes que nada a Dios, por permitirme realizar las metas que me he propuesto y otorgarme el don de la salud así como el llenar mi vida de dicha.

…A mis padres quienes me han apoyado incondicional mente para que yo logre esta meta, pero sobre todo por su amor, cariño y comprensión, en todo momento los llevo conmigo en las enseñanzas que me han brindado.

…A mis hermanas por la compañía, el apoyo que me brindan, motivarme a superarme y demostrarme que puedo contar con ellas en todo momento.

….. A mis amigos que también forman parte de mi familia y que me han apoyado en las buenas y en las malas, a ellos que me brindan su confianza, amistad, y sobre todo por aconsejarme en los momentos que más se necesitan.

…Finalmente pero no menos importante agradezco de manera especial y sincera a la Dra. María Josefa Somodevilla García por aceptarme para realizar esta tesis bajo su dirección. Su apoyo y confianza en mi trabajo y su capacidad para guiar mis ideas ha sido un aporte invaluable.

Este trabajo es para todos ustedes.

# Resumen

El tema del desempleo, ha sido objeto de declaraciones, opiniones y comentarios, para dar paso a la polémica, muchas de ellas dignas de valor, peso y mucho respeto, otras muy superfluas, poco serias y que confluyen en la minimización del problema que acarrea, pasando desapercibidos e indolentes. Ante el triste panorama que se tiene, se ha lastimado con ello a un gran número de personas desempleadas y subempleadas, cuyas expectativas de mejorar sus niveles de vida se alejan cada vez más de sus posibilidades.

Actualmente el desempleo es un problema socioeconómico que afecta seriamente a la República Mexicana. Dicho problema hoy día también se ve agudizado por los niveles educativos de las personas en busca de empleo y principalmente por la deserción escolar.

El problema que se abordará en este proyecto, tiene sus orígenes, en datos obtenidos por el Instituto Nacional de Estadística Geográfica e Informática (INEGI) con respecto al desempleo así como los niveles de educación que hay en cada estado de la República Mexicana. Estos datos han sido recabados del censo del año 2005 al año 2011. Debido al volumen de información con que se cuenta, se llevará a cabo un proceso de extracción de conocimientos, basado en almacenes y minería de datos, con el objetivo de sentar bases para analizar el impacto de la educación respecto la obtención de empleo.

# Índice General

Agradecimientos ..... 2
Resumen ..... 3
Índice General ..... 4
Índice de Figuras ..... 7
Índice de Tablas ..... 8
CAPÍTULO 1 ..... 9
INTRODUCCIÓN ..... 9
1.1 PLANEAMIENTO DE LA INVESTIGACIÓN ..... 10
1.1.1 PROBLEMA A RESOLVER ..... 10
1.1.2 OBJETIVOS DE LA INVESTIGACIÓN ..... 10
1.1.3 JUSTIFICACIÓN DE LA INVESTIGACIÓN ..... 11
1.2 PRESENTACIÓN DE LA SOLUCIÓN ..... 11
1.2.1 PROPUESTA DE SOLUCIÓN ..... 12
1.3 APORTACIONES DE LA INVESTIGACIÓN ..... 12
1.4 ORGANIZACIÓN DE LA TESIS ..... 12
1.5 CONCLUSIONES ..... 13
CAPÍTULO 2 ..... 14
ESTADO DEL ARTE ..... 14
2.1 PROYECTOS Y ESFUERZOS PARA LA MITIGACIÓN DEL DESEMPLEO EN MEXICO ..... 15
2.1.1 OCDE ..... 15
2.1.2 INDICADORES OPORTUNOS DE OCUPACIÓN Y EMPLEO ..... 17
2.1.3 CEPAL ..... 17
2.2 COMPARATIVA DE LOS TRABAJOS. ..... 19
2.3 DESEMPLEO Y LA TOMA DE DECISIONES. ..... 19
2.4 CONCLUSIONES ..... 20
CAPÍTULO 3 ..... 21

MARCO TEÓRICO ..... 21
3.1 ALMACENES DE DATOS ..... 21
3.1.1 OLAP y OLTP ..... 22
3.1.2 ALMACENES DE DATOS Y BASES DE DATOS TRANSACCIONALES ..... 24
3.2 RELACIÓN DE LA MINERIA DE DATOS CON OTRAS DISCIPLINAS. ..... 25
3.3 METODOLOGÍA KDD ..... 27
3.3.1 LOS SEIS PASOS DEL KDD ..... 27
3.4 TAREA Y TÉCNICAS DE MINERÍA DE DATOS ..... 29
3.4.1 TAREAS PREDICTIVAS ..... 30
3.4.2 TAREAS DESCRIPTIVAS ..... 31
3.4.2 TAREAS DESCRIPTIVAS ..... 32
3.5 SOFTWARE PARA DATA MINING ..... 33
3.5.1 WEKA (Waikato Environment for Knowledge Analysis) ..... 33
3.6 CONCLUSIONES ..... 36
CAPÍTULO 4 ..... 37
ANÁLISIS Y DISEÑO ..... 37
4.1 PLANTEAMIENTO Y REQUERIMIENTOS ..... 37
4.2 FASE DE INTEGRACIÓN Y RECOPILACIÓN ..... 38
4.3 FASE DE SELECCIÓN Y LIMPIEZA Y TRANSFORMACIÓN ..... 40
4.3.1 TRATAMIENTO DE DATOS ..... 42
4.4 CONSTRUCCIÓN DEL ALMACÉN DE DATOS ..... 44
4.5 FASE DE MINERÍA DE DATOS ..... 46
4.5 CONCLUSIONES ..... 47
CAPÍTULO 5 ..... 48
EXPERIMENTOS ..... 48
5.1 EXPERIMENTO 1 – EM ..... 49
5.2 EXPERIMENTO 2 – K-MEANS ..... 50
5.3 EXPERIMENTO 3 – SELECCIÓN DE ATRIBUTOS ..... 53
5.4 EXPERIMENTO 4 – REMOVER ATRIBUTOS DERIVADOS ..... 55
5.5 COMPARACIÓN DE RESULTADOS ..... 58
CAPÍTULO 6 ..... 61
CONCLUSIONES ..... 61
6.1 CONOCIMIENTOS ADQUIRIDOS ..... 61

6.2 TRABAJO FUTURO ..... 62
6.3 CONCLUSIONES FINALES ..... 63
REFERENCIAS ..... 64

# Índice de Figuras

Figura 2. 1 Logotipo de OCDE ..... 16
Figura 2. 2 Logotipo de INEGI ..... 17
Figura 2. 3 Logotipo de CEPAL ..... 18

Figura 3. 1 Disciplinas que se relacionan con la Minería de Datos. ..... 26
Figura 3. 2 Procesos de extracción de conocimiento ..... 29
Figura 3. 3 Interfaz GUI de WEKA ..... 34
Figura 3. 4 Interfaz del entorno EXPLORER ..... 35

Figura 4. 1 Fases del proceso de descubrimiento en bases de datos, KDD ..... 38
Figura 4. 2 Visualización con WEKA de personas desempleadas con respecto al estado ..... 41
Figura 4. 3 Datos originales obtenidos de INEGI "Ocupacional" ..... 42
Figura 4. 4 Datos Finales ..... 43
Figura 4. 5 Modelo de la Base de Datos ..... 44

Figura 5. 1 Atributos de la vista minable. ..... 48
Figura 5. 2 Resultados del método EM. ..... 49
Figura 5. 3 Resultados de SimpleKMeans con K=5. ..... 50
Figura 5. 4 Resultados de SimpleKMeans con K=4. ..... 51
Figura 5. 5 Resultados de SimpleKMeans con K=6. ..... 52
Figura 5. 6 Resultados de aplicar selección de atributos supervisada. ..... 53
Figura 5. 7 Atributos restantes de pues de aplicar la selección de atributos. ..... 54
Figura 5. 8 Resultados de SimpleKMeans con K=6 des pues de la selección de atributos. ..... 55
Figura 5. 9 Atributos finales ..... 56
Figura 5. 10 Resultados de SimpleKMeans con K=6 con atributos finales. ..... 57

# Índice de Tablas

Tabla 3. 1 Diferencias entre una base de datos transaccional y un almacén de datos. ..... 25
Tabla 3. 2 Aplicaciones de las técnicas de minería de datos [1] ..... 32

Tabla 4. 1 Atributos de los datos de educación obtenidos por INEGI ..... 39
Tabla 4. 2 Atributos de los datos de desempleo obtenidos por INEGI ..... 39
Tabla 4. 3 Atributos de la tabla Hechos ..... 45
Tabla 4. 4 Atributos de la tabla UG ..... 45
Tabla 4. 5 Atributos de la tabla Tiempo ..... 45
Tabla 4. 6 Atributos de la tabla Datos ..... 46

Tabla 5. 1 Comparativa de resultados. ..... 58

# CAPÍTULO 1

## INTRODUCCIÓN

Actualmente el desempleo es un problema socioeconómico que afecta seriamente a la República Mexicana definamos al desempleo como “Situación en la que se encuentran las personas que teniendo edad, capacidad y deseo de trabajar no pueden conseguir un puesto de trabajo viéndose sometidos a una situación de paro forzoso” [1]. Dicho problema hoy día también se ve afectado por los niveles educativos que las personas en busca de empleo tienen y principalmente por la deserción escolar.

Con las nuevas tecnologías, se dispone de una gran cantidad de información histórica, la cual puede ser utilizada para plantear estrategias de trabajo las cuales ayudarían con la toma de decisiones, es decir, usar la información como una materia prima de la cual se pueda extraer conocimiento e información útil y novedosa. En las herramientas actuales se considera a los *DataWarehouse* (Almacenes de Datos) y Data Mining (Minería de Datos).

Un *DataWarehouse* es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta [2].La información almacenada es fiable y homogénea. Se necesita de una herramienta para la toma de decisiones basándose en información integrada y global, que facilite la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén, proporcione la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios. En concreto tener una herramienta que ofrezca una optimización tecnológica y económica en entornos de centro de información, estadística o de generación de informes. [3]

En esta tesis se desarrolla un *DataWarehouse* para tomar decisiones con respecto al desempleo y el impacto de los niveles educativos, el cual posteriormente será explotado por medio de técnicas de minería de datos. La información utilizada para la realización de dicho proyecto fue tomada de las bases de datos del SIMBAD (Sistema Estatal y Municipal de Bases de Datos). Dicha información es histórica, es decir, representa hechos que se han producido a lo largo de los años.

Este proyecto conlleva dos retos para la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos, entre otros). Y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. La combinación de las herramientas antes mencionadas facilita y acelera el procesamiento de la información obteniendo información clara, concisa y confiable.

## 1.1 PLANEAMIENTO DE LA INVESTIGACIÓN

En esta sección se describe el problema a resolver, se precisan los objetivos que se desean alcanzar, así como el planteamiento de la solución propuesta para dicho problema, para finalmente describir la organización de la tesis.

### 1.1.1 PROBLEMA A RESOLVER

El problema que se abordará en este proyecto, tiene sus orígenes, en datos obtenidos por el Instituto Nacional de Estadística Geográfica e Informática (INEGI) con respecto al desempleo así como los niveles de educación que hay en cada estado de la República Mexicana. Estos datos han sido recabados del censo del año 2005 al año 2011. Debido al volumen de información se aplicará un proceso de extracción de conocimientos, con el objetivo de sentar bases para analizar el impacto que la educación y sus niveles tiene con la obtención de empleo.

### 1.1.2 OBJETIVOS DE LA INVESTIGACIÓN

El objetivo general es el siguiente:

Desarrollar un sistema para la toma de decisiones para analizar el caso de desempleo en cualquier zona de la República Mexicana, utilizando la metodología KDD (*Knowledge Discovery in Databases*).

Los objetivos particulares se enlistan a continuación:

1. Desarrollar el almacén de datos usando de la metodología de Kimball para su posterior implementación en MySQLServer 5.0.
2. Descubrir los patrones en el DW para la predicción del impacto del desempleo con respecto al nivel educativo alcanzado por año, mediante el uso de técnicas de minería de datos.
3. Realizar un análisis estadístico de la población desempleada centrándose en una división de zonas.

### 1.1.3 JUSTIFICACIÓN DE LA INVESTIGACIÓN

El desempleo, es un problema que afecta a todos los países, independientemente que sean pobres, en desarrollo e industrializados, en México es una de las consecuencias el rezago económico así como en el rezago educativo y se ha convertido en un problema grave por el impacto a nivel social y económico.

Ante la duda de saber realmente el impacto que tiene la educación con que las personas estén desempleadas, se busca la utilización de nuevas herramientas de tratamiento de información (bases de datos, *Data Warehouse*, *Data Mining*.) que mejoran la calidad, cantidad y eficiencia de los datos, así como el análisis, procesamiento y comunicación de los mismos. En otras palabras, pueden aportar a la institución la base tecnológica necesaria para afrontar los nuevos retos de la situación actual y las perspectivas a futuro. De ahí, que en este trabajo, se resalte el hecho de que las bases de datos y el DW permiten en primera instancia el almacenamiento adecuado de los datos obtenidos de las tazas de desempleo así como los niveles de aprovechamiento educativo.[4]

Pero, además se incide en otro hecho, que es el que a través de dichas herramientas la institución puede extraer de dichos datos, la información y el conocimiento que necesitan para identificar y responder estratégicamente sus necesidades y retos que enfrentan.

## 1.2 PRESENTACIÓN DE LA SOLUCIÓN

La propuesta de solución al problema definido en la sección 1.1.1 y los productos desarrollados son comentados en la siguiente subsección.

### 1.2.1 PROPUESTA DE SOLUCIÓN

Se propone el diseño e implantación de un *Data Warehouse* multidimensional para la toma de decisiones, referente al tema de Factores Educativos que influyen en el desempleo en México. Posteriormente a la creación del *Data Warehouse*, éste se explotará mediante técnicas adecuadas de minería de datos.

El sistema a desarrollar se fundamenta en Bases de Datos Relacionales y la construcción de cubos OLAP en almacenes de datos. Todo esto es desarrollado en MySQL para posteriormente ser explotado por Weka (*Waikato Environment for Knowledge Analysis*. Weka es un Entorno para Análisis del Conocimiento de la Universidad de Waikato), es una plataforma de software para aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es un software libre distribuido bajo licencia GNU-GPL.

## 1.3 APORTACIONES DE LA INVESTIGACIÓN

Las aportaciones se derivan de la comparación de las técnicas de Minería de Datos, aplicadas sobre las vistas minables, enriqueciéndose por medio de la visualización de ejemplos, y de los alcances de trabajos similares presentados en el Capítulo 2, “Estado del Arte”.

## 1.4 ORGANIZACIÓN DE LA TESIS

En este trabajo de tesis el contenido será organizado en capítulos, los cuales serán presentados de la siguiente manera y con el orden correspondiente:

- **Capítulo 1. Introducción:** En este capítulo se plantea el problema a resolver, los objetivos que se plantean alcanzar, las razones que justifican el proyecto, así como el planteamiento de la propuesta que dará solución al problema y los resultados obtenidos.

- **Capítulo 2. Estado del Arte:** Se presentan los trabajos que se han realizado en torno al tema del desempleo en México, que tienen como base de desarrollo la creación de *Data Warehouse* y/o explotación mediante técnicas de Minería de Datos.

- **Capítulo 3. Marco teórico:** Este capítulo contiene información sobre los conceptos que soportaron el desarrollo de este proyecto de investigación, así como información básica de las herramientas a utilizar.

- **Capítulo 4. Análisis y diseño:** En este capítulo se muestra el preprocesamiento de los datos, para después hacer un análisis de los datos de Desempleo y niveles educativos.

- **Capítulo 5. Implementación y resultados:** Se presentan los resultados obtenidos, al aplicar Técnicas de Minería de datos sobre las vistas minables obtenidas en el Capítulo 4.

- **Capítulo 6. Conclusiones y trabajo a futuro:** Se exponen las conclusiones sobre la funcionalidad y factibilidad del sistema, así como las aportaciones hechas al área y puntos de vista sobre el posible trabajo futuro que podría mejorar el sistema para hacerlo más robusto.

- Finalmente se presentan las referencias consultadas, para el desarrollo de la tesis.

