UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA
ÁREA ACADÉMICA DE COMPUTACIÓN
Y ELECTRÓNICA

Aplicación de Minería de Datos para la identificación de perfiles de los factores migratorios en el estado de Hidalgo, México

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRÍA EN CIENCIAS COMPUTACIONALES

PRESENTA:
Kristell Daniella Franco Sánchez

DIRECTORA: Dra. Anilu Franco Arcega
CODIRECTOR: MCC. Luis Heriberto García Islas

Pachuca de Soto Hgo., Mayo de 2015

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

Área Académica de Computación y Electrónica

Maestría en Ciencias Computacionales

Oficio No. MCC/27/2015

Kristell Daniella Franco Sánchez

Presente.

Por este conducto le comunico que el jurado asignado para la revisión de su trabajo de tesis titulado "**Aplicación de Minería de Datos para la identificación de perfiles de los factores migratorios en el estado de Hidalgo, México**", que para obtener el grado de Maestro en Ciencias Computacionales fue presentado por usted ha tenido a bien, en reunión de sinodales autorizarlo para impresión.

PRESIDENTE:	DRA. MARÍA DE LOS ÁNGELES ALONSO LAVERNIA
VOCAL:	DRA. ANILU FRANCO ARCEGA
SECRETARIO:	M. EN C. FÈLIX A. CASTRO ESPINOZA
SUPLENTES:	M. EN C. LUÍS HERIBERTO GARCÍA ISLAS

A continuación se integran las firmas de conformidad de los integrantes del jurado

ATENTAMENTE.
"AMOR, ORDEN Y PROGRESO"
Mineral de la Reforma, Hgo., a 12 de mayo de 2015

Dra. Anilu Franco Arcega
Coordinadora de la Maestría en Ciencias Computacionales

c.c.p AFA/ apl

Instituto de Ciencias Básicas e Ingeniería,
Carretera Pachuca - Tulancingo Km. 4.5, Ciudad del Conocimiento,
Colonia Carboneras, Mineral de la Reforma, Hidalgo, México, C.P. 42184
Tel. +52 771 7172000 ext. 6734
afranco@uaeh.edu.mx

Con todo mi amor y cariño para mi esposo,
por ser mi fuerza y mi inspiración.
Te amo con toda el alma Heriberto.

A mi pequeñito Daniel,
por llenar mi vida de inmensa
felicidad y luz. Te amo hijito.

A nuestro bebe que viene en camino,
por ser una razón más para seguir luchando.
Te amo aún sin conocerte.

A Padre Jesús, por amarme y
consentirme tanto, por iluminar
en todo momento mi camino.

Les dedico todo el esfuerzo, tiempo,
dedicación y sacrificios, invertidos en esta tesis.

Agradecimientos

Le doy gracias a Dios, por todo el amor que me tiene, por estar siempre a mi lado, incluso en los momentos más obscuros y tristes, por guiar mi vida hacia la luz y colocarme siempre en el camino correcto.

Le agradezco infinitamente a mi esposo, Heriberto, sin ti nada de esto hubiera sido posible. Gracias por apoyarme y ser el motor que me daba fuerza para alcanzar mi sueño, por toda la paciencia que me has tenido y todos los momentos que tuviste que sacrificar para poderme apoyar, también te doy gracias por ser mi compañero de aventuras, mi cómplice, mi alma gemela. Te amo con toda el alma y no me alcanzan las palabras para agradecerte todo lo que has hecho por mi.

A mi pequeñito Daniel Heriberto, gracias por acompañarme en cada momento de la escritura de esta tesis, desde que estabas en mi vientre hasta cuando me hacías tomar una pausa para jugar un rato con tus carritos, gracias por ser mi motivación y enseñarme que tan grande puede ser un corazón para amar tanto y tan intensamente.

A mi mami, por todos los sacrificios que tuviste que hacer para sacarme adelante, por darme la herencia más grande que me pudiste haber dado: una carrera y las armas para defenderme en la vida, me hiciste una persona de bien, una mujer trabajadora y responsable.

A mi papi, que desde el cielo me sigues cuidando, porque nunca me has dejado sola, por tantos recuerdos tan bonitos que guardo en mi memoria.

A mi hermana Mery, mi cuñado Carlos y mis sobrinos Carlos Alberto y Lizeth Naomi; porque siempre creyeron en mi, por darme ánimos y escucharme cuando más lo necesitaba.

A mis suegros Heriberto y Ma. Luisa, por apoyarme en cada momento, por adoptarme y quererme como si fuera su hija.