## 1.5 CONCLUSIONES

El término desempleo indica la falta de trabajo. Un desempleado es aquel sujeto que forma parte de la población activa (se encuentra en edad de trabajar) y que busca empleo sin conseguirlo. Esta situación se traduce en la imposibilidad de trabajar pese a la disponibilidad de la persona. El problema de la falta de empleo ocasiona problemas a nivel socioeconómico.

# CAPÍTULO 2

# ESTADO DEL ARTE

El término desempleo puede tener dos puntos de vista: el sociológico (aquel que se relaciona con el desempleo como fenómeno social) y el psicológico (aquel que se relaciona con la historia o experiencia particular de cada persona). Entendemos por desempleo la situación por la que pasa una persona cuando no tiene un trabajo fijo y, por tanto, no cuenta con los medios para subsistir de manera independiente (es decir, sin la asistencia de sus conocidos o del Estado). El conjunto de las situaciones particulares de desempleo hace que se hable de desempleo como problema sociológico, quizás uno de los problemas más graves que debe enfrentar una sociedad en lo que respecta a su bienestar social**[1]**.

El desempleo puede generarse por varias razones. Cuando entendemos el desempleo como un fenómeno social, las crisis económicas y las medidas de ajuste son por lo general las responsables de situaciones de desempleo que hacen que baje la demanda de trabajadores, que aumente la inseguridad de las inversiones y por tanto haya menos empresas o empleadores disponibles a tomar trabajadores. Al mismo tiempo, podemos señalar que el desempleo se ha vuelto cada vez más un elemento característico de las sociedades modernas en las cuales la tecnología es utilizada en gran parte para reemplazar a la mano de obra humana. Por otro lado, altos valores de desempleo afectan a la economía de una sociedad ya que la misma se puede llegar a ver paralizada en una cantidad importante de sus actividades.

Sin embargo, el desempleo no debe ser entendido nunca a través de la visión limitada de la economía o de la sociología ya que, en el fondo, el desempleo es un problema que afecta gravemente a la realidad de las personas que lo sufren. Por lo general, mantener una situación de desempleo o de inestabilidad laboral puede transformar el carácter, el estado anímico, las capacidades y los intereses de una persona, haciendo que se vuelva una persona falta de esperanzas, depresiva, negativa y mucho más estresada. Esto es así porque normalmente el término desempleo implica involuntariedad en la situación de falta de trabajo.

## 2.1 PROYECTOS Y ESFUERZOS PARA LA MITIGACIÓN DEL DESEMPLEO EN MEXICO

En esta sección mencionaremos algunos trabajos que han realizado las instituciones, gubernamentales nacionales como internacionales, para poder dar una solución o mitigar el problema, ya sea con proyectos que tratan de preverlo o diseñar medidas que se apliquen cuando este se presente. Algunos ejemplos de trabajos se mencionan a continuación en las siguientes secciones.

### 2.1.1 OCDE

Es la Organización para la Cooperación y el Desarrollo Económicos (OCDE) que agrupa a 34 países miembros y su misión es promover políticas que mejoren el bienestar económico y social de las personas alrededor del mundo. Esta ofrece un foro en el cual los gobiernos trabajen conjuntamente compartiendo experiencias y buscando soluciones a problemas comunes. Sus trabajos se ocupan de entender que es lo que conduce al cambio económico, social y ambiental. Mide la productividad y los flujos globales del comercio e inversión. Analiza y compara datos para realizar pronósticos de tendencias. Además fija estándares internacionales dentro de un amplio rango de temas de políticas públicas.

Según estudios de la OCDE (ver figura 2.1), tanto en España, como en la Unión Europea, a mayor nivel de formación corresponde una mayor tasa de ocupación y una menor tasa de desempleo, así como un nivel salarial más elevado.

En el informe español “Panorama de la Educación. Indicadores de la OCDE 2013” **[6]** se describen los beneficios sociales y económicos de la educación relacionados con el mercado de trabajo. Haciendo uso de los datos publicados por el Instituto Nacional de Estadística (INE) del año 2011 la OCDE obtuvo que:

- Las tasas de desempleo más elevadas se registran en individuos cuya educación es solo básica.
- Las tasas de empleo más elevadas las registran en personas que han finalizado la enseñanza secundaria superior y se han formado en ciencias: mecánica y electrónica, industria manufacturera y construcción, en agricultura, salud y servicios sociales.

Los temas que aborda la OCDE son:

- Administración Pública
- Agricultura y Alimentación
- Asuntos Sociales, Migración y Salud
- Ciencia y Tecnología
- Comercio
- Desarrollo Urbano, Rural y Regional
- Economía
- Educación
- Empleo
- Energía
- Energía Nuclear
- Finanzas e Inversión
- Impuestos
- Industria y Servicios
- Medio Ambiente
- Transporte


*Figura 2. 1 Logotipo de OCDE*

### 2.1.2 INDICADORES OPORTUNOS DE OCUPACIÓN Y EMPLEO

El **Instituto Nacional de Estadística y Geografía** (**INEGI**) es un organismo autónomo del gobierno mexicano, dedicado a la coordinación del Sistema Nacional de Información Estadística y Geográfica del país. Es la institución encargada de realizar los censos de población cada diez años, así como los censos económicos cada cinco años y los censos agropecuarios del país.

El trabajo de compendio de información estadística por parte del Instituto incluye producto nacional mensual, encuestas de confianza de los consumidores y muestras de proporción de comercios. Por otra parte se muestran estadísticas de ocupación y empleo, educación, de violencia intrafamiliar y de pareja; así como muchos trabajos más que dan fundamento a los estudios y proyecciones de diversas instituciones gubernamentales, recientemente el instituto está haciendo una encuesta a las escuelas. La sede del Instituto está ubicada en la ciudad de Aguascalientes, en el Estado de Aguascalientes, México.

EL boletín de prensa que lleva por nombre "Indicadores Oportunos de Ocupación y Empleo" es un trabajo descriptivo realizado en el año 2012 por el INEGI (figura 2.2) el cual informa sobre los principales resultados de la Encuesta Nacional de Ocupación y Empleo (ENOE) del año 2011, la cual informa los porcentajes de población económicamente activa la cual es considerada a partir de los 14 años en adelante **[7].** Se realiza con una comparación de un mismo estado con respecto al tiempo para observar los cambios.


*Figura 2. 2 Logotipo de INEGI*

### 2.1.3 CEPAL

La **Comisión Económica para América Latina y el Caribe** (**CEPAL**) es el asociación dependiente de la Organización de las Naciones Unidas (ONU) responsable de promover el desarrollo económico y social. Sus labores se concentran en el campo de la investigación económica [8].

La sede de la Comisión se encuentra en Santiago de Chile, la cual coordina dos sedes subregionales: una para América Central, con sede en la Ciudad de y otra para los países del Caribe, situada en Puerto España. Tiene oficinas nacionales en Bogotá, Montevideo, Brasilia, Buenos Aires, y una oficina de enlace en la ciudad de Washington D.C.

En la revista CEPAL 107 publicada en agosto de 2012 se presenta el trabajo de investigación llamado "México: ¿Cómo inciden las políticas monetarias en las tasas de desempleo?" (Escrito por: Alejandro Islas C. y Willy Walter Cortez [8]) el cual se adentra en la investigación del impacto de la política monetaria en las tasas de desempleo de México tomando un marco de tiempo que abarca del año 1988 al año 2004.

Se realizó un análisis exhaustivo con dos series de tasas de desempleo, donde la primera es la tasa oficial que maneja el gobierno mexicano y la segunda que es una tasa alternativa calculada según la metodología utilizada por la Encuesta Continua de Población de la Oficina de Estadísticas Laborales de los estados Unidos de América. Los resultados arrojados haciendo uso de la tasa oficial son que al principio del periodo de tiempo hay un aumento rápido de desempleo, después de un tiempo este se normaliza y en cierto punto este regresa a la línea de salida lo que indicaría que el mercado laboral mexicano es muy fluido. Cuando se usa la tasa alternativa se observa que la política monetaria no es neutral en el corto y largo plazo.

Unas de las funciones que se evalúan son la de impulso-respuesta y la de perturbación inicial con las cuales se observa que no existe un efecto a largo plazo de la política monetaria en el desempleo.

Finalmente los modelos basados en la medida alternativa de desempleo utilizados en el trabajo sugieren que la política monetaria puede explicar hasta un 27% de la varianza del desempleo y que con medida oficial solo es posible explicar un 2,2%.


*Figura 2. 3 Logotipo de CEPAL*

## 2.2 COMPARATIVA DE LOS TRABAJOS.

Entre los proyectos hay diferencias y similitudes, una similitud es que estos estudios hacen uso de datos recabados a lo largo de la historia del cada país. Una diferencia primordial es que el primero es de un enfoque comparativo con datos de otros países, el segundo se enfoca a datos comparativos de un solo país aunque si hace distinción en la información de los estados que lo conforman y el tercero se centra en la comparativa de valores oficiales y alternativos que son obtenidos con el uso de métodos específicos.

El proyecto descrito en la sección 2.1.1 maneja técnicas computacionales propias de la minería de datos, no se enfoca únicamente al desempleo, sino que buscan diferentes temas que también son descritos en la sección.

El proyecto de la sección 2.1.2 es más un artículo que una explicación completa del tema y solo abarca un periodo de tiempo relativamente corto y finalmente la sección 2.1.3. Este se refiere a un trabajo realizado con el propósito de entender si la política monetaria puede explicar la tasa de desempleo a lo largo de un periodo de tiempo extenso.

## 2.3 DESEMPLEO Y LA TOMA DE DECISIONES.

Los sistemas de toma de decisión para los sectores públicos y privados de países en vías de desarrollo, típicamente enfrentan la presión de actuar en respuesta a problemas que requieren acción inmediata. Además, el efecto de tales decisiones debe ser evidente durante plazos, usualmente cortos, en los cuales esos ejecutivos operan. Consecuentemente, dan prioridad relativamente baja a los asuntos que se perciben como problemas de un futuro distante o simplemente se pretende minimizar, tal es el caso del “Desempleo”. La Organización para la Cooperación y el Desarrollo Económicos impulsada por su misión trata de promover políticas que mejoren el bienestar económico y social de las personas alrededor del mundo.

## 2.4 CONCLUSIONES

El desempleo es un problema social que se ha minimizado y para cual no se han puesto en práctica estrategias lo suficientemente efectivas o que al menos tratasen de contrarrestar los efectos perjudiciales que ocasiona el desempleo.

Es necesario tomar en cuenta las experiencias de otros países para poder desarrollar proyecto que realmente ayuden a mitigar los problemas que acarrea el desempleo.

# CAPÍTULO 3

# MARCO TEÓRICO

En este capítulo se presentan los fundamentos teóricos que ofrecen el sustento formal al desarrollo de la tesis. Exponiendo como primer punto la tecnología de los DataWarehouse (DW) como una herramienta para analizar la información y como segundo punto, la Minería de Datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados. Con respecto al segundo punto que involucra la Minería de Datos, cabe mencionar que es la parte más extensa de este trabajo de tesis y la más técnica, ya que se trata de describir el funcionamiento y la conveniencia de la aplicación de las técnicas de Minería de Datos.

## 3.1 ALMACENES DE DATOS

El aumento de la cantidad y variedad de información que se encuentra computarizada en bases de datos digitales y otras fuentes ha crecido exponencialmente en las últimas décadas. La mayor parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Además de su función de "memoria de la organización", la información histórica es útil para explicar el pasado, entender el presente y predecir la información futura. La mayoría de las decisiones de empresas, organizaciones e instituciones se basan en información sobre experiencias pasadas extraídas de fuentes muy diversas **[2]**.

Por lo tanto, estas se preocuparon por unificar las diversas fuentes de información de las cuales disponían, en un único lugar, al que sólo se le agregaría información importante, sobre la base de una estructura organizada, integrada, lógica, dinámica y de fácil explotación. La respuesta a esta problemática fueron los Almacenes de Datos.

Existen muchas definiciones para el Almacén de Datos, la más conocida fue propuesta por William Inmon- considerado el padre del *DataWarehouse* en 1992 **[9]**:

*"Un DW es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales". William Inmon indicó que un DataWarehouse se caracterizaba por ser:*

**Integrado:** Los datos almacenados en el *DataWarehouse* deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

**Temático:** Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

**Histórico:** El tiempo es parte implícita de la información contenida en un *DataWarehouse.* En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el *DataWarehouse* sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el *DataWarehouse* se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

**No volátil:** La información es útil sólo cuando es estable. La información no se modifica ni se elimina, una vez almacenado el dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas. Es lógico debido a que el designio del *DataWarehouse* es ser capaz de analizar lo que ya ha sucedido **[10]**.

### 3.1.1 OLAP y OLTP

Existen dos tipos de procesamientos orientados al análisis de datos, los cuales corresponden a las siglas OLTP y OLAP los cuales se detallan a continuación:

- **OLTP** (*On-Line Transactional Processing*). Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones en tiempo real, constituye el trabajo primario en un sistema de información. Este trabajo consiste en realizar transacciones, es decir, actualizaciones y consultas a la base de datos con un objetivo operacional.

- **OLAP** (*On-Line Analytical Processing*). Son bases de datos orientadas al procesamiento analítico en tiempo real, que engloba un conjunto de operaciones, exclusivamente de consulta, en las que se requiere agregar y cruzar gran cantidad de información. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil. Ejemplos de este tipo de trabajo analítico pueden ser: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos entre los más significativos.

Una característica de ambos procesamientos es que se plantea que sean "*on-line*", es decir, que sean relativamente "instantáneos" y se puedan realizar en cualquier momento (en tiempo real). Esto parece evidente e imprescindible para el OLTP, pero no está tan claro que esto sea posible para algunas consultas muy complejas realizadas por el OLAP.

Hasta hace pocos años, y todavía existente en muchas organizaciones y empresas, ambos tipos de procesamiento (OLTP y OLAP) se realizaran sobre la misma base de datos transaccional, con lo cual se generan dos problemas fundamentales:

- Las consultas OLAP perturban el trabajo transaccional diario de los sistemas de información originales. Al ser consultas complejas y que involucran muchas tablas y agrupaciones, suelen consumir gran parte de los recursos del sistema de gestión de base de datos. El resultado es que durante la ejecución de estas consultas, las operaciones OLTP, se resienten: las aplicaciones van más lentas, las actualizaciones consumen mucho tiempo y el sistema puede incluso llegar a colapsarse. De este hecho viene el nombre familiar que se les da a las consultas OLAP: *"killer queries"* (consultas asesinas). Como consecuencia, muchas de estas consultas se deben realizar por la noche o en fines de semana, con lo que en realidad dejan de ser "*on-line"*.

- La base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos. Esto sugiere que, aun cuando tuviéramos el sistema exclusivamente para realizar una consulta OLAP, dicha consulta puede requerir demasiado tiempo, pero no solo por ser compleja intrínsecamente, sino porque el esquema de la base de datos no es el más adecuado para este tipo de consultas.

Ambos problemas indican que será prácticamente imposible realizar un análisis complejo de la información en tiempo real si ambos procesamientos se realizan sobre una misma base de datos. Desde esta perspectiva, se separa definitivamente la base de datos con fines transaccionales de la base de datos con fines analíticos y es así que nacen los almacenes de datos **[2]**.

### 3.1.2 ALMACENES DE DATOS Y BASES DE DATOS TRANSACCIONALES

Tradicionalmente el análisis para la toma de decisiones se realizaba sobre las mismas bases de datos de trabajo o base de datos transaccionales. Esto implicaba combinar el trabajo transaccional diario de los sistemas de información originales (OLTP), con el análisis de datos en tiempo real sobre la misma base de datos (OLAP), esto provoca problemas con que:

- Disturba el trabajo transaccional diario de los sistemas de información originales.
- Se realizan consultas muy pesadas (killer queries).
- En situaciones de carga alta, la perturbación es tal que el proceso analítico se debe realizar por la noche o en periodos festivos.

Todos estos problemas son provocados porque la base de datos está diseñada para el trabajo transaccional y no para el análisis de datos, por lo que el análisis es lento.

La ventaja fundamental de un almacén de datos es su diseño específico y su separación de la base de datos transaccional. Un almacén de datos:

- Facilita el análisis de los datos en tiempo real (OLAP).
- No disturba el OLTP de las bases de datos originales.

Por lo anterior no debemos confundirlas y debemos diferenciar claramente entre las bases de datos transaccionales y los almacenes de datos. Las diferencias entre ambas son las que se muestran a continuación en la tabla 3.1.

*Tabla 3. 1 Diferencias entre una base de datos transaccional y un almacén de datos.*

| | **BASE DE DATOS TRANSACCIONAL** | **ALMACÉN DE DATOS** |
|---|---|---|
| **Propósito** | Operaciones diarias. Soporte a las aplicaciones. | Recuperación de información, informes, análisis y minería de datos. |
| **Tipo de datos** | Datos de funcionamiento de la organización. | Datos útiles para el análisis, la sumarización, etc. |
| **Características de los datos** | Datos de funcionamiento, cambiantes, internos, incompletos... | Datos históricos, datos internos y externos, datos descriptivos... |
| **Modelos de datos** | Datos normalizados. | Esquema estrella, copo de nieve, parcialmente desnormalizados, multidimensionales,… |
| **Número y tipo de usuarios** | Cientos/miles: aplicaciones, operarios, administrador de la base de datos. | Decenas: directores, ejecutivos, analistas (granjeros, mineros). |
| **Acceso** | SQL. Lectura y escritura. | SQL y herramientas propias *(slice & dice, drill, roll, pivot...).* Lectura. |

## 3.2 RELACIÓN DE LA MINERIA DE DATOS CON OTRAS DISCIPLINAS.

La Minería de Datos es un campo multidimensional que se ha desarrollado en paralelo o como prolongación de otras disciplinas. Por ello, la investigación y los avances en la Minería de Datos se nutren de los avances que se generan con las que tiene.

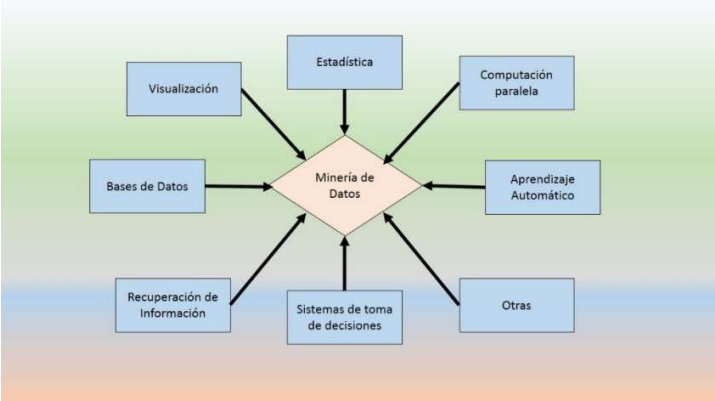



*Figura 3. 1 Disciplinas que se relacionan con la Minería de Datos.*

Algunas de las disciplinas que se muestran en la Figura 3.1 se describen a continuación:

- Las bases de datos: los almacenes de datos y el procesamiento analítico en línea (OLAP) están relacionados ampliamente con la Minería de Datos, donde el procesamiento OLAP, no se trata únicamente de obtener informes avanzados, como se incluyen en muchas herramientas de *Bussiness Intelligence*, sino que se refieren a la extracción de conocimiento novedoso, útil y comprensible.

- La recuperación de información: consiste en la obtención de información relevante desde los datos textuales, donde una tarea típica es encontrar documentos a partir de palabras claves, lo cual puede verse como un proceso de clasificación de los documentos en función de estas palabras clave.

- La estadística: en la Minería de Datos se utilizan mucho los conceptos, algoritmos y técnicas de esta disciplina. Por mencionar algunos tenemos, la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, etcétera.

- El aprendizaje automático: es una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender, por medio de algoritmos y programas; constituye junto con la estadística, el corazón del análisis inteligente de los datos.

- Los sistemas para la toma de decisión: (DSS por sus siglas en inglés *Decision Support System*) puede adoptar muchas formas diferentes, y en general es un sistema informático utilizado para servir de apoyo, más que automatizar, el proceso de la toma de decisiones.

- La visualización de los datos: en general son una herramienta para el análisis de los datos. Estas herramientas permiten al usuario descubrir, intuir o entender patrones que serían más difíciles de "ver" a partir de descripciones matemáticas o textuales de los resultados **[5].**

## 3.3 METODOLOGÍA KDD

Se refiere al Descubrimiento del Conocimiento, es un campo de rápido crecimiento en la investigación, debido a la demanda creciente de herramientas que ayudan a revelar y comprender información oculta en enormes cantidades de datos esta metodología se ha hecho realmente popular. Los datos se generan diariamente ya sea por las agencias federales, los bancos, las empresas de seguros, tiendas minoristas y en la WWW. Esto genera una explosión de datos que se produce a través del uso creciente de computadoras, escáneres, cámaras digitales, códigos de barras entre otras.

Actualmente se cuenta con una gran cantidad de información, almacenada en base de datos, hojas de cálculo y otros repositorios de datos, que están disponibles pero no fácilmente analizables. Lo que se necesita es una metodología clara y sencilla para extraer el conocimiento oculto en los datos.

En el siguiente apartado se explica el modelo de KDD, mostraremos los seis pasos que conlleva este modelo, para permitir el fortalecimiento del conocimiento y entablar una comunicación entre diversas herramientas, como son Minería de datos, bases de datos y algunos otros repositorios de información.

### 3.3.1 LOS SEIS PASOS DEL KDD

Se denomina KDD al descubrimiento de conocimientos en una base de datos, esto claro al seguir una serie de pasos concretamente. Este modelo debe ayudar a planear y reducir el costo al detallar procedimientos que se realizan en cada uno de los pasos **[2]**. Los seis pasos del modelo se describen a continuación:

**1. Comprendiendo el dominio del problema.** En este paso se trabaja con el apoyo de expertos del área del tema para identificar el problema y delimitar los objetivos que se pretenden alcanzar con el proyecto, así como aprende sobre soluciones y temas actuales que tengan relación con el problema.

En resumen este punto se centra en aprender la terminología específica del tema y a la vez crear una descripción del problema, incluyendo sus restricciones, para que este sea llevado a cabo.

**2. Comprensión de los datos.** En este paso comprende la recolección de datos, descartar los datos que no nos son útiles y decidir cuáles serán el formato y tamaño de los datos que nos son útiles. Si el conocimiento existe, podremos clasificar algunos atributos como más importantes. Después de definir nuestros datos útiles debemos verificar la utilidad de los datos en relación con los objetivos del KDD. Los datos deben ser verificados en cuanto a integridad, redundancia, falta de valores, la plausibilidad del valor de los atributos y cuestiones similares.

**3. Preparación de los datos.** Fallar en este paso conllevará a que el proyecto no funcione, dado que el descubrimiento del conocimiento depende de este proceso además que este paso es el más extenuante ya que comprende aproximadamente la mitad del esfuerzo del proyecto.

**4. Minería de Datos.** Esta es la parte en la cual se descubre el conocimiento. Mediante el uso de herramientas de Minería de Datos se puede descubrir nueva información, además este proceso dura mucho menos que la preparación de datos. Este paso no solo implica el uso de herramientas de MD, sino que puede complementarse de otras si es necesario. Las herramientas de MD comprenden muchos tipos de algoritmos como por ejemplo: métodos bayesianos, computación evolutiva, aprendizaje automático, redes neuronales, Clustering y técnicas de preprocesamiento.

La principal dificultad en este paso es que muchas de las herramientas no pueden ser aplicadas a un enorme volumen de datos. Las herramientas están caracterizadas por un aumento lineal en el tiempo de ejecución, dentro de una cantidad fija de memoria disponible. La mayoría de las herramientas de Minería de Datos, no son escalables, pero hay ejemplos de herramientas de Minería que sí lo son, como por ejemplo el Clustering, el aprendizaje automático y las reglas de asociación.

**5. Evaluación del Conocimiento Descubierto.** En este paso los expertos analizan los resultados obtenidos, para verificar que verdaderamente sean resultados novedosos, útiles e interesantes. En todo proceso KDD se pueden volver a realizar análisis de los datos, para así poder identificar alternativas y poder mejorar los resultados.

**6. Uso del conocimiento descubierto.** En este paso intervienen los propietarios de las bases de datos, ya que les toca a ellos, llevar a cabo la aplicación del conocimiento descubierto, para ello se requiere hacer una planeación de donde y como se aplicara. El área de aplicación no solo comprende un dominio sino que deberá ampliarse a otros que sean parte de la organización. Se debe crear un plan para vigilar el conocimiento descubierto y todo el proyecto debe ser documentado.

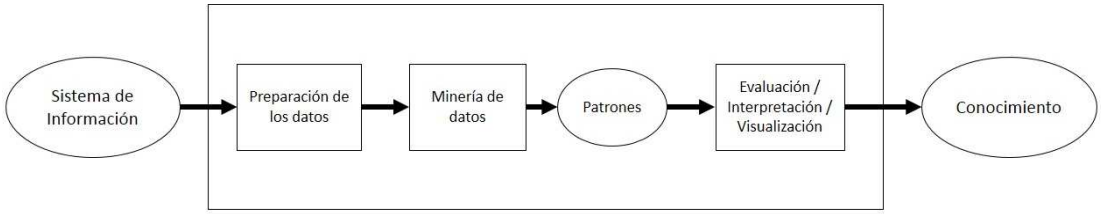



*Figura 3. 2 Procesos de extracción de conocimiento*

Teniendo en cuenta el proceso de extracción de conocimiento (figura 3.2) podemos notar, la estrecha relación que existe entre KDD y la minería de datos. La MD se refiere a la aplicación de métodos de aprendizajes y estadísticos para la obtención de patrones y modelos, mientras que el KDD es el proceso para la extracción de conocimiento desde las bases de datos.

## 3.4 TAREA Y TÉCNICAS DE MINERÍA DE DATOS

Una tarea de Minería de datos es un tipo de problema de Minería de Datos. Por ejemplo, “clasificar las piezas de una empresa de autopartes en: óptimas, defectuosas, reparables y defectuosas irreparables” es una tarea. Concretamente el tipo de tarea es de clasificación. Está tarea se podría resolver mediante árboles de decisión o redes neuronales, entre otros métodos.

Dentro de las tareas de Minería de Datos existen tipos, cada una de las cuales se puede considerar como un ejemplo de problema a ser resuelto por un algoritmo de Minería de Datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra **[2]**.

Como hemos mencionado anteriormente, las tareas de minería de datos se pueden clasificar como predictivas o descriptivas. Entre las tareas predictivas encontramos la clasificación y la regresión, mientras que el agrupamiento (*clustering*), las reglas de asociación secuenciales y las correlaciónales son tareas descriptivas.

### 3.4.1 TAREAS PREDICTIVAS

Este tipo de tareas se conoce también como "aprendizaje supervisado", son problemas y técnicas en los que hay que predecir uno o más valores para uno o más ejemplos. La minería de datos puede mostrar cómo se comportaran ciertos atributos en el futuro de cierto conjunto de datos. Entre los ejemplos de predicciones de datos podemos incluir el análisis de transacciones de compra para predecir que compraran los consumidores ante determinados descuentos, que volumen de ventas se generará en una tienda en un periodo determinado y si la eliminación de una línea de productos generará más beneficios. En aplicaciones de ese tipo, la lógica de negocio se utiliza unida a la minería de datos. En un contexto científico, determinados patrones de ondas sísmicas podrían predecir un terremoto con una probabilidad alta **[11].**

No obstante una vez más podemos analizar estos problemas con más detenimiento para observar que la variable a predecir puede ser una variable categórica (por ejemplo si compra o no un producto). Esta distinción nos lleva a que podemos ver este tipo de problemas en dos que son:

- Problemas de clasificación: hacen referencia a los problemas en los que la variable a predecir tiene un número finito de valores, esto es la variable es categórica.

- Problemas de predicción de valores: se refieren a los problemas en los que la variable a predecir es numérica.

Realizar estas distinciones es necesario, ya que de acuerdo a la problemática, se debe buscar la técnica adecuada que se utilizará para solucionar el problema **[12].**

### 3.4.2 TAREAS DESCRIPTIVAS

Estas técnicas o métodos son también llamados “no supervisados”, utilizados frecuentemente cuando una aplicación no está lo suficientemente preparada y no tiene el potencial necesario para una solución predictiva, para descubrir patrones y tendencias en los datos, que permitan explorar las propiedades de los datos examinados.

La descripción es normalmente usada para realizar un análisis preliminar de los datos. Busca derivar descripciones concisas de características de los datos: medias, desviaciones estándares, etc.

La meta principal de todas estas tareas es una descripción del conjunto de datos origen. Pero las tareas descriptivas las podemos dividir en dos que son:

- **Análisis de segmentación**: que se refiere a los problemas donde la meta es encontrar grupos homogéneos en la población de objetos origen. A estos problemas también se les denomina problemas de aprendizaje no supervisado o *clustering* **[12]**.

- **Análisis de asociaciones**: hace referencia a los problemas en los que se persigue obtener relaciones entre los valores de atributos de una base de datos **[12]**.

### 3.4.2 TAREAS DESCRIPTIVAS

A continuación (Tabla 3.2) se muestran de manera ilustrativa, algunas tareas (clasificación, regresión, agrupamiento, reglas de asociación, correlaciones/factorizaciones) y algunas técnicas o algoritmos.

*Tabla 3. 2 Aplicaciones de las técnicas de minería de datos [1]*

| NOMBRE | PREDICTIVO | | DESCRIPTIVO | | |
|---|---|---|---|---|---|
| | Clasificación | Regresión | Agrupamiento | Reglas de Asociación | Correlaciones y Factorizaciones |
| Redes Neuronales | X | X | X | | |
| Árboles de decisión: ID3, C4.5, C5.0 | X | | | | |
| Árboles de decisión CART | X | X | | | |
| Otros árboles de decisión | X | X | X | X | |
| Redes de Kohonen | | | X | | |
| Regresión lineal y logarítmica | | X | | | X |
| Regresión logística | X | | | X | |
| Kmeans | | | X | | |
| A priori | | | | X | |
| Naïve Bayes | X | | | | |
| Vecinos más próximos | X | X | X | | |
| Análisis factorial y de Componentes principales. | | | | | X |
| Twostep, Cobweb | | | X | | |
| Algoritmos genéticos y evolutivos | X | X | X | X | X |
| Máquinas de vectores soporte | X | X | X | | |
| CN2 rules (cobertura) | X | | | X | |
| Análisis discriminante multivariante | X | | | | |

## 3.5 SOFTWARE PARA DATA MINING

Actualmente podemos encontrar tanto en el ámbito comercial como en el académico una serie de entornos software diseñado para dar soporte al ejercicio de la minería de datos. La mayoría de estos sistemas de software integran en un mismo entorno capacidades para el procesado de datos, diferentes modelos de análisis, facilidades para el diseño de experimentos y soporte gráfico para la visualización de resultados. Su manejabilidad no se haya condicionada a que el usuario posea conocimientos de programación, ya que existe una interfaz que facilita la interacción entre el usurario y la herramienta.

Entre los sistemas más utilizados se pueden nombrar: SPSS, *Clementine*, WEKA, *Kepler*, DMS, *DBMiner*, YALE, *DB2 Intelligent Miner*, *SAS Enterprise Miner*, *STATISTICA* y *Data Miner*, entre otros.

### 3.5.1 WEKA (Waikato Environment for Knowledge Analysis)

Es un entorno para experimentación de análisis de datos, desarrollada por un equipo de investigadores de la universidad de Waikato (Nueva Zelanda), que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Para ello se requiere que los datos a analizar se almacenen en un formato conocido como ARFF (Attribute Relation File Format) o en algún otro formato de texto permitido.

WEKA es un software de libre distribución, bajo licencia GNU, desarrollado en Java. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos.

WEKA se caracteriza por:

- **Acceso a los datos**: Los datos son cargados desde un archivo ARFF (archivo plano organizado en filas y columnas). El usuario puede observar en los diferentes componentes gráficos, información de interés sobre el conjunto de muestras (tamaño del conjunto en número de registros, número de atributos, tipo de datos, medias y varianzas de los atributos numéricos, distribución de frecuencias en los atributos nominales, etc.)

- **Preprocesamiento de datos:** selección de atributos, discretización, tratamiento de valores desconocidos y transformación de datos numéricos.

- **Modelos de aprendizaje:** árboles de decisión (J4.8, versión propia del algoritmo C4.5), tablas de decisión, vecinos más próximos, máquinas de vectores de soporte, reglas de asociación, métodos de agrupamiento y modelos combinados.