A mi cuñada Cynthia, por todo su cariño y apoyo desde siempre.

A mis sinodales, porque sin duda, forman parte fundamental en esta tesis; por todo su tiempo y dedicación invertidos en este trabajo. Dra. Anilu, Mtro. Félix y Dra. María de los Ángeles, mi sincero e infinito agradecimiento por haber creído en mi, por brindarme su amistad, su apoyo incondicional, sus palabras de aliento, por toda la paciencia y todo su tiempo dedicado para enriquecer y mejorar mi trabajo; siempre les estaré agradecida.

A mi Directora de tesis, Anilu, por todo tu tiempo, por guiarme en esta experiencia, por tu amistad.

Al Dr. Aurelio Granados Alcantar, por su apoyo escencial en la elaboración de esta tesis, por haberme proporcionado el conjunto de datos de INEGI y por la validación de los resultados obtenidos.

A mis profesores de la Maestría, gracias por todas sus enseñanzas, por exigirme para sacar lo mejor de mi.

A mis amigos y mis compañeros de trabajo, gracias por alentarme y darme ánimos, por no dejar de creer en mi.

A mis alumnos, por motivarme para seguirme preparando, para ser mejor profesional y mejor persona.

Resumen

La Minería de Datos es una disciplina que ha sido aplicada exitósamente en diversos ámbitos, tales como la administración de negocios, marketing y ventas, diagósticos médicos, procesos de manufactura, astronomía, por mencionar algunos. En específico, el análisis de problemas sociales, donde la toma de decisiones es importante para implementar acciones respecto a estos problemas, es una de las áreas donde la Minería de Datos contribuye significativamente, al encontrar soluciones que permitan a los expertos comprender de manera más eficiente y efectiva los comportamientos humanos que derivan dichos problemas.

El presente trabajo muestra un análisis del fenómeno de la migración en el estado de Hidalgo, visto desde dos perspectivas diferentes: *los factores de tipo demográfico* que describen las viviendas de los migrantes y *los factores sociales* que detallan su perfil. Este análisis comprende la aplicación de técnicas computacionales de Minería de Datos para encontrar conocimiento útil, novedoso y comprensible, seleccionando aquellas técnicas que permitan describir y clasificar mejor estos datos.

Este estudio muestra, para ambos factores, sus características más relevantes, describe los grupos de migrantes obtenidos con técnicas de minería de datos y además revela el comportamiento de los mismos mediante la obtención de reglas lingüísticas. Este conocimiento puede ser potencialmente usado por Gobierno y agencias de servicios sociales en el estado de Hidalgo, para la creación de programas sociales específicos que ayuden a disminuir o prevenir la migración de la población.

Contenido

Índice de figuras

Índice de tablas

*

Introducción

En los últimos años, la Minería de Datos ha tenido gran impacto en la industria de la información, debido a la amplia disponibilidad de grandes volúmenes de datos, los cuales son almacenados en bases de datos de diferentes tipos. Además, con el fin de tomar buenas decisiones, las instituciones han desarrollado la imperiosa necesidad de convertir esos datos en información y conocimiento útil, que puedan ser utilizados en diversas aplicaciones que van desde la gestión empresarial, control de producción, análisis de mercado, diseño de la ingeniería, exploración de la ciencia, etc. En este sentido, las herramientas de Minería de Datos contribuyen enormemente a las estrategias de negocio, bases de conocimiento y a la investigación científica y médica [34].

La Minería de Datos es una de las fases del proceso de Descubrimiento de Conocimiento en las Bases de Datos (KDD por sus siglas en inglés: Knowledge Discovery in Databases) [24], que permite procesar automáticamente grandes cantidades de datos con la finalidad de poder identificar patrones que generen conocimiento y de esta manera permitir al usuario el uso de esta información valiosa para su contexto.

Diversas disciplinas aportan métodos e ideas a los procesos de Minería de Datos, con lo cual se desarrollan modelos y soluciones más completas y eficientes [34]. Entre estas disciplinas se pueden encontrar la estadística, aprendizaje automático, bases de datos, reconocimiento de patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, análisis de datos espaciales, etc.

Las técnicas de Minería de Datos pueden crear y resolver problemas de dos tipos: predictivo y descriptivo. Los modelos generados a partir de los métodos predictivos estiman valores futuros o desconocidos utilizando variables o campos de una base de datos conocidos para hacer la predicción.