- **Visualización:** la interfaz gráfica se compone de diversos entornos. El entorno *Explorer* permite controlar todas las operaciones anteriores. El entorno consola (CLI) posibilita la invocación textual de las operaciones anteriores. El entorno *Experimenter* facilita el diseño y la realización de experimentos complejos. El proceso global de minería de datos en WEKA se acelera considerablemente gracias al entorno *KnowledgeFlow* que, de una forma gráfica y a modo de flujos de operaciones, permite definir la totalidad del proceso.


*Figura 3. 3 Interfaz GUI de WEKA*

Como podemos observar en la Figura 3.3 la interfaz de inicio de WEKA está compuesta por diferentes entornos:

- El entorno *Explorer* permite controlar todas las operaciones anteriores (filtrado, selección y especificación del modelo, diseño de experimentos, etc.).

- El entorno *Experimenter* facilita el diseño y la realización de experimentos complejos

- El proceso global de Minería de datos en WEKA se acelera considerablemente gracias al entorno *KnowledgeFlow* que, de una forma gráfica y a modo de flujos de operaciones, permite definir la totalidad del proceso (carga de datos, preprocesamiento, obtención de modelos, comprobación y visualización de resultados).

- El entorno consola (CLI) posibilita la invocación textual de las operaciones anteriores. (También es posible acceder directamente a los métodos que implementan dichas tareas e incorporarlos en el código fuente de la aplicación de Minería de Datos que se esté programando.)

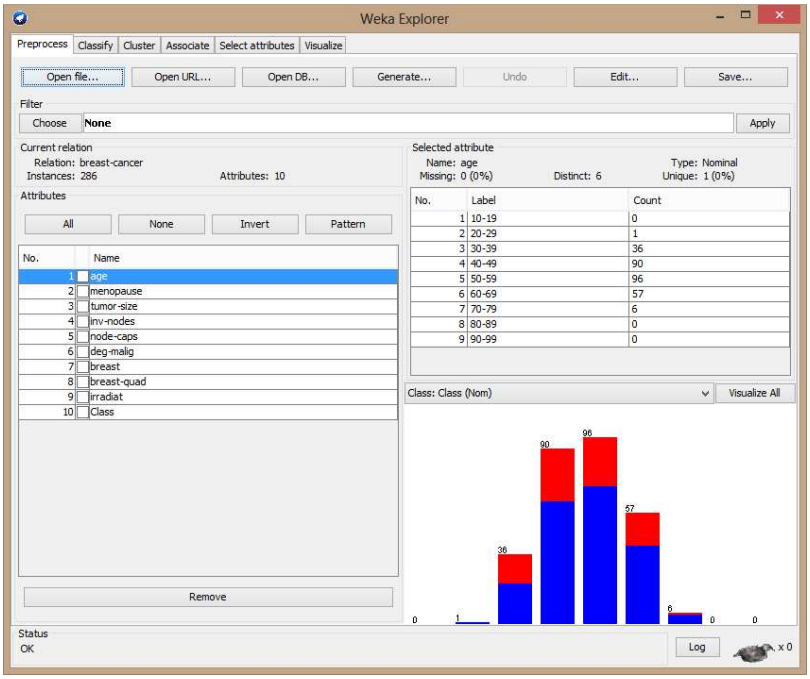



*Figura 3. 4 Interfaz del entorno EXPLORER*

La figura 3.4 nos permite apreciar el entorno de trabajo de la herramienta WEKA. Dese aquí podemos seleccionar los datos que usaremos para la extracción de conocimientos, además que nos da la opción de aplicar filtros necesarios para poder generar la vista minable con la cual trabajaremos.

## 3.6 CONCLUSIONES

El marco teórico antes descrito, amplia los conocimientos acerca de DataWarehouse, Minería de datos y de las herramientas que son utilizadas para el desarrollo de esta tesis que tiene como fin resolver el problema plateado en la sección 1.1.1. Con el marco teórico se tiene un panorama de los pasos a seguir para el desarrollo de la tesis.

# CAPÍTULO 4

# ANÁLISIS Y DISEÑO

En el presente capítulo se describe la metodología a utilizar para el diseño e implementación del almacén de datos para el análisis de factores educativos que influyen en el desempleo en México. También se detallan las técnicas de Minería de Datos que serán aplicadas a dicho Sistema.

## 4.1 PLANTEAMIENTO Y REQUERIMIENTOS

En base al problema plateado en la sección 1.1.1 y considerando los objetivos que se desean alcanzar, mismos que fueron especificados en la sección 1.1.2, se determina que la metodología a seguir es el proceso KDD, el cual fue planteado en la sección 3.3.1 como se muestra a continuación:

1. **Preparación de los datos:**
   - Integrar y recopilar los datos de Bases de Datos y otras fuentes muy diversas tanto internas como externas, para la construcción del *DataWarehouse.*
   - Desarrollar un proceso de selección, limpieza y transformación de la información, y de esta forma eliminar y corregir aquellos datos que sean incorrectos, además esto permitirá decidir la estrategia a seguir con los datos incompletos. En este paso consideraremos únicamente aquellas variables o atributos que son relevantes, con la finalidad de crear una vista minable y que los resultados de la misma sean útiles.

2. **Aplicación de las técnicas de Minería de Datos**
   - En esta fase se decidirá cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elegirá el método a utilizar.

3. **Evaluación e interpretación**
   - En esta fase los resultados obtenidos se evalúan y analizan por expertos, para verificar si el conocimiento es válido, si es útil, de ser necesario se realiza el proceso desde cualquiera de las fases anteriores.

4. **Difusión y uso**
    - Se hará uso del nuevo conocimiento y será distribuido a todos los posibles usuarios.

Para ilustrar las fases ya explicadas se muestra la figura 4.1.

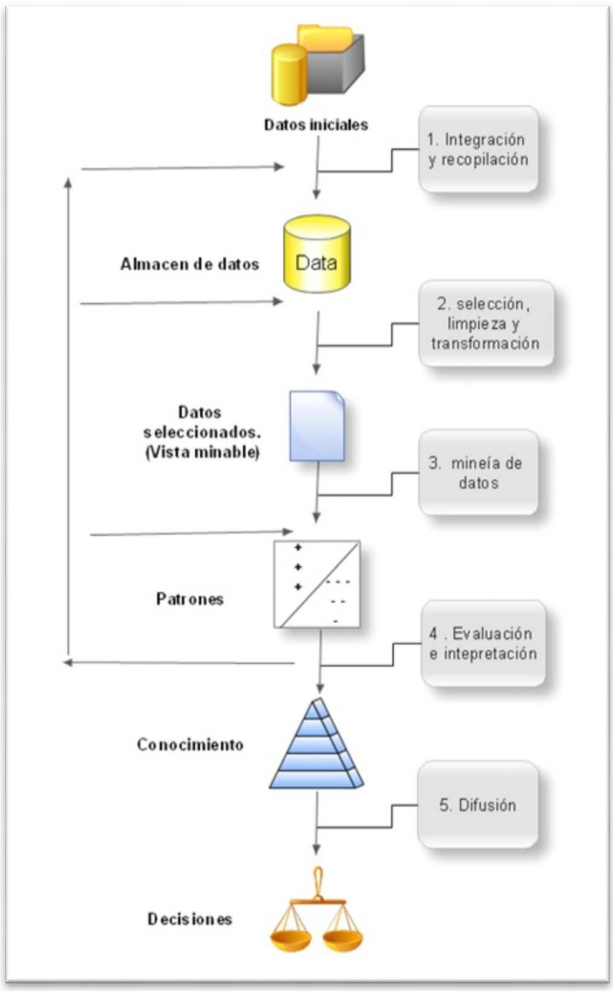



*Figura 4. 1 Fases del proceso de descubrimiento en bases de datos, KDD*

## 4.2 FASE DE INTEGRACIÓN Y RECOPILACIÓN

En esta sección se describe la integración de los datos, obtenidos por el Instituto Nacional de Estadística Geográfica e Informática (INEGI). Dichos datos comprenden el periodo de tiempo de 2005-2011, los atributos correspondientes a los datos de desempleo se describen en la Tabla 4.2 y los datos referentes a niveles educativos se describen en la Tabla 4.1. Cabe mencionar que ambas tablas estaban almacenadas en hojas de cálculo.

*Tabla 4. 1 Atributos de los datos de educación obtenidos por INEGI*

| ATRIBUTOS |
|---|
| Año |
| Nombre del Estado |
| Estudiantes activos en bachillerato del sistema abierto |
| Egresados en bachillerato del sistema abierto |
| Secundaria Existencias |
| Secundaria Aprobados |
| Secundaria Egresados |
| Profesional técnico Existencias |
| Profesional técnico Aprobados |
| Profesional técnico Egresados |
| Bachillerato Existencias |
| Bachillerato Aprobados |
| Bachillerato Egresados |
| Profesionales |
| PIB |

*Tabla 4. 2 Atributos de los datos de desempleo obtenidos por INEGI*

| ATRIBUTOS |
|---|
| Año |
| Estado |
| Personas Ocupada |
| Personas Desocupada |
| Ocupada Hombres |
| Ocupada Mujeres |
| Desocupada Hombres |
| Desocupada Mujeres |

## 4.3 FASE DE SELECCIÓN Y LIMPIEZA Y TRANSFORMACIÓN

La calidad del conocimiento obtenido depende en gran parte de la calidad de los datos minados, es por ello, que después de haber seleccionado los datos, el siguiente paso en la metodología KDD es preparar el subconjunto de datos que se va a minar, los cuales van a constituir lo que se conoce como vista minable.

En el transcurso de este proceso se eliminaron valores *outliers* (valores que no se ajustaron al comportamiento general de los datos). Posteriormente se derivaron algunos atributos que eran relevantes, se rellenaron algunos valores faltantes así como se cambiaron los datos de tipo numérico a su representación porcentual.

Todas las acciones descritas anteriormente se efectuaron con la finalidad de mejorar la eficiencia de la herramienta de minería de datos, y de esta manera mejorar la calidad del conocimiento obtenido. Existen maneras diversas para llevar a cabo este proceso. Algunas de ellas son hacerlo manualmente registro a registro y también es posible hacer uso de herramientas automatizadas.

En la figura 4.2 se muestra el histograma del parámetro de “Personas desocupadas” con respecto al estado correspondiente que comprende de los años 2005 al 2011 y se puede apreciar que los valores van desde 4670 a 433296. Esto ocasionaría problemas al tratar de aplicar las técnicas de minería de datos ya que aun cuando son valores del mismo tipo sus rangos son muy variados. Por ejemplo, no es posible tratar de comparar de manera absoluta la cantidad de desempleados que tiene el estado de México con el estado de Zacatecas, ya que en este último los desempleados no superan los 40 mil, mientras que en el estado de México se cuentan 245 mil personas en esta categoría.

La figura 4.2 nos muestra claramente la diferencia que tiene el estado de México con el resto de los estados. Por lo tanto, el estado en el que se presentaban los datos no era el que necesitábamos para poder aplicar minería de datos, por lo cual se procedió a su modificación. Este proceso se explica más detalladamente en la siguiente sección.

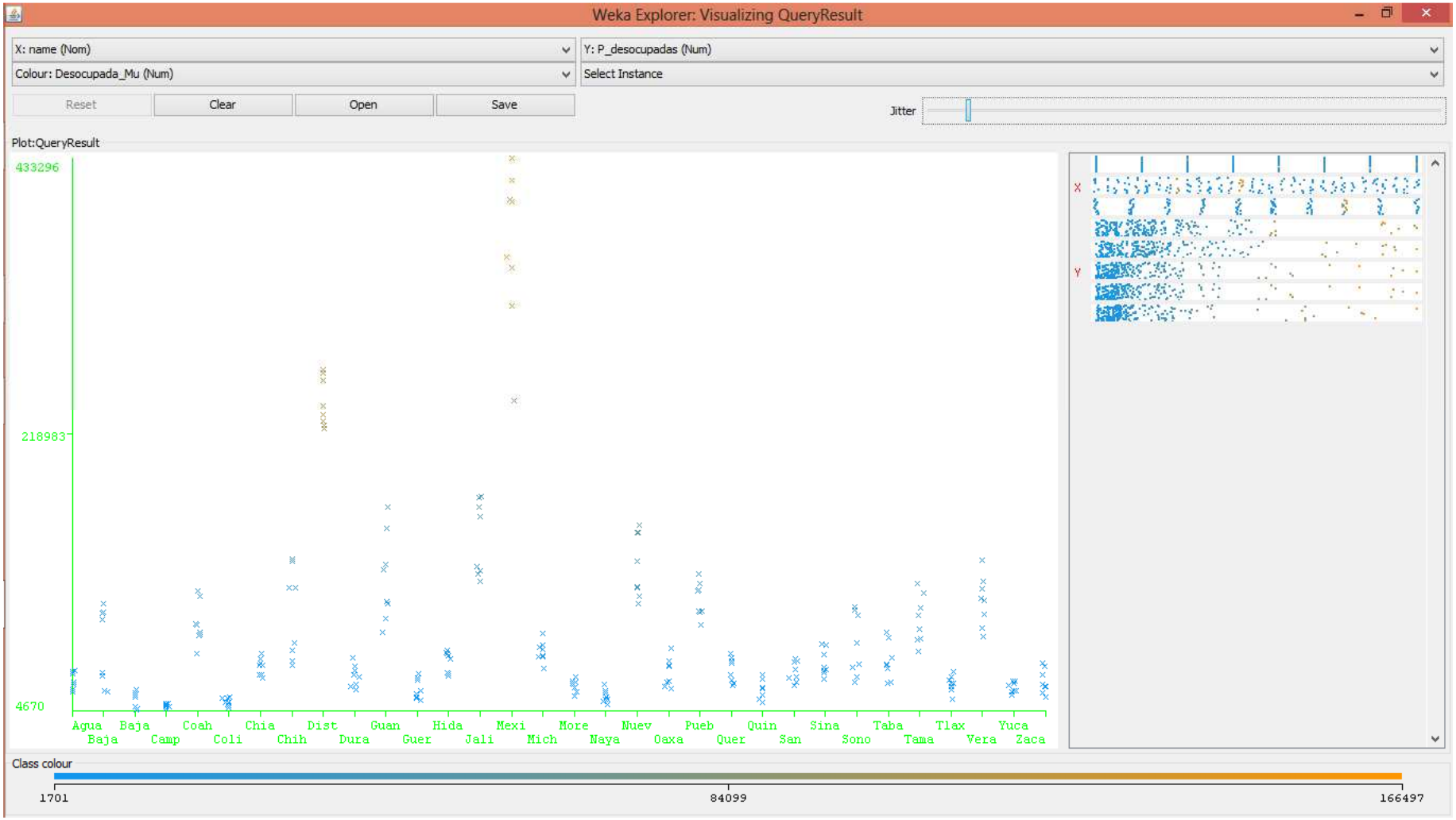
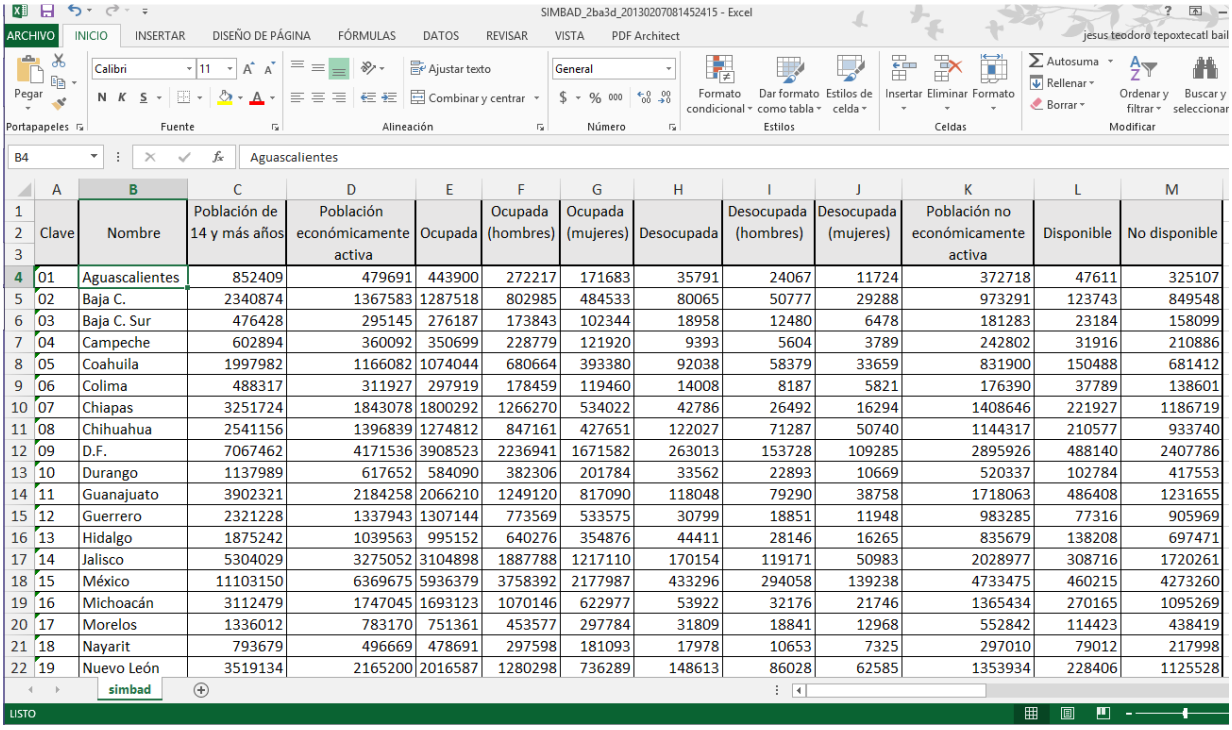



*Figura 4. 2 Visualización con WEKA de personas desempleadas con respecto al estado*

### 4.3.1 TRATAMIENTO DE DATOS

En esta sección daremos a conocer, todo el preprocesamiento de los datos, hasta obtener la información que se usó para las pruebas.

Entre las modificaciones que se realizaron a los datos, se encuentra la conversión de ciertos atributos, separación de columnas, y creación de nuevos campos, así como también la eliminación de datos que no eran de utilidad. En la Figura 4.3 se muestran algunos de los datos obtenidos por INEGI referentes a desempleo en México.

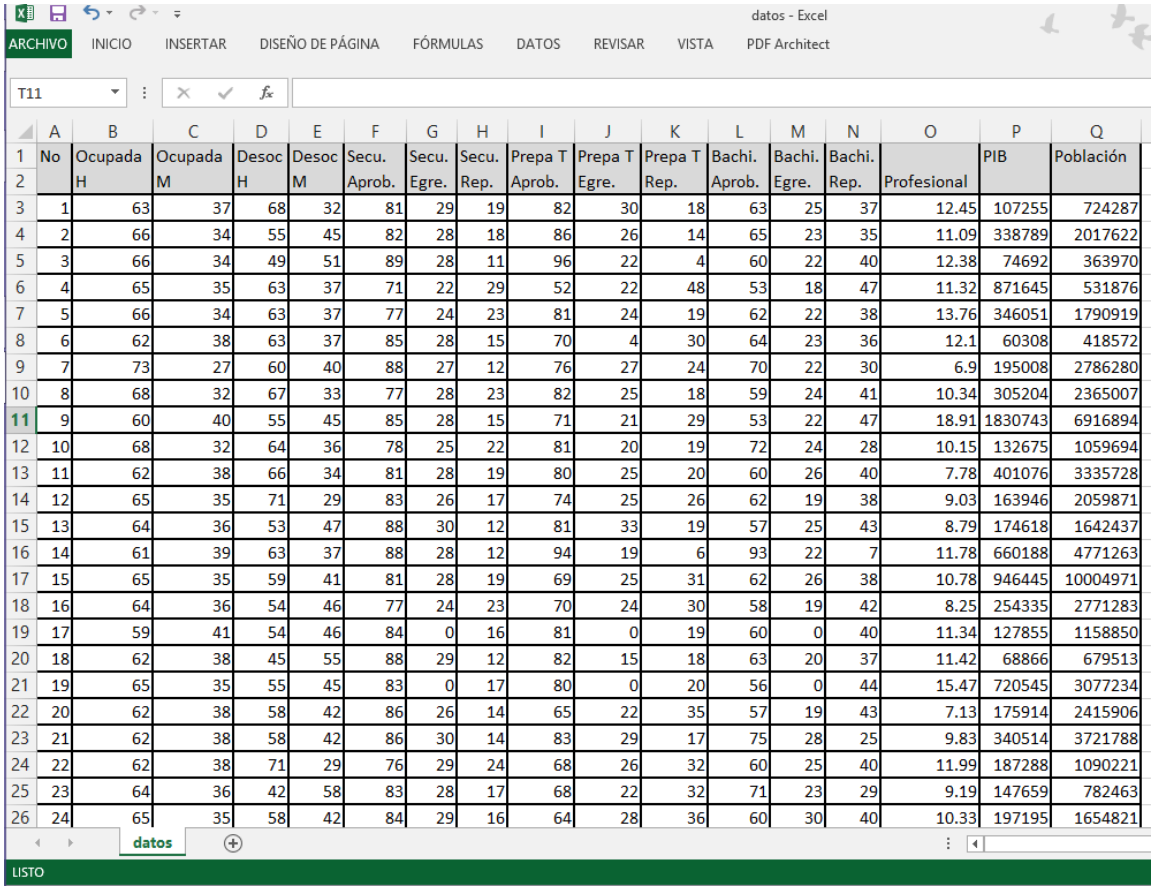



| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | |
| 2 | Clave | Nombre | Población de 14 y más años | Población económicamente activa | Ocupada | Ocupada (hombres) | Ocupada (mujeres) | Desocupada | Desocupada (hombres) | Desocupada (mujeres) | Población no económicamente activa | Disponible | No disponible |
| 3 | | | | | | | | | | | | | |
| 4 | 01 | Aguascalientes | 852409 | 479691 | 443900 | 272217 | 171683 | 35791 | 24067 | 11724 | 372718 | 47611 | 325107 |
| 5 | 02 | Baja C. | 2340874 | 1367583 | 1287518 | 802985 | 484533 | 80065 | 50777 | 29288 | 973291 | 123743 | 849548 |
| 6 | 03 | Baja C. Sur | 476428 | 295145 | 276187 | 173843 | 102344 | 18958 | 12480 | 6478 | 181283 | 23184 | 158099 |
| 7 | 04 | Campeche | 602894 | 360092 | 350699 | 228779 | 121920 | 9393 | 5604 | 3789 | 242802 | 31916 | 210886 |
| 8 | 05 | Coahuila | 1997982 | 1166082 | 1074044 | 680664 | 393380 | 92038 | 58379 | 33659 | 831900 | 150488 | 681412 |
| 9 | 06 | Colima | 488317 | 311927 | 297919 | 178459 | 119460 | 14008 | 8187 | 5821 | 176390 | 37789 | 138601 |
| 10 | 07 | Chiapas | 3251724 | 1843078 | 1800292 | 1266270 | 534022 | 42786 | 26492 | 16294 | 1408646 | 221927 | 1186719 |
| 11 | 08 | Chihuahua | 2541156 | 1396839 | 1274812 | 847161 | 427651 | 122027 | 71287 | 50740 | 1144317 | 210577 | 933740 |
| 12 | 09 | D.F. | 7067462 | 4171536 | 3908523 | 2236941 | 1671582 | 263013 | 153728 | 109285 | 2895926 | 488140 | 2407786 |
| 13 | 10 | Durango | 1137989 | 617652 | 584090 | 382306 | 201784 | 33562 | 22893 | 10669 | 520337 | 102784 | 417553 |
| 14 | 11 | Guanajuato | 3902321 | 2184258 | 2066210 | 1249120 | 817090 | 118048 | 79290 | 38758 | 1718063 | 486408 | 1231655 |
| 15 | 12 | Guerrero | 2321228 | 1337943 | 1307144 | 773569 | 533575 | 30799 | 18851 | 11948 | 983285 | 77316 | 905969 |
| 16 | 13 | Hidalgo | 1875242 | 1039563 | 995152 | 640276 | 354876 | 44411 | 28146 | 16265 | 835679 | 138208 | 697471 |
| 17 | 14 | Jalisco | 5304029 | 3275052 | 3104898 | 1887788 | 1217110 | 170154 | 119171 | 50983 | 2028977 | 308716 | 1720261 |
| 18 | 15 | México | 11103150 | 6369675 | 5936379 | 3758392 | 2177987 | 433296 | 294058 | 139238 | 4733475 | 460215 | 4273260 |
| 19 | 16 | Michoacán | 3112479 | 1747045 | 1693123 | 1070146 | 622977 | 53922 | 32176 | 21746 | 1365434 | 270165 | 1095269 |
| 20 | 17 | Morelos | 1336012 | 783170 | 751361 | 453577 | 297784 | 31809 | 18841 | 12968 | 552842 | 114423 | 438419 |
| 21 | 18 | Nayarit | 793679 | 496669 | 478691 | 297598 | 181093 | 17978 | 10653 | 7325 | 297010 | 79012 | 217998 |
| 22 | 19 | Nuevo León | 3519134 | 2165200 | 2016587 | 1280298 | 736289 | 148613 | 86028 | 62585 | 1353934 | 228406 | 1125528 |

*Figura 4. 3 Datos originales obtenidos de INEGI "Ocupacional"*

Como se puede observar en la figura 4.3 en la columna “Población de 14 y más años” los valores difieren mucho entre ellos, por lo tanto el obtener conocimiento útil con la minería de datos sería muy difícil o el conocimiento sería de poca relevancia. Lo mismo pasa con las demás columnas; esto se debe a que el número de habitantes varía considerablemente en los diferentes estados de la república mexicana. Parar corregir este problema se reemplaza la representación numérica de la población por su correspondiente valor porcentual.

En la figura 4.4 podemos observar cómo quedan los datos después de someterlos a los cambios mencionados anteriormente, es decir los porcentajes de hombres y mujeres ocupadas se calcula en referencia al total de personas ocupadas.  Cabe mencionar que las transformaciones hechas siguen representando los datos que se tenían anteriormente y las únicas afectaciones de este proceso son las de facilitar la extracción de conocimiento y que este sea aún más novedoso.




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | No | Ocupada H | Ocupada M | Desoc H | Desoc M | Secu. Aprob. | Secu. Egre. | Secu. Rep. | Prepa T Aprob. | Prepa T Egre. | Prepa T Rep. | Bachi. Aprob. | Bachi. Egre. | Bachi. Rep. | Profesional | PIB | Población |
| 2 | | | | | | | | | | | | | | | | | |
| 3 | 1 | 63 | 37 | 68 | 32 | 81 | 29 | 19 | 82 | 30 | 18 | 63 | 25 | 37 | 12.45 | 107255 | 724287 |
| 4 | 2 | 66 | 34 | 55 | 45 | 82 | 28 | 18 | 86 | 26 | 14 | 65 | 23 | 35 | 11.09 | 338789 | 2017622 |
| 5 | 3 | 66 | 34 | 49 | 51 | 89 | 28 | 11 | 96 | 22 | 4 | 60 | 22 | 40 | 12.38 | 74692 | 363970 |
| 6 | 4 | 65 | 35 | 63 | 37 | 71 | 22 | 29 | 52 | 22 | 48 | 53 | 18 | 47 | 11.32 | 871645 | 531876 |
| 7 | 5 | 66 | 34 | 63 | 37 | 77 | 24 | 23 | 81 | 24 | 19 | 62 | 22 | 38 | 13.76 | 346051 | 1790919 |
| 8 | 6 | 62 | 38 | 63 | 37 | 85 | 28 | 15 | 70 | 4 | 30 | 64 | 23 | 36 | 12.1 | 60308 | 418572 |
| 9 | 7 | 73 | 27 | 60 | 40 | 88 | 27 | 12 | 76 | 27 | 24 | 70 | 22 | 30 | 6.9 | 195008 | 2786280 |
| 10 | 8 | 68 | 32 | 67 | 33 | 77 | 28 | 23 | 82 | 25 | 18 | 59 | 24 | 41 | 10.34 | 305204 | 2365007 |
| 11 | 9 | 60 | 40 | 55 | 45 | 85 | 28 | 15 | 71 | 21 | 29 | 53 | 22 | 47 | 18.91 | 1830743 | 6916894 |
| 12 | 10 | 68 | 32 | 64 | 36 | 78 | 25 | 22 | 81 | 20 | 19 | 72 | 24 | 28 | 10.15 | 132675 | 1059694 |
| 13 | 11 | 62 | 38 | 66 | 34 | 81 | 28 | 19 | 80 | 25 | 20 | 60 | 26 | 40 | 7.78 | 401076 | 3335728 |
| 14 | 12 | 65 | 35 | 71 | 29 | 83 | 26 | 17 | 74 | 25 | 26 | 62 | 19 | 38 | 9.03 | 163946 | 2059871 |
| 15 | 13 | 64 | 36 | 53 | 47 | 88 | 30 | 12 | 81 | 33 | 19 | 57 | 25 | 43 | 8.79 | 174618 | 1642437 |
| 16 | 14 | 61 | 39 | 63 | 37 | 88 | 28 | 12 | 94 | 19 | 6 | 93 | 22 | 7 | 11.78 | 660188 | 4771263 |
| 17 | 15 | 65 | 35 | 59 | 41 | 81 | 28 | 19 | 69 | 25 | 31 | 62 | 26 | 38 | 10.78 | 946445 | 10004971 |
| 18 | 16 | 64 | 36 | 54 | 46 | 77 | 24 | 23 | 70 | 24 | 30 | 58 | 19 | 42 | 8.25 | 254335 | 2771283 |
| 19 | 17 | 59 | 41 | 54 | 46 | 84 | 0 | 16 | 81 | 0 | 19 | 60 | 0 | 40 | 11.34 | 127855 | 1158850 |
| 20 | 18 | 62 | 38 | 45 | 55 | 88 | 29 | 12 | 82 | 15 | 18 | 63 | 20 | 37 | 11.42 | 68866 | 679513 |
| 21 | 19 | 65 | 35 | 55 | 45 | 83 | 0 | 17 | 80 | 0 | 20 | 56 | 0 | 44 | 15.47 | 720545 | 3077234 |
| 22 | 20 | 62 | 38 | 58 | 42 | 86 | 26 | 14 | 65 | 22 | 35 | 57 | 19 | 43 | 7.13 | 175914 | 2415906 |
| 23 | 21 | 62 | 38 | 58 | 42 | 86 | 30 | 14 | 83 | 29 | 17 | 75 | 28 | 25 | 9.83 | 340514 | 3721788 |
| 24 | 22 | 62 | 38 | 71 | 29 | 76 | 29 | 24 | 68 | 26 | 32 | 60 | 25 | 40 | 11.99 | 187288 | 1090221 |
| 25 | 23 | 64 | 36 | 42 | 58 | 83 | 28 | 17 | 68 | 22 | 32 | 71 | 23 | 29 | 9.19 | 147659 | 782463 |
| 26 | 24 | 65 | 35 | 58 | 42 | 84 | 29 | 16 | 64 | 28 | 36 | 60 | 30 | 40 | 10.33 | 197195 | 1654821 |

*Figura 4. 4 Datos Finales*

## 4.4 CONSTRUCCIÓN DEL ALMACÉN DE DATOS

Una vez realizado el proceso de selección, limpieza y transformación, se procedió a la construcción del *DataWarehouse* con los datos modificados. El diseño se presenta en la figura 4.5.