Por su parte, los modelos descriptivos exploran las propiedades de los datos examinados para identificar patrones que expliquen o resuman los datos, o bien para encontrar relaciones entre ellos [21, 35].

Una de las áreas donde se ha aplicado exitosamente la Minería de Datos es en la solución y análisis de problemas sociales. Un problema social puede considerarse desde tres enfoques diferentes; en el primero se destacan los aspectos perjudiciales de los problemas, otro enfoque es el que concibe un problema social cuando un grupo significativo de la sociedad lo percibe y pone en marcha acciones para solucionarlo y por último, el enfoque que trata de integrar los aspectos perjudiciales y la percepción de la sociedad [10]. Desde el punto de vista del primer enfoque, un problema social puede ser definido como una situación que incumple con una o más normas generales compartidas y aprobadas por una parte del sistema social. Por su parte, desde la percepción de la sociedad, un problema social es el resultado de un proceso de definición colectiva que surge cuando una parte importante de la población considera como no deseadas a ciertas situaciones sociales y además puede transmitir esa percepción a otros sectores. Finalmente, desde el enfoque que integra los componentes objetivos y subjetivos, un problema social es algún aspecto de la sociedad acerca del cual un amplio número de personas están preocupadas.

El hecho de estudiar y analizar las causas, consecuencias y/o factores de los problemas sociales hace posible identificar conocimiento novedoso y a su vez relaciones que surgen de distintos problemas sociales como son la pobreza, la marginación, el narcotráfico, la delicuencia, etc. [3]. Además, con estos estudios se pueden planear efectivamente programas sociales [62], los cuales beneficien a grupos vulnerables de cualquier país.

La Minería de Datos se ha aplicado en diversos estudios sociales, tales como la identificación de características de una población [33], la caracterización de estudiantes [20], la segregación residencial [2], etc.; donde generalmente se utilizan variables de tipo socio-económico, para describir los comportamientos de estos fenómenos sociales.

De manera específica, el problema social de la migración humana es un fenómeno socio-espacial, que ocurre a causa y consecuencia de diversos cambios en ámbitos interdependientes, las estructuras sociales y las relaciones espaciales [26, 28]. Desde el enfoque demográfico, indica el cruce de los límites

de una división geográfica para cambiar de residencia [25]. En una migración intervienen dos áreas geográficas: aquella donde se inicia el desplazamiento, se le denomina "región de origen", y en la que finaliza, la "región destino" [72]; es considerado un emigrante al individuo que sale de su región de origen y un inmigrante al individuo que llega a la región destino.

Estudiar el fenómeno de migración en cualquier país es de gran importancia, principalmente porque puede influir en la dinámica y magnitud de la población, repercute en las estructuras sociales, culturales y económicas de una región o país, y es considerada en las políticas de desarrollo debido a que es un hecho trascendente en la vida de las personas y de las sociedades que buscan un mayor crecimiento, equidad y calidad de vida [25]. Por ejemplo, en el caso de los productores agrícolas, la migración es considerada como una alternativa que permite mejorar la calidad de vida de sus familias, debido a las condiciones de marginación y pobreza a las que están sometidos [42, 43, 51].

El análisis de la migración, permite además la elaboración de propuestas sobre la implementación de programas que estimulen a la inversión de los migrantes en sus comunidades de origen, la creación o mejora de las propuestas de los organismos gubernamentales, etc.

La migración de México a Estados Unidos se ha convertido en el mayor circuito migratorio entre dos países del mundo, lo que ha llevado a ambos países a centrar su atención a este fenómeno, ya que hoy en día ha experimentado una gran penetración en aspectos económicos, sociales y políticos [76]. En particular, en el estado de Hidalgo la movilidad hacia Estados Unidos inició desde los años treinta y, en los años cuarenta muchos hidalguenses se incorporaron al programa "bracero", posterior a esto la migración disminuyó considerablemente quedando casi pausada en estos lugares y se reinició hasta principios de los años ochenta masificándose la salida de la población [64, 66].

Debido a los resultados del Censo de Población y Vivienda del INEGI del año 2000, el estudio de la migración en el estado de Hidalgo se vuelve relevante [64], ya que se coloca como una entidad importante en migración internacional, principalmente hacia Estados Unidos, formando parte de la región "nueva" o "emergente" de migración en el país [4] junto con el estado

de Morelos, debido a que la intensidad migratoria de ambos estados es muy similar a la del estado de Jalisco, el cual tiene una tradición migratoria de más de cien años.

El estado de Hidalgo se ubica entre las diez principales entidades de origen de la migración internacional; es considerado por el Consejo Nacional de Población [17] un estado de alto grado de intensidad migratoria a los Estados Unidos, según la medición del índice de intensidad migratoria de los hogares mexicanos, realizado en el año 2000 [65]. También, en ese mismo año, de acuerdo con los datos de la muestra, del 10 % del Censo de Población y Vivienda del año 2000 del INEGI, el 8.7 % de los hogares hidalguenses tenían uno o más miembros migrantes [56]. De los hidalguenses que migraron permanentemente a los Estados Unidos en el periodo 1995- 2000, el 82.6 % son hombres y el 17.4 % son mujeres, de los hombres el 24.7 % están ubicados en las edades de 15 a 19 años y el 22.4 % en las edades de 20 a 24 años; de las mujeres migrantes, se ubican el 47.7 % y el 5.6 % respectivamente en los rangos de edades mencionados [57].

Como se observa, el estudio del fenómeno de migración puede realizarse a través de un análisis estadístico. Este análisis puede usarse cuando el estudio requiera describir aspectos o características específicos, por ejemplo determinar la edad promedio de un grupo social en particular o el grado de escolaridad más frecuente en los grupos. Típicamente, se emplean técnicas como el análisis de varianza, prueba de t apareada, correlación de Pearson, prueba de Kruskal-Wallis, correlación de Spearman, etc. [58]. Sin embargo, este tipo de técnicas no son las más apropiadas cuando se desea obtener de la información patrones de comportamiento de estos grupos sociales. Esto es una limitante de las técnicas que se utilizan tradicionalmente para el análisis de los problemas de tipo social.

Para resolver esta limitante, se propone el uso de técnicas de Minería de Datos. En este trabajo de investigación se busca generar un modelo de los perfiles de los migrantes que permita el análisis de la migración internacional en el estado de Hidalgo, utilizando la información del Censo de Población y Vivienda 2010 del INEGI y la Encuesta sobre Migración en las Fronteras (EMIF) del Colegio de la Frontera Norte (COLEF). Con ello se busca obtener conocimiento que permita explicar la composición de los grupos de migrantes en el estado de Hidalgo.

El objetivo general de este trabajo es:

Encontrar perfiles que describen a los factores demográficos y sociales, para el estudio de la migración en el estado de Hidalgo, basándose en técnicas descriptivas y predictivas de Minería de Datos.

Para poder cumplir con dicho objetivo se plantearon los siguientes objetivos específicos:

- Integrar y recopilar los datos útiles para el estudio de los factores demográficos y sociales causados por la migración.
- Limpiar y transformar los datos a un formato común, detectando y resolviendo inconsistencias en ellos.
- Crear las vistas minables necesarias para la aplicación de técnicas de Minería de Datos.
- Aplicar algoritmos de agrupamiento a los conjuntos de datos.
- Evaluar los resultados de los algoritmos de agrupamiento, para elegir los que tengan un mejor desempeño.
- Aplicar algoritmos de clasificación a los agrupamientos elegidos.
- Encontrar un conjunto de reglas que describa los factores demográficos y sociales de los migrantes en el estado de Hidalgo.
- Interpretar los perfiles con ayuda del experto.

Alcance

Se aplicaron técnicas de Minería de Datos para procesar información referente al problema de migración del estado de Hidalgo. Se abordó y analizó el problema desde dos perspectivas diferentes, pero relacionadas entre sí:

- La situación en que se encuentran las familias de los migrantes del estado de Hidalgo, principalmente las condiciones de sus viviendas (factor demográfico).

- El perfil de los migrantes del estado, identificando grupos con diferente patrón de comportamiento (factor social).

A partir de estos resultados se encontraron características que permiten hacer predicciones relacionadas al problema de migración en el estado de Hidalgo.

Limitaciones

Debido a la naturaleza del problema social tomado como caso de estudio, los datos disponibles acerca del fenómeno de migración son escasos; por lo que el presente proyecto está limitado a la información de los migrantes que viajaron a la frontera Norte del país y cruzaron a Estados Unidos (EMIF-Norte) y de la información de las viviendas que dejaron en el estado (Censo INEGI 2010).

Estructura del documento

El presente documento se encuentra dividido en cuatro capítulos que muestran el desarrollo de la investigación durante la estancia en el programa de Maestría en Ciencias Computacionales.