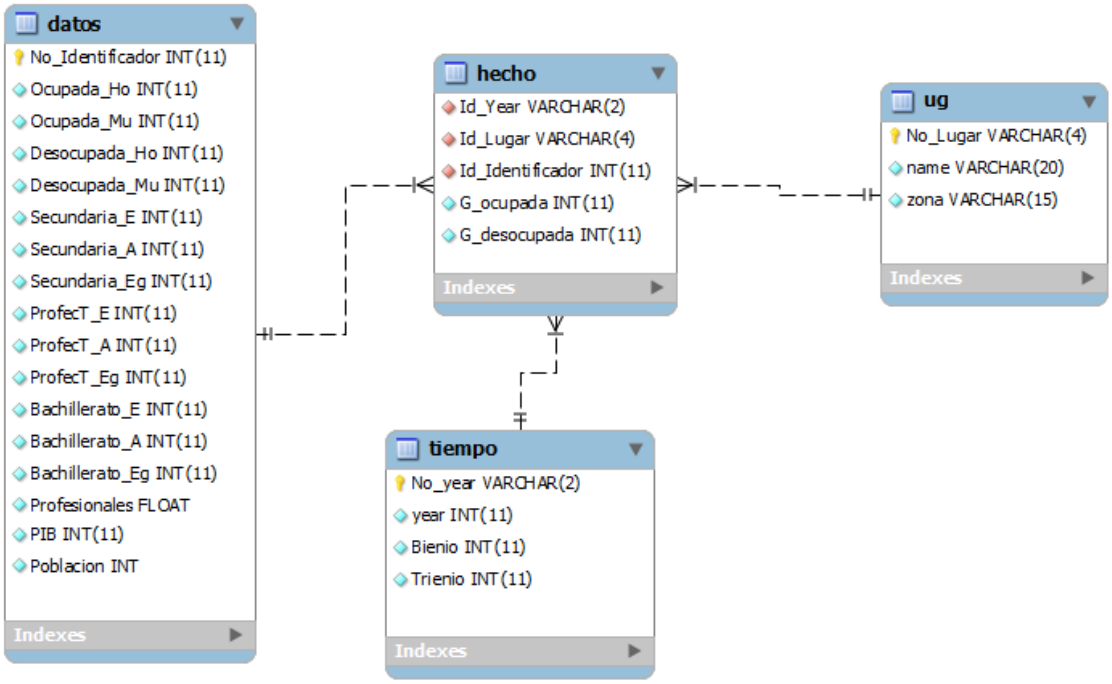



*Figura 4. 5 Modelo de la Base de Datos*

Se puede observar en la figura 4.5, que existe la tabla con el nombre de hechos la cual se relaciona con las tres restantes. Con esto se satisfacen requerimientos mínimos de minería de datos correspondientes a Información relacionada con la temática (datos), lugar (ug) y periodo (tiempo).

La tabla 4.3 representa los campos de la tabla Hechos, que es la tabla que se relaciona con las tres tablas de dimensión. La tabla 4.4 representa a la tabla ug (Ubicación geográfica), la tabla 4.5 muestra los datos contenidos por tiempo y finalmente la tabla 4.6 se refiere a los atributos de datos, la cual contiene la mayor cantidad de información acerca del desempleo y educación. Entonces estas cuatro tablas conforman el *DataWarehouse*.

*Tabla 4. 3 Atributos de la tabla Hechos*

| Nombre del campo | Descripción |
|---|---|
| Id_Year | Representa el atributo con el que se relaciona con la tabla Tiempo |
| Id_Lugar | Representa el atributo con el que se relaciona con la tabla UG |
| Id_Identificador | Representa el atributo con el que se relaciona con la tabla datos |
| G_ocupada | Se corresponde con la cantidad de personas que cuentan con un empleo |
| G_desocupada | Se corresponde con la cantidad de personas que no cuentan con un empleo |

*Tabla 4. 4 Atributos de la tabla UG*

| Nombre del campo | Descripción |
|---|---|
| No_Lugar | Representa el número asignado al estado |
| name | Nombre del estado |
| zona | La zona a la cual pertenece el estado |

*Tabla 4. 5 Atributos de la tabla Tiempo*

| Nombre del campo | Descripción |
|---|---|
| No_year | Representa el numero asignado al año |
| year | Año |
| Bienio | Bienio al que pertenece |
| Trienio | Trienio al que pertenece |

*Tabla 4. 6 Atributos de la tabla Datos*

| Nombre del campo | Descripción |
|---|---|
| No_Identificador | Representa el número asignado a datos |
| Ocupada_Ho | Representa el porcentaje de personas que trabajan y son hombres |
| Ocupada_Mu | Representa el porcentaje de personas que trabajan y son mujeres |
| Desocupada_Ho | Representa el porcentaje de personas que no trabajan y son hombres |
| Desocupada_Mu | Representa el porcentaje de personas que no trabajan y son mujeres |
| Secundaria_A | Representa el porcentaje de estudiantes de secundaria que aprueban |
| Secundaria_E | Representa el porcentaje de estudiantes de secundaria que egresan |
| Secundaria_R | Representa el porcentaje de estudiantes de secundaria que reprueban |
| PrepaT_A | Representa el porcentaje de estudiantes de preparatoria técnica que aprueban |
| PrepaT_E | Representa el porcentaje de estudiantes de preparatoria técnica que egresan |
| PrepaT_R | Representa el porcentaje de estudiantes de preparatoria técnica que reprueban |
| Bachillerato_A | Representa el porcentaje de estudiantes de bachillerato que aprueban |
| Bachillerato_E | Representa el porcentaje de estudiantes de bachillerato que egresan |
| Bachillerato_R | Representa el porcentaje de estudiantes de bachillerato que reprueban |
| Profesionales | Representa el porcentaje de profesionales existentes. |
| PIB | Representa el producto interno bruto |
| Poblacion | Representa a población total del estado. |

## 4.5 FASE DE MINERÍA DE DATOS

Esta fase tiene como objetivo extraer y presentar conocimiento implícito, que podrá ser utilizado por el usuario.

El proceso comienza a partir de un modelo basado en un gran volumen de datos, que han sido recopilados y procesados para dicho fin. En estos datos se encuentran patrones, reglas y relaciones que no se aprecian a simple vista, y pueden utilizarse para explicar situaciones pasadas, entender los datos y hacer predicciones.

La aplicación de técnicas de minería de datos se basó en el hecho de la naturaleza descriptiva del problema a resolver. Ninguna de los atributos del conjunto de datos representa una variable de clases que permita hacer predicciones acerca de la cantidad de desempleados. [14]

***K-Means*:** fue la primera técnica aplicada, (a partir de un número *k* de *clusters*, obtenidos por medio de una previa visualización de los datos, que sugiere un número *k=4)*. Dado que el conjunto de datos es numérico, la aplicación de *k-Means* encaja perfectamente. La técnica también permite dividir los datos en grupos teniendo en cuenta el criterio de la distancia Euclidiana. [13]

Una explicación completa de los métodos aplicados se presenta en el Capítulo 5.

## 4.5 CONCLUSIONES

En este capítulo se presentó la metodología KDD la cual fue empleada, para el diseño, construcción e implantación de las vistas minables de datos para el análisis de factores educativos que influyen en el desempleo en México. Además se explica a grandes rasgos la forma en que se modificaron los datos en cada etapa.

Finalmente se describieron las tareas, técnicas y métodos de Minería de Datos que se emplearon para la obtención del conocimiento contenido en el Almacén de Datos antes construido.

# CAPÍTULO 5

# EXPERIMENTOS

En este capítulo se mostraran los resultados del *clustering* o agrupamiento el cual es un procedimiento de reunión de una serie de vectores según criterios habitualmente basados en distancias. Se trata de disponer los vectores de entrada, de forma que estén más cercanos aquellos que tengan características comunes [17] y hacer una comparativa final con respecto de los resultados obtenidos.

La figura 5.1 muestra los 19 atributos que son considerados para la realización de los experimentos.

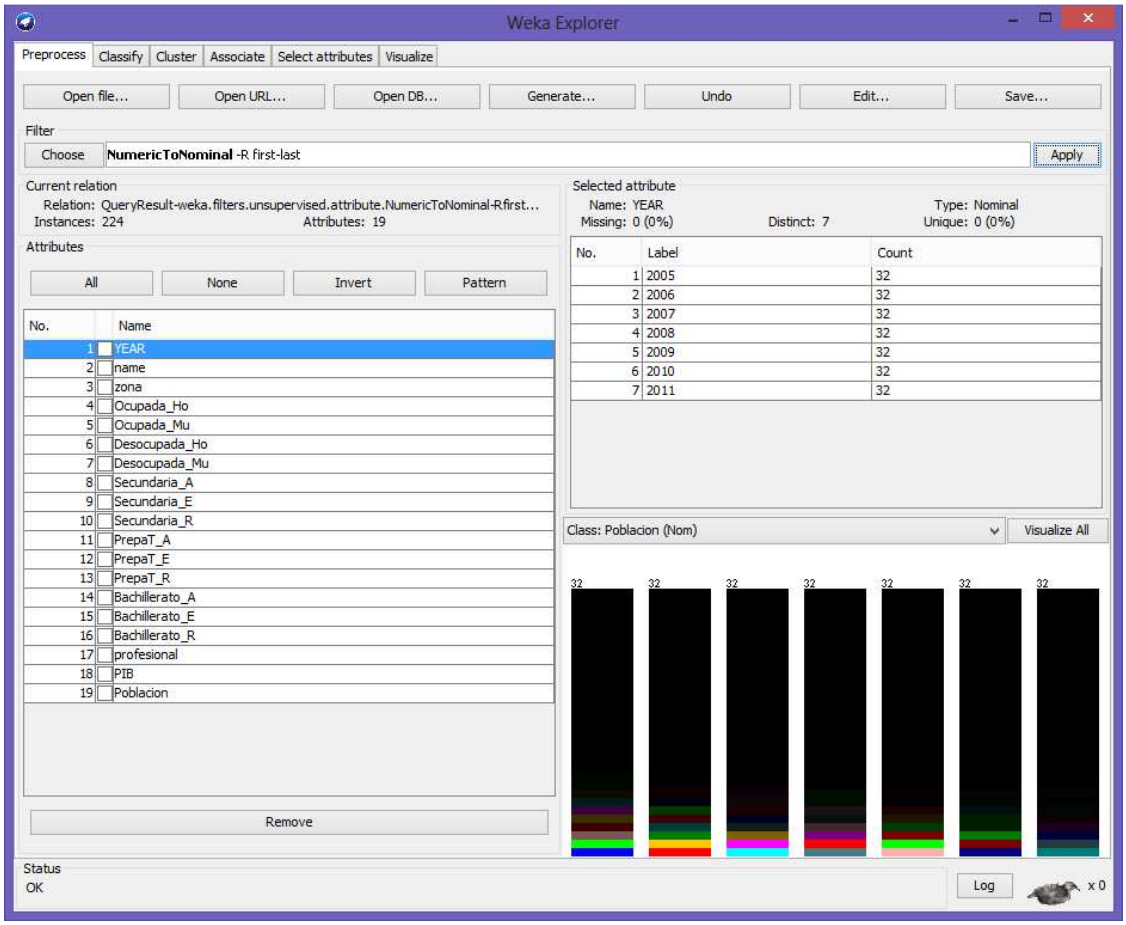



*Figura 5. 1 Atributos de la vista minable.*

## 5.1 EXPERIMENTO 1 – EM

En este primer experimento para conocer el número inicial de *clusters* se realiza una prueba con el método EM con un parámetro cantidad de *clusters* de -1. Esto con la finalidad de obtener el valor K del algoritmo *Kmeans*, donde la K indica el número de *clusters* a obtener por dicho método. En la figura 5.2 podemos observar el resultado de aplicar el método EM el cual nos propone usar 5 *clusters*.

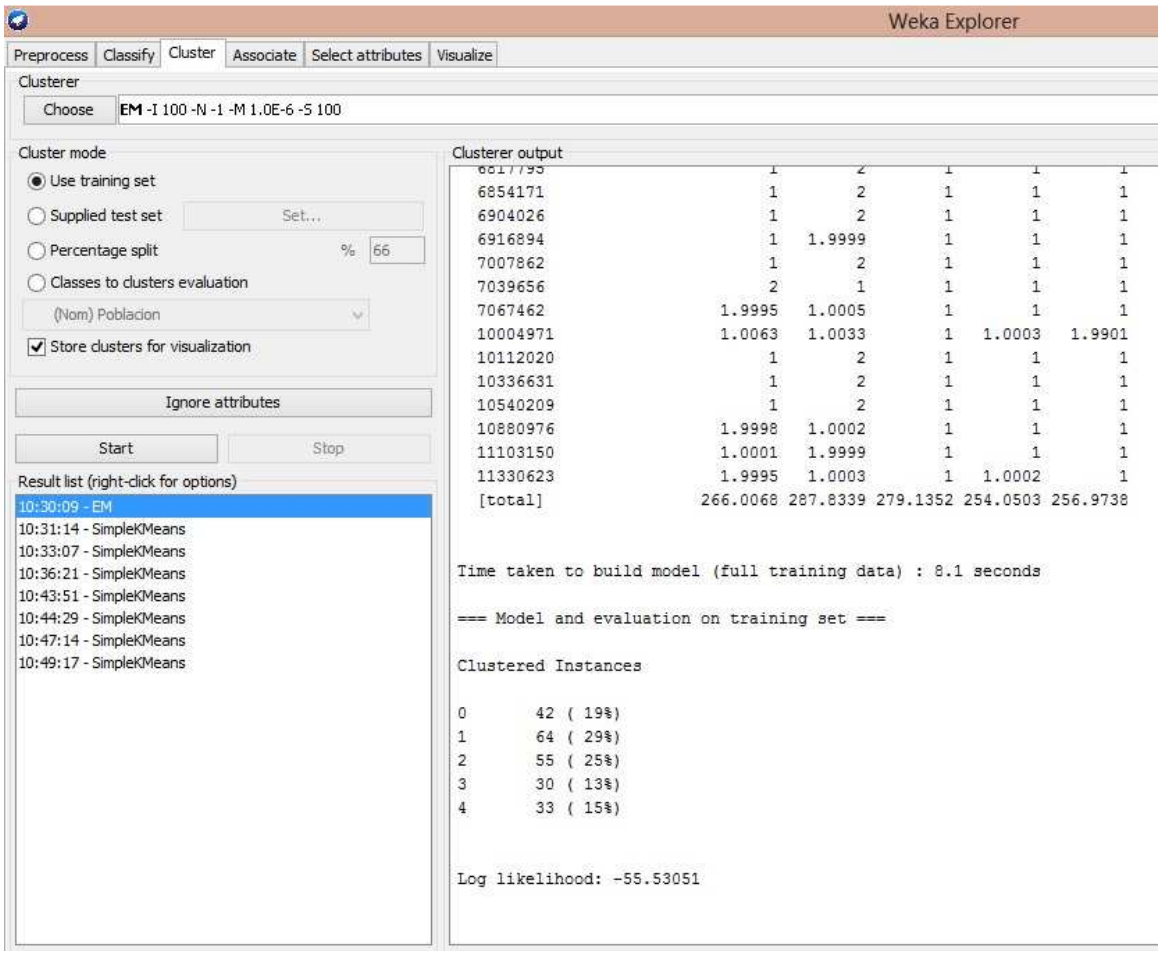



*Figura 5. 2 Resultados del método EM.*

## 5.2 EXPERIMENTO 2 – K-MEANS

En esta sección veremos los resultados obtenidos al hacer uso del método "SimpleKMeans". Se tomaron en cuenta los resultados del método anterior, es decir se considera K=5. Los resultados se muestran en la figura 5.3.

14:09:02 - SimpleKMeans

```
Number of iterations: 5
Within cluster sum of squared errors: 3341.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2           3           4
                             (224)        (78)        (48)        (30)        (32)        (36)
=============================================================================================================
YEAR                           2005        2007        2011        2006        2009        2005
name                 Aguascalientes Baja California    Hidalgo     Jalisco  Nuevo Leon     Nayarit
zona                         Centro      Noroeste Oriente centro  Occidente     Noreste  Occidente
Ocupada_Ho                       63          63          61          61          64          62
Ocupada_Mu                       37          37          39          39          36          38
Desocupada_Ho                    63          67          63          57          58          70
Desocupada_Mu                    37          33          37          43          42          30
Secundaria_A                     85          85          88          86          83          90
Secundaria_E                     29          29          28          28          30          24
Secundaria_R                     15          15          12          14          17          10
PrepaT_A                         69          73          59          83          79          69
PrepaT_E                         22          22          27          29          25          24
PrepaT_R                         31          27          41          17          21          31
Bachillerato_A                   60          59          65          75          71          60
Bachillerato_E                   26          24          26          23          30          19
Bachillerato_R                   40          41          35          25          29          40
profesional                    9.83       12.74        8.79        9.83       15.47        8.25
PIB                           58586       63525       58586       67373       62696       60308
Poblacion                    363970      363970      416958      384257      447675      418572



Time taken to build model (full training data) : 0.33 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       78 ( 35%)
1       48 ( 21%)
2       30 ( 13%)
3       32 ( 14%)
4       36 ( 16%)
```

*Figura 5. 3 Resultados de SimpleKMeans con K=5.*

Como podemos observar en la Figura 5.3 se obtuvo un error de 3341. Además se puede apreciar que los valores para los atributos entre los diferentes clusters son muy parecidos excepto para el atributo PrepaT_A. Por otra parte es notable, que para las zonas Noreste y Occidente se generan 2 clusters para cada una pero ningún clúster fue generado para describir las zonas del sur del país. Tomando en cuenta los resultados obtenidos se dispuso hacer una prueba de clustering pero en esta ocasión con k=4.