El Capítulo 1, Marco Teórico, aborda los conceptos fundamentales de Minería de Datos, incluyendo la parte teórica de algunas técnicas de clustering y clasificación utilizadas en este proyecto. En este capítulo, también se presenta la descripción de algunas de las metodologías de Minería de Datos existentes que sirven como guía en el desarrollo de la solución propuesta.

El Capítulo 2 presenta el Estado del arte, el cual está compuesto por trabajos que resuelven problemas relacionados a la migración utilizando técnicas estadísticas tradicionales y por trabajos que utilizan la Minería de Datos para resolver diferentes problemas sociales.

Los Capítulos 3 y 4 describen los pasos que se siguieron en la metodología elegida para llevar a cabo el análisis de la información de los migrantes en el estado de Hidalgo, de cada factor a analizar: demográfico y social, respectivamente. De cada factor, se muestran los experimentos realizados en cada paso y los resultados obtenidos.

Finalmente, se presenta una sección de Conclusiones del trabajo de investigación, así como el Trabajo Futuro que se ha planteado resolver.

Capítulo 1

Marco Teórico

El objetivo principal de la Minería de Datos es encontrar patrones y relaciones entre grandes volúmenes de datos, con la finalidad de crear modelos que sean abstracciones de la realidad [70]. Dada esta concepción y la gran diversidad de modelos que puede crear, la Minería de Datos hoy en día se aplica en diferentes áreas como la astronomía, educación, aspectos climatológicos, medicina, industria, telecomunicaciones, manufactura, mercadotecnia, detección de fraudes, análisis de mercado, etc. [68].

En particular, en Medicina la Minería de Datos se ha usado para diagnosticar y estudiar diversos padecimientos tales como arritmias o diabetes en pacientes con diferentes características [67, 69], o bien para analizar resultados de estudios médicos, por ejemplo, se ha llevado a cabo un estudio de mastografías para predecir cáncer de mama [54]. Mercadotecnia ha utilizado la Minería de Datos para analizar el comportamiento de clientes en instituciones bancarias [52, 74] y en aplicaciones de E-business [7]. En Educación, las técnicas de esta disciplina han ayudado a analizar diversos factores que benefician al mejoramiento de esta área; ha contribuido por ejemplo a descubrir como las personas aprenden, a predecir cuáles son las mejores técnicas de aprendizaje y a entender el comportamiento real de estudiantes [9, 1, 14]. En Telecomunicaciones, la Minería ha analizado el comportamiento que presentan las comunicaciones para proveer servicios personalizados a los usuarios de una red [47, 59] o bien ha estudiado los hábitos de uso de Internet entre estudiantes de algunas universidades [39]. En astronomía, se han aplicado técnicas para clasificación de objetos astronómicos como galaxias y estrellas, clasificación por los tipos morfológicos y las edades [23, 27]. De la misma

manera, la Minería de Datos ha sido aplicada exitosamente a la solución de diferentes problemas sociales, como el estudio de diversas poblaciones con el fin de analizar el comportamiento de sus integrantes [33, 20, 2].

La arquitectura de un sistema de minería de datos típico [34] puede tener los siguientes componentes principales:

1. Bases de datos, almacén de datos u otro repositorio: Conjunto de datos dispuestos en bases de datos, almacenes de datos o algún otro tipo de repositorio de información.

2. Base de datos o almacén de datos del servidor: Es el servidor de bases de datos o almacén de datos que se encarga de buscar los datos pertinentes, de acuerdo a la petición de minería de datos que realice el usuario.

3. Base de conocimiento: Es el conocimiento de dominio que se utiliza para guiar la búsqueda o evaluar el grado de interés de los patrones resultantes.

4. Motor de Minería de Datos: Es un componente esencial para un sistema de minería de datos, puede estar compuesto de un conjunto de módulos para tareas tales como la asociación, clasificación, análisis de conglomerados, etc.

5. Módulo de evaluación del patrón: Este componente puede emplear medidas de intereses con el fin de filtrar los patrones descubiertos y centrarse en los patrones interesantes.

6. Interfaz gráfica de usuario: Se encarga de comunicar a los usuarios con el sistema de extracción de datos, permitiendo al usuario interactuar con el sistema, haciendo consultas de minería de datos o algunas tareas como la navegación entre la base de datos o almacén de datos, evaluación de los patrones extraídos y visualizar los patrones en diferentes formas.

Esta disciplina utiliza diversos algoritmos que cumplen con diferentes tareas. Los algoritmos examinan los datos y determinan un modelo, dicho modelo debe ser lo más fiel posible a las características de los datos examinados [21]. Los algoritmos de minería de datos constan de tres partes:

- Modelo: El objetivo del algoritmo es ajustar un modelo a los datos.

- Preferencia: Algunos criterios se deben utilizar para ajustar un modelo sobre otro.
- Búsqueda: Todos los algoritmos requieren alguna técnica para buscar los datos.

Los modelos resultantes de aplicar Minería de Datos, pueden ser de dos tipos, modelos predictivos y modelos descriptivos. El modelo predictivo utiliza resultados conocidos de diferentes datos para hacer una predicción, además se puede basar en el uso de datos históricos. Las tareas predictivas de minería de datos incluyen la clasificación, regresión, análisis de series de tiempo y predicción. El modelo descriptivo identifica patrones o relaciones entre los datos; explora las propiedades de los datos examinados. Algunas tareas de Minería de Datos que tienen un enfoque descriptivo son el agrupamiento, el resumen y las reglas de asociación. En ambos modelos es necesaria la ayuda de un experto en el tema para que los valide y les dé un significado [21].

El diagrama de la Figura 1.1 muestra los tipos de modelos y las tareas de Minería de Datos más comunes.

Aún cuando la Minería de Datos es una herramienta poderosa, existen algunos problemas de aplicación asociados a ella [21], entre los cuales se encuentran: la interacción humana, ya que es necesaria la intervención de los expertos técnicos para la interpretación de los resultados; y el overfitting, el cual ocurre cuando el modelo no se ajusta a estados futuros de la base de datos y los outliers, que ocurre cuando se encuentran muchas entradas de datos que no encajan con el modelo derivado. Otros problemas de aplicación comunes, son los siguientes:

- Interpretación de los resultados
- Visualización de los resultados
- Grandes conjuntos de datos
- Alta dimensionalidad
- Datos multimedia
- Datos faltantes

Figura 1.1: Modelos y tareas de Minería de Datos.

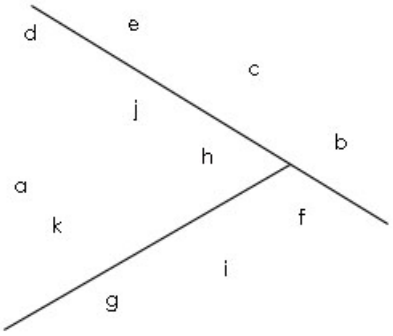
- Datos irrelevantes
- Datos ruido
- Modificación de los datos
- Integración
- Aplicación

1.1. Agrupamiento

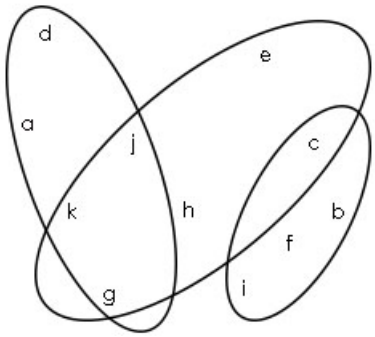
Agrupamiento (o clustering en inglés) se le conoce también como aprendizaje no supervisado, debido a que el conjunto de entrenamiento no tiene definida una partición apriori y por lo tanto los objetos no tienen asignada una etiqueta de clase. Este modelo divide o particiona los datos en clusters o grupos que pueden o no ser disjuntos. Los grupos se van formando considerando la similitud o la distancia, dependiendo del algoritmo, a manera que los individuos más semejantes formen un grupo. En el clustering es muy común que sea necesaria la ayuda de un experto de domino para interpretar el significado de los grupos creados, esto como consecuencia de que los grupos no están predefinidos [21].

Comúnmente, los clusters se representan mediante un diagrama que muestre como la instancia cae dentro del grupo, así como se muestra en la Figura 1.2 [73]. En el más simple de los casos se tendría una instancia asociada con un número de cluster, lo cual se puede representar sobre un plano de dos dimensiones y particionando el espacio para mostrar cada grupo, un ejemplo es como el que se muestra en la Figura 1.2(a). En algunas ocasiones los algoritmos permiten que una instancia pertenezca a más de un grupo, por lo que el diagrama de representación de los grupos dibuja subconjuntos superpuestos para representarlos, como un diagrama de Venn, tal como se ilustra en la Figura 1.2(b). Algunos algoritmos agrupan las instancias probabilísticamente y no de manera categórica; por ejemplo, la instancia siempre tiene una probabilidad o un grado de pertenencia con el que pertenece a cada uno de los clusters, como se muestra en la Figura 1.2(c). Otros algoritmos producen una estructura jerárquica de grupos, que son llamadas dendogramas, este término significa diagramas de árbol y se deriva del vocablo griego dendron que significa “un árbol”, un ejemplo de éste se muestra en la Figura 1.2(d).