La figura 5.4 ilustra los resultados obtenidos del método SimpleKMeans con K=4 y se observa que el error se incrementa pasando de 3341 a 3433 lo que significa que hay una diferencia de 92.

14:15:47 - SimpleKMeans

```
Number of iterations: 5
Within cluster sum of squared errors: 3433.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2           3
                             (224)        (87)        (60)        (30)        (47)
=============================================================================================
YEAR                         2005        2007        2010        2006        2005
name                Aguascalientes Baja California  Nayarit      Puebla   Nuevo Leon
zona                        Centro      Noroeste  Occidente  Oriente centro  Noreste
Ocupada_Ho                  63          63          61          61          64
Ocupada_Mu                  37          37          39          39          36
Desocupada_Ho               63          67          63          57          58
Desocupada_Mu               37          33          37          43          42
Secundaria_A                85          85          88          86          83
Secundaria_E                29          29          28          28          30
Secundaria_R                15          15          12          14          17
PrepaT_A                    69          69          59          83          79
PrepaT_E                    22          22          27          29          25
PrepaT_R                    31          31          41          17          21
Bachillerato_A              60          68          65          75          60
Bachillerato_E              26          24          26          23          26
Bachillerato_R              40          32          35          25          40
profesional                 9.83        10.15       11.42       9.83        15.47
PIB                         58586       63465       58586       62696       65112
Poblacion                   363970      363970      375614      384257      447675



Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       87 ( 39%)
1       60 ( 27%)
2       30 ( 13%)
3       47 ( 21%)
```

*Figura 5. 4 Resultados de SimpleKMeans con K=4.*

También podemos observar que en los clusters la zona Noreste se sigue repitiendo y que al igual del caso anterior la mayor divergencia de datos se aprecia para el atributo PrepaT_A.

Después de corroborar que al disminuir el número K de clusters el error se incrementa, optaremos por incrementar el número de clusters a 6 y observar el comportamiento del error cuadrático medio.

14:18:24 - SimpleKMeans

```
Number of iterations: 6
Within cluster sum of squared errors: 3273.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                    Cluster#
Attribute                 Full Data         0             1             2             3             4             5
                            (224)        (64)          (45)          (29)          (26)          (34)          (26)
=============================================================================================================================
YEAR                          2005         2007          2010          2006          2009          2005          2011
name                Aguascalientes Baja California      Hidalgo        Puebla     Nuevo Leon       Michoac      Morelos
zona                        Centro     Noroeste  Oriente centro       Sureste       Noreste    Occidente  Metropolitana
Ocupada_Ho                      63           63            61            61            64            62            59
Ocupada_Mu                      37           37            39            39            36            38            41
Desocupada_Ho                   63           67            63            57            58            70            55
Desocupada_Mu                   37           33            37            43            42            30            45
Secundaria_A                    85           85            88            86            83            90            85
Secundaria_E                    29           29            28            28            30            24            28
Secundaria_R                    15           15            12            14            17            10            15
PrepaT_A                        69           73            81            83            79            69            57
PrepaT_E                        22           22            25            29            25            24            15
PrepaT_R                        31           27            19            17            21            31            43
Bachillerato_A                  60           59            62            75            71            60            65
Bachillerato_E                  26           24            26            23            30            19            21
Bachillerato_R                  40           41            38            25            29            40            35
profesional                   9.83        12.74          8.79          9.83         15.47          8.25         11.34
PIB                          58586        63525         58586         67373         62696         60308         69265
Poblacion                   363970       363970        384257        430837        447675        418572        443485



Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       64 ( 29%)
1       45 ( 20%)
2       29 ( 13%)
3       26 ( 12%)
4       34 ( 15%)
5       26 ( 12%)
```

*Figura 5. 5 Resultados de SimpleKMeans con K=6.*

La figura 5.5 muestra los resultados el experimento con K=6, se aprecia que el error decrece pasando de 3341 a 3273 para un total de 68 puntos menos. Cabe mencionar que a diferencia del caso con k=5, donde se repetían dos zonas en este solo se repite la zona Noroeste y se generan clusters que describen las zonas Metropolitana y Sureste.

Observemos que al incrementar el número de *clusters* a 6, el error disminuye, por tanto en las siguientes pruebas usaremos una K=6.

## 5.3 EXPERIMENTO 3 – SELECCIÓN DE ATRIBUTOS

Debido a que la suma de errores es bastante grande es necesario reducirlo, por lo cual se emplearon métodos de selección de atributos para poder descartar los menos significativos. En la figura 5.6 podemos apreciar que el método empleado es el *"Ranker"* y para que este pueda funcionar es necesario que el evaluador de atributo sea el *"InfoGainAttributeEval"*.

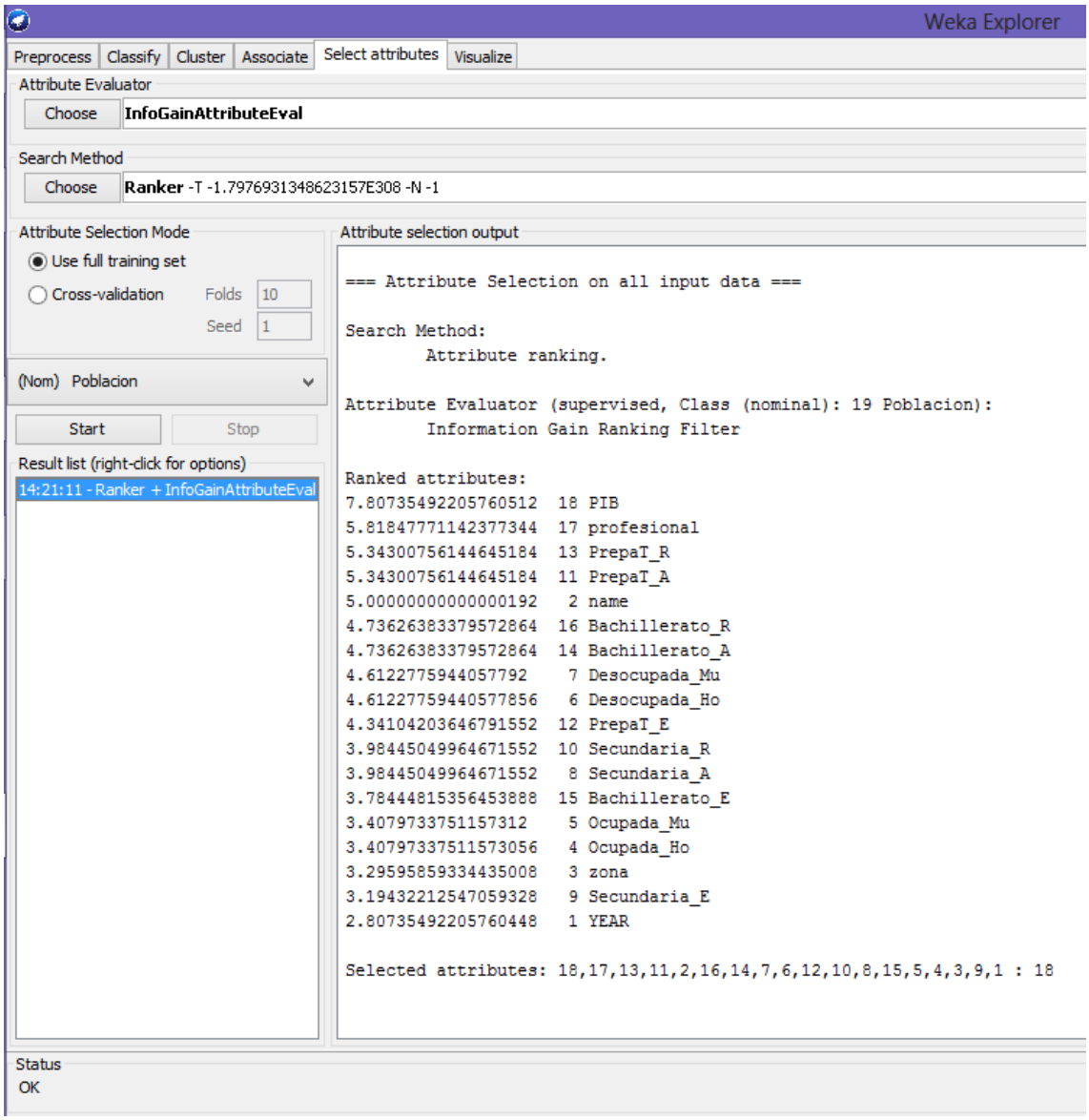



*Figura 5. 6 Resultados de aplicar selección de atributos supervisada.*

Considerando el resultado de la evaluación de atributos, se procedió a remover los últimos 6 atributos los cuales tienen menor peso en el proceso de agrupamiento y también el atributo numero 12 ya que es un valor calculable. El nuevo conjunto de datos se presenta en la figura 5.7.

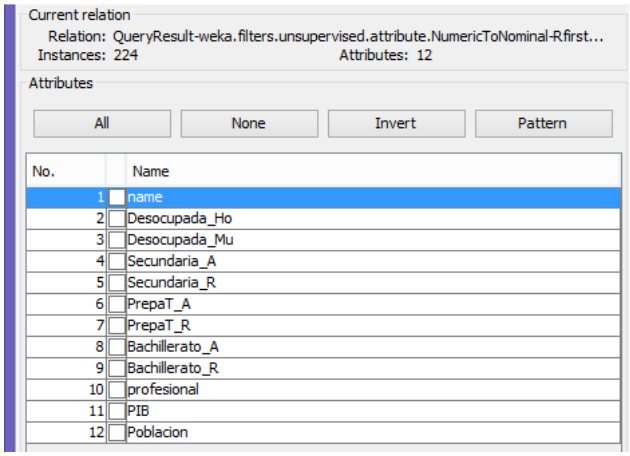



*Figura 5. 7 Atributos restantes de pues de aplicar la selección de atributos.*

Con el conjunto de datos de la figura 5.7 se aplica nuevamente SimpleKMeans con K=6 dado que este número es el que mejores resultados reportó en las pruebas anteriores. Como se puede notar en la Figura 5.8, la suma del error obtenida en este experimento decrece en 1113 puntos con respecto al experimento 2.

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 2160.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5
                            (224)       (94)       (20)       (27)       (31)       (30)       (22)
=============================================================================================================
name                  Aguascalientes   Campeche     Puebla    Jalisco Nuevo Leon    Michoac    Morelos
Desocupada_Ho                 63         67         66         57         58         70         64
Desocupada_Mu                 37         33         34         43         42         30         36
Secundaria_A                  85         85         87         86         83         81         84
Secundaria_R                  15         15         13         14         17         19         16
PrepaT_A                      69         70         82         83         79         69         80
PrepaT_R                      31         30         18         17         21         31         20
Bachillerato_A                60         59         79         75         71         60         65
Bachillerato_R                40         41         21         25         29         40         35
profesional                 9.83      10.34       9.83      11.78      15.47       8.25      11.34
PIB                        58586      58586      68866      62696      62903      65221      63525
Poblacion                 363970     375614     476428     384257     447675     363970     692460



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       94 ( 42%)
1       20 (  9%)
2       27 ( 12%)
3       31 ( 14%)
4       30 ( 13%)
5       22 ( 10%)
```

*Figura 5. 8 Resultados de SimpleKMeans con K=6 des pues de la selección de atributos.*

## 5.4 EXPERIMENTO 4 – REMOVER ATRIBUTOS DERIVADOS

Como resultado de la previa eliminación de los atributos poco relevantes la suma del error del agrupamiento decrece cerca de una tercera parte. Considerando estos resultados, una vez más se decide eliminar otro subconjunto de atributos, cuyo peso es menor en relevancia con respecto al resto de ellos. En esta ocasión fueron removidos aquellos atributos que pueden ser derivados de algún otro atributo, como lo son los correspondientes a los reprobados de los niveles Bachillerato y Preparatoria técnica, ya que se cuenta con los atributos que representan a los que resultaron aprobados.

Después de remover los atributos, el conjunto de datos resultante se muestran en la figura 5.9. El método a agrupamiento SimpleKMeans fue aplicado a este nuevo conjunto de datos y el resultado se observa en la figura 5.10.

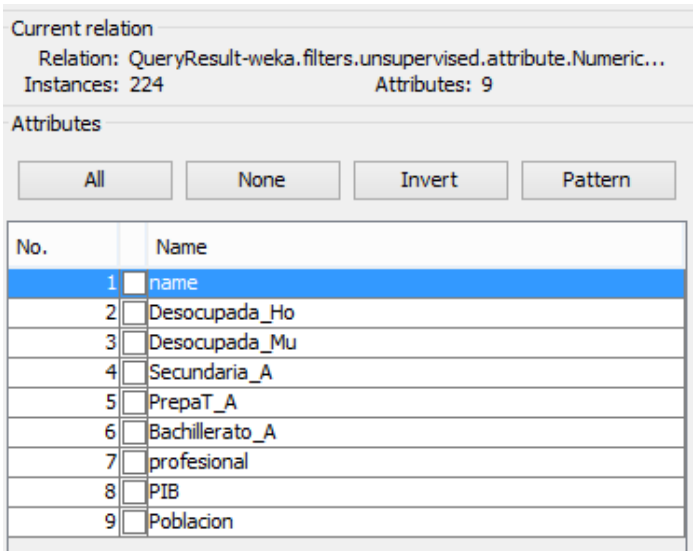



*Figura 5. 9 Atributos finales*

En la Figura 5.10 se aprecia que al descartar nuevamente atributos poco significativos, la suma del error se reduce y comprado con el experimento inicial con k=6 del apartado 5.2, el error representa el 50% de esa primera prueba.

```
=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 1644.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2           3           4           5
                              (224)        (91)        (22)        (24)        (29)        (33)        (25)
=============================================================================================================================
name                  Aguascalientes Baja California      Puebla     Guerrero   Nuevo Leon     Michoac      Morelos
Desocupada_Ho                   63          63          66          57          58          70          64
Desocupada_Mu                   37          37          34          43          42          30          36
Secundaria_A                    85          85          87          86          83          81          84
PrepaT_A                        69          70          82          83          79          69          57
Bachillerato_A                  60          59          79          75          56          60          65
profesional                   9.83       10.34        9.83        8.24       15.47        8.25       11.34
PIB                          58586       58586       68866       62696       62903       63465       63525
Poblacion                   363970      375614      476428      384257      447675      363970      553796