La interpretación de los grupos se puede realizar a través de su descripción conceptual, que es un predicado lógico de las características de los individuos, siendo verdadera para los individuos que pertenecen al grupo y falsa para los que están fuera de él. La descripción conceptual se puede formalizar encontrando descripciones breves y claras de los grupos, manteniendo los falsos errores positivos y los falsos negativos bajos, tanto como sea posible. Los falsos errores positivos, se refiere a los individuos que no pertenecen al grupo, pero que cumplen con su descripción y los falsos negativos son los individuos



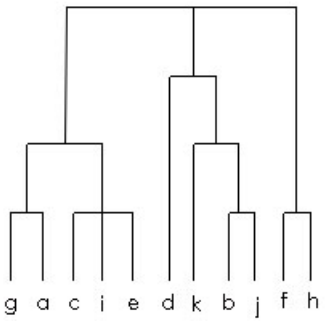
(a) Partición



(b) Diagrama de Venn

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

(c) Tabla



(d) Dendograma

Figura 1.2: Representación de resultados de clustering.

que si pertenecen al grupo, pero que no cumplen con la descripción [11].

El clustering es una técnica útil que permite la generación de clases en poblaciones donde no se cuenta con una clasificación previa; una vez obtenidas las clases, se pueden determinar patrones de la población y profundizar en las características de los individuos que integran cada grupo. Estos patrones y características ayudan a entender el comportamiento de la población, y de esta forma poder hacer una toma de decisiones en base a los modelos generados.

1.1.1. Tipos de algoritmos de clustering

La manipulación de grandes volúmenes de información, como son datos financieros, de telecomunicaciones, de medicina, imágenes, etc., persigue di-

ferentes objetivos, entre los cuales se encuentra la identificación de grupos (agrupamiento) a través de la aplicación de algunos métodos como algoritmos jerárquicos, particionales y basados en densidad.

En los algoritmos de tipo jerárquico, se toma al conjunto de datos y se particiona por niveles, por lo general, en cada nivel se fusionan o se dividen dos grupos según el tipo de algoritmo, ya sea aglomerativo o divisivo [50].

En los algoritmos particionales la división de los datos se realiza tomando en cuenta un conocimiento previo a la cantidad de grupos, moviendo a los individuos al grupo que les corresponde, satisfaciendo así una función objetivo. Un tipo de algoritmo particional es el $k - means$, que consiste en definir k centros (uno para cada grupo) y luego tomar cada objeto de entrenamiento y situarlo en el grupo de su centro más cercano, posteriormente se recalculan los centros y se vuelven a distribuir todos los objetos según el centro más cercano. Este proceso se repite hasta que ya no hay cambio en los grupos formados [22].

Los algoritmos basados en densidad toman en cuenta la distribución de los individuos, de tal forma que los grupos que se formen tengan una alta densidad entre sus individuos, y baja densidad entre grupos. Uno de los algoritmos más destacados de este tipo es el DBSCAN; éste comienza con la selección de un objeto p arbitrario y en el caso de que p sea un objeto central se construye un grupo con todos los individuos denso-alcanzables por él, en caso contrario se toma otro objeto del conjunto de datos; este proceso se realiza hasta que todos los objetos hayan sido procesados; este algoritmo considera los conceptos de ruido y puntos borde. Los puntos ruido son aquellos puntos que se quedan fuera de los grupos formados y los puntos borde son aquellos que no son puntos ruido ni puntos centrales [50].

Existen algunos algoritmos que son considerados híbridos, ya que fusionan los distintos métodos; entre ellos se encuentra el algoritmo Chameleon y el algoritmo CURE. El primero de ellos se conforma de dos fases; en la primera se crea un grafo de los k vecinos más cercanos y se aplica un algoritmo de particionamiento para generar subgrupos, en la segunda fase se utiliza un algoritmo jerárquico aglomerativo donde se combinan dichos subgrupos, considerando la interconectividad y cercanía hasta encontrar los grupos correctos. Este algoritmo tiene la capacidad de adaptarse a los cambios de las

características internas de los subgrupos y se pueden fusionar varios subgrupos en un mismo paso. Por su parte, el algoritmo CURE es una combinación entre los algoritmos jerárquicos y los particionales. Su proceso consiste en obtener un número c de objetos representativos del grupo, seleccionando los c objetos más dispersos del grupo y atrayéndolos hacia el centro por un factor de contracción. En cada iteración se fusionan los dos grupos más cercanos y se vuelve a calcular el nuevo centro del grupo formado y los c objetos representativos.

A continuación se describen brevemente los algoritmos de agrupamiento SOM, EM, MDBC y K-means, los cuales son utilizados a lo largo de este proyecto.

1.1.2. Algoritmo SOM (Self- Organizing Maps)

Este algoritmo también es conocido como Red Kohonen, debido a su creador Teuvo Kohonen. Es un algoritmo para el agrupamiento y la visualización de los datos, que toma el enfoque de una red neuronal, ya que está formado por neuronas organizadas en un mapa [8].

Se llama mapa al espacio enrejado de dos dimensiones, donde es asignada la capa de neuronas de salida; esto permite dividir los datos de entrada en grupos que sean similares. El mapa se describe como un grafo (C, r) donde C es un conjunto de neuronas interconectadas con una topología discreta definida por r y la distancia entre cada par de neuronas impone una relación de vecindad.

Como entrada se toma un conjunto de vectores de muestra etiquetados y da como salida un conjunto de neuronas con las etiquetas de los vectores de entrada conectados a estas neuronas.

El entrenamiento de la red consiste en ajustar iterativamente los pesos de conexión de la entrada a la salida [44]; es decir, en cualquier etapa de iteración, ajusta los pesos de la neurona ganadora, con la finalidad de que se vuelva más similar al patrón de entrada; se denomina neurona ganadora, a la neurona que esté más cercana al patrón de entrada. De manera simultánea, en las iteraciones iniciales, el conjunto de vecinos de la neurona ganadora ajustan sus pesos de forma similar. Es común utilizar una medida de radio

para definir el tamaño de la vecindad, haciendo una analogía a una ciudad, para una rejilla cuadrada, el radio sería la distancia de las cuadras al centro. Después de un número suficiente de épocas los pesos se agruparán, a manera que la red de neuronas de salida constituyan un mapa topológico de las entradas, de ahí el nombre de Self- Organizing Map (Mapa de Auto-Organización).

El rendimiento del mapeo se evalúa por una medida de error promedio, para todos los patrones, la distancia de cada patrón a partir de la neurona de salida ganadora.

1.1.3. Algoritmo EM

Es conocido como el algoritmo de Expectativa de Maximización, ya que sigue una secuencia de etapas de Estimación (E) y Maximización (M) [11]. Este algoritmo inicia con conjeturas de parámetros que permiten calcular de cada objeto, las probabilidades de pertenecer a un grupo y después utiliza esas probabilidades para estimar los parámetros de inicio, este procedimiento lo realiza de manera iterativa [73].

La expectativa se refiere al primer paso, cuando el algoritmo calcula los valores esperados de la clase con las probabilidades del clúster. En el segundo paso, cuando se calculan los parámetros de distribución, se alcanza la maximización de la probabilidad de las distribuciones de los datos dados.

Estas probabilidades actúan como pesos; si w_i es la probabilidad de la instancia i de pertenecer al grupo A, la media y la desviación estándar para el grupo A serían como en la ecuación 1.1 y en la ecuación 1.2, respectivamente.

$$\mu_A = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n} \tag{1.1}$$

$$\sigma_A^2 = \frac{w_1 (x_1 - \mu)^2 + w_2 (x_2 - \mu)^2 + ... + w_n (x_n - \mu)^2}{w_1 + w_2 + ... + w_n} \tag{1.2}$$

Donde x_i son todos los casos, no sólo aquellos que pertenecen al grupo A.

El algoritmo se detiene cuando los valores de los parámetros actuales se aproximan o coinciden con los valores anteriores.

1.1.4. Make Density Based Clusterer (MDBC)

Este algoritmo crea los grupos con la idea de la agrupación humana natural, basándose en las propiedades de densidad. Por esta razón, los grupos se detectan fácilmente debido a su alta densidad con respecto a los objetos que posee. Los grupos se componen de puntos centrales, los cuales se ubican en la región interna del grupo; y de dos puntos fronterizos, los cuales se sitúan en el borde del grupo [29]. Este algoritmo se puede usar cuando los grupos son irregulares, con ruido y cuando se encuentran valores atípicos [60].

Este algoritmo utiliza otro algoritmo de clustering para que regrese la distribución y la densidad; además se adapta a las distribuciones normales y las distribuciones discretas