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       91 ( 41%)
1       22 ( 10%)
2       24 ( 11%)
3       29 ( 13%)
4       33 ( 15%)
5       25 ( 11%)
```

*Figura 5. 10 Resultados de SimpleKMeans con K=6 con atributos finales.*

## 5.5 COMPARACIÓN DE RESULTADOS

En esta sección se presentan a modo de resumen los resultados obtenidos. Se muestran en la Tabla 5.1, los datos relevantes generados por el agrupamiento, así como el número de clusters definidos, suma del error obtenida y el número de atributos del conjunto de datos.

*Tabla 5. 1 Comparativa de resultados.*

<table>
<tr><th>EXPERIMENTO</th><th>K</th><th>Error</th><th>#Atributos</th></tr>
<tr><td>

Cluster centroids:

| Attribute | Full Data (224) | Cluster# 0 (87) | 1 (60) | 2 (30) | 3 (47) |
|---|---|---|---|---|---|
| YEAR | 2005 | 2007 | 2010 | 2006 | 2005 |
| name | Aguascalientes | Baja California | Nayarit | Puebla | Nuevo Leon |
| zona | Centro | Noroeste | Occidente | Oriente centro | Noreste |
| Ocupada_Ho | 63 | 63 | 61 | 61 | 64 |
| Ocupada_Mu | 37 | 37 | 39 | 39 | 36 |
| Desocupada_Ho | 63 | 67 | 63 | 57 | 58 |
| Desocupada_Mu | 37 | 33 | 37 | 43 | 42 |
| Secundaria_A | 85 | 85 | 88 | 86 | 83 |
| Secundaria_E | 29 | 29 | 28 | 28 | 30 |
| Secundaria_R | 15 | 15 | 12 | 14 | 17 |
| PrepaT_A | 69 | 69 | 59 | 83 | 79 |
| PrepaT_E | 22 | 22 | 27 | 29 | 25 |
| PrepaT_R | 31 | 31 | 41 | 17 | 21 |
| Bachillerato_A | 60 | 68 | 65 | 75 | 60 |
| Bachillerato_E | 26 | 24 | 26 | 23 | 26 |
| Bachillerato_R | 40 | 32 | 35 | 25 | 40 |
| profesional | 9.83 | 10.15 | 11.42 | 9.83 | 15.47 |
| PIB | 58586 | 63465 | 58586 | 62696 | 65112 |
| Poblacion | 363970 | 363970 | 375614 | 384257 | 447675 |

</td><td>4</td><td>3433</td><td>19</td></tr>
<tr><td>

Cluster centroids:

| Attribute | Full Data (224) | Cluster# 0 (78) | 1 (48) | 2 (30) | 3 (32) | 4 (36) |
|---|---|---|---|---|---|---|
| YEAR | 2005 | 2007 | 2011 | 2006 | 2009 | 2005 |
| name | Aguascalientes | Baja California | Hidalgo | Jalisco | Nuevo Leon | Nayarit |
| zona | Centro | Noroeste | Oriente centro | Occidente | Noreste | Occidente |
| Ocupada_Ho | 63 | 63 | 61 | 61 | 64 | 62 |
| Ocupada_Mu | 37 | 37 | 39 | 39 | 36 | 38 |
| Desocupada_Ho | 63 | 67 | 63 | 57 | 58 | 70 |
| Desocupada_Mu | 37 | 33 | 37 | 43 | 42 | 30 |
| Secundaria_A | 85 | 85 | 88 | 86 | 83 | 90 |
| Secundaria_E | 29 | 29 | 28 | 28 | 30 | 24 |
| Secundaria_R | 15 | 15 | 12 | 14 | 17 | 10 |
| PrepaT_A | 69 | 73 | 59 | 83 | 79 | 69 |
| PrepaT_E | 22 | 22 | 27 | 29 | 25 | 24 |
| PrepaT_R | 31 | 27 | 41 | 17 | 21 | 31 |
| Bachillerato_A | 60 | 59 | 65 | 75 | 71 | 60 |
| Bachillerato_E | 26 | 24 | 26 | 23 | 30 | 19 |
| Bachillerato_R | 40 | 41 | 35 | 25 | 29 | 40 |
| profesional | 9.83 | 12.74 | 8.79 | 9.83 | 15.47 | 8.25 |
| PIB | 58586 | 63525 | 58586 | 67373 | 62696 | 60308 |
| Poblacion | 363970 | 363970 | 416958 | 384257 | 447675 | 418572 |

</td><td>5</td><td>3341</td><td>19</td></tr>
</table>

Cluster centroids:

| Attribute | Full Data (224) | Cluster# 0 (64) | 1 (45) | 2 (29) | 3 (26) | 4 (34) | 5 (26) |
|---|---|---|---|---|---|---|---|
| YEAR | 2005 | 2007 | 2010 | 2006 | 2009 | 2005 | 2011 |
| name | Aguascalientes | Baja California | Hidalgo | Puebla | Nuevo Leon | Michoac | Morelos |
| zona | Centro | Noroeste | Oriente centro | Sureste | Noreste | Occidente | Metropolitana |
| Ocupada_Ho | 63 | 63 | 61 | 61 | 64 | 62 | 59 |
| Ocupada_Mu | 37 | 37 | 39 | 39 | 36 | 38 | 41 |
| Desocupada_Ho | 63 | 67 | 63 | 57 | 58 | 70 | 55 |
| Desocupada_Mu | 37 | 33 | 37 | 43 | 42 | 30 | 45 |
| Secundaria_A | 85 | 85 | 88 | 86 | 83 | 90 | 85 |
| Secundaria_E | 29 | 29 | 28 | 28 | 30 | 24 | 28 |
| Secundaria_R | 15 | 15 | 12 | 14 | 17 | 10 | 15 |
| PrepaT_A | 69 | 73 | 81 | 83 | 79 | 69 | 57 |
| PrepaT_E | 22 | 22 | 25 | 29 | 25 | 24 | 15 |
| PrepaT_R | 31 | 27 | 19 | 17 | 21 | 31 | 43 |
| Bachillerato_A | 60 | 59 | 62 | 75 | 71 | 60 | 65 |
| Bachillerato_E | 26 | 24 | 26 | 23 | 30 | 19 | 21 |
| Bachillerato_R | 40 | 41 | 38 | 25 | 29 | 40 | 35 |
| profesional | 9.83 | 12.74 | 8.79 | 9.83 | 15.47 | 8.25 | 11.34 |
| PIB | 58586 | 63525 | 58586 | 67373 | 62696 | 60308 | 69265 |
| Poblacion | 363970 | 363970 | 384257 | 430837 | 447675 | 418572 | 443485 |

| | | |
|---|---|---|
| 6 | 3273 | 19 |

Cluster centroids:

| Attribute | Full Data (224) | Cluster# 0 (94) | 1 (20) | 2 (27) | 3 (31) | 4 (30) | 5 (22) |
|---|---|---|---|---|---|---|---|
| name | Aguascalientes | Campeche | Puebla | Jalisco | Nuevo Leon | Michoac | Morelos |
| Desocupada_Ho | 63 | 67 | 66 | 57 | 58 | 70 | 64 |
| Desocupada_Mu | 37 | 33 | 34 | 43 | 42 | 30 | 36 |
| Secundaria_A | 85 | 85 | 87 | 86 | 83 | 81 | 84 |
| Secundaria_R | 15 | 15 | 13 | 14 | 17 | 19 | 16 |
| PrepaT_A | 69 | 70 | 82 | 83 | 79 | 69 | 80 |
| PrepaT_R | 31 | 30 | 18 | 17 | 21 | 31 | 20 |
| Bachillerato_A | 60 | 59 | 79 | 75 | 71 | 60 | 65 |
| Bachillerato_R | 40 | 41 | 21 | 25 | 29 | 40 | 35 |
| profesional | 9.83 | 10.34 | 9.83 | 11.78 | 15.47 | 8.25 | 11.34 |
| PIB | 58586 | 58586 | 68866 | 62696 | 62903 | 65221 | 63525 |
| Poblacion | 363970 | 375614 | 476428 | 384257 | 447675 | 363970 | 692460 |

| | | |
|---|---|---|
| 6 | 2175 | 12 |

Cluster centroids:

| Attribute | Full Data (224) | Cluster# 0 (91) | 1 (22) | 2 (24) | 3 (29) | 4 (33) | 5 (25) |
|---|---|---|---|---|---|---|---|
| name | Aguascalientes | Baja California | Puebla | Guerrero | Nuevo Leon | Michoac | Morelos |
| Desocupada_Ho | 63 | 63 | 66 | 57 | 58 | 70 | 64 |
| Desocupada_Mu | 37 | 37 | 34 | 43 | 42 | 30 | 36 |
| Secundaria_A | 85 | 85 | 87 | 86 | 83 | 81 | 84 |
| PrepaT_A | 69 | 70 | 82 | 83 | 79 | 69 | 57 |
| Bachillerato_A | 60 | 59 | 79 | 75 | 56 | 60 | 65 |
| profesional | 9.83 | 10.34 | 9.83 | 8.24 | 15.47 | 8.25 | 11.34 |
| PIB | 58586 | 58586 | 68866 | 62696 | 62903 | 63465 | 63525 |
| Poblacion | 363970 | 375614 | 476428 | 384257 | 447675 | 363970 | 553796 |

| | | |
|---|---|---|
| 6 | 1644 | 9 |

Al observar los resultados encontramos que en todos los casos siempre se obtuvo como mínimo un estado perteneciente a la zona Norte y en ninguno de los casos se obtuvo un estado que se encontrase al sur de la zona Centro. Estos resultados podrían justificarse debido a que en la zona Norte se encuentra una gran cantidad de población flotante, es decir la población compuesta por aquellas personas que residen temporalmente en la zona, en este caso se refiere a todos los migrantes y los deportados que están a la espera de poder pasar a frontera.

Por otra parte el hecho de que en los resultados obtenidos no se contemple ningún estado de la zona Sur, es que en esta zona los índices de profesionistas es el más bajo en comparación de otras zonas al igual que de aprovechamiento del nivel medio superior, curiosamente en el nivel secundaria sus índices de aprovechamiento son bastante altos.

Con todo lo anterior podemos denotar que la relación específica entre la educación y el desempleo es que a mayor índice de aprovechamiento el índice de desempleo también crece en relación. Además al revisar los datos año por año la zona sur destaca muy poco en los índices de aprovechamiento menores los tiene la zona sur y oriente.

# CAPÍTULO 6

# CONCLUSIONES

Este capítulo presenta las conclusiones y comentarios finales del presente trabajo. Discute brevemente las lecciones aprendidas en durante su desarrollo, así como de las aportaciones alcanzadas. Por último ofrece un panorama del posible trabajo que pudiera seguir este proyecto en el futuro.

Durante el desarrollo del trabajo de tesis, fueron diversos los obstáculos a los cuales me tuve que enfrentar, desde el momento de definir el tema, hasta el comenzar a reportar los resultados obtenidos. Algunos de estos obstáculos fueron, por ejemplo, conseguir la información ya que algunos de los datos utilizados en el proceso de pruebas, fueron difíciles de conseguir ya que algunos de estos eran obtenidos de encuestas que solo se realizan cada 5 años lo cual retrasaba la elaboración del DataWarehouse.

## 6.1 CONOCIMIENTOS ADQUIRIDOS

En el desarrollo de la tesis se tuvieron que resolver distintos problemas, los cuales nos llevaron a adquirir y recordar una serie de conocimientos en los siguientes temas:

- Excel ya que la información proporcionada por INEGI se encuentra en formato ctv.
- SQL Server para la construcción del Almacén de Datos.
- WEKA para la aplicación de métodos que nos proporcionen información.

Los conocimientos anteriores con respecto a las herramientas pueden servir para realizar tareas de nivel profesional por ejemplo el tratamiento de grandes volúmenes de información mediante SQL server o la extracción de conocimiento mediante la herramienta WEKA.

Enfocándonos a los resultados obtenidos con respecto del último experimento ilustrado en la figura 5.10 podemos describirlo de la siguiente manera:

**Grupo 1:** grana aprovechamiento en preparatoria técnica, bajo nivel de aprovechamiento en bachillerato, cantidad de profesionistas media y población pequeña nos da más hombres desempleados.

**Grupo 2:** excelente aprovechamiento en preparatoria técnica, gran nivel de aprovechamiento en bachillerato y una población regular nos da el doble de hombres desempleados que de mujeres.

**Grupo 3:** excelente aprovechamiento en preparatoria técnica, gran nivel de aprovechamiento en bachillerato y una población pequeña nos da una cantidad de mujeres desempleadas muy parecida al de hombres desempleados.

**Grupo 4:** excelente aprovechamiento en preparatoria técnica, bajo nivel de aprovechamiento en bachillerato, una muy alta cantidad de profesionistas y una población regular nos da que la cantidad de desempleados hombres no es tan diferida de desempleadas mujeres.

**Grupo 5:** buen nivel de aprovechamiento en preparatoria técnica, bajo nivel de aprovechamiento en bachillerato, muy poca población profesionista y una población pequeña nos da que los desempleados hombres son más del doble que de desempleadas mujeres.

**Grupo 6:** bajo nivel de aprovechamiento en preparatoria técnica y bachillerato, una alta población de profesionistas y una gran población nos da que los hombres desempleados son el doble que de mujeres desempleadas.

## 6.2 TRABAJO FUTURO

Entre los posibles caminos que pudiera seguir este trabajo de tesis en un futuro están el trabajar en la recolección de información de manera más personal tal como sexo, nivel educativo, estado ocupacional, salario, entre los más significativos. Esta nueva información facilitará la construcción de modelos predictivos que son necesarios para saber en qué zonas probablemente se generará un gran índice de desempleo y cuáles serían las afectaciones a nivel socioeconómico.

Esto permitiría a los especialistas crear campañas de prevención y planes de contingencia adecuados para contrarrestar los efectos negativos de la falta de ingresos.

## 6.3 CONCLUSIONES FINALES

El haber tenido la oportunidad de desarrollar un trabajo de investigación de este tipo me genera una gran satisfacción, ya que me permitió adquirir experiencia y habilidades nuevas. La experiencia obtenida en el desarrollo del proyecto, permite concluir:

- Mucha de la información que actualmente se tiene es desaprovechada ya que no es procesada parar obtener conocimiento y así generar mejores beneficios ya sea en el ámbito privado como público.
- El uso de herramientas de software libre para los procesos de extracción y exploración por sus bajos costos resultó satisfactorio.
- El manejo de dichas herramientas no requiere de conocimientos avanzados de computación ya que con una capacitación básica personas ajenas al campo informático pudieran hacer uso de estas.
- El desarrollo de los procesos de extracción, transformación y carga son los apropiados según la información requerida.
- El uso de WEKA permite un manejo intuitivo y sencillo, sin requerir de un especialista informático.

# REFERENCIAS


**[1]** Aguilar Centeno Jaime, Lozano Cruz Eduardo. (2007). *El empleo y desempleo en México 2000-2006.* Tesis de Licenciatura, Benemerita Universidad Autonoma de Puebla, México.

**[2]** J. Hernández-Orallo, M. J. Ramírez-Quintana & C. Ferri. (2008). *Introducción a la Minería de Datos.* España. Pearson, Prentice Hall.

**[3]** Navarrete Carrasco, R.C. (2001). *Business Intelligence: La necesidad actual*. Recuperado el 10 agosto de 2013 de: http://www.monografias.com/trabajos10/busi/busi.shtml

**[4]** Pérez César & Santín Daniel. (2006). *Data Mining Soluciones con Enterprise Miner*. España. Ra-Ma.

**[5]** José C. Riquelme, Roberto Ruiz & Karina Gilbert. (2006*). Minería de datos: Conceptos y tendencias*. Recuperado el 12 agosto de 2013 de: http://www.redalyc.org/articulo.oa?id=92502902

**[6]** OCDE. (Mayo 2012). *Organización para la Cooperación y el Desarrollo Económicos*. Recuperado el 10 noviembre 2013 de: http://www.oecd.org/centrodemexico/laocde/

**[7]** INSTITUTO NACIONAL DE ESTADISTICA Y GEOGRAFIA. (19 DE ENERO DE 2012). *Boletines de Prensa*. Recuperado el 9 de noviembre 2013 de: http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/comunicados/ocupbol.asp

**[8]** CEPAL. (Octubre 2013). *COMISION ECONOMMICAPARA AMERICA LATINA Y EL CARIBE*. Recuperado el 10 de diciembre 2013 de: http://www.eclac.cl/

**[9]** Angoitia Espinoza Itziar. (2008). *DataWarehouse para la gestión de lista de espera sanitaria.* Tesis Licenciatura en Informática, Facultad de Informática Universidad Politécnica de Madrid, España.

**[10]** Clauidio. (2003). *Datawarehousing: Teoría. Introducción al Concepto Data Warehousing* . Recuperado el 15 de diciembre 2013 de: http://personal.lobocom.es/claudio/gen006.htm

**[11]** Ramez E, (2007), *Fundamentos de Sistemas de Bases de Dato (*Madrid.) Addison Wesley. 5ta Edición

**[12]** Pérez C, Santín D, (2006) *Data Mining Soluciones con Enterprise Miner... (*México). Alfaoemega. Primera edición.

**[13]** Gabits. (2009). *Algoritmos de Minería de Datos. Portal web. Actualizada: diciembre 2009*. Recuperado el 10 de febrero 2014 de: http://algoritmosmineriadatos.blogspot.mx/2009/12/algoritmo-naive-bayes.html

**[14]**. Ricardo Aler. (2010). *Introducción al aprendizaje automático y a la minería de datos con Weka herramientas de la inteligencia artificial ingeniería informática.* Recuperado el 7 de marzo 2014 de: http://ocw.uc3m.es/ingenieria-informatica/herramientas-de-la-inteligencia-artificial/contenidos/transparencias/MDWEBHIA-clase.pdf

**[15].** López Cumplido A, (Agosto 2010), *Sistema para toma de decisiones acerca del cambio climático.* Tesis de Licenciatura, BUAP, Puebla, México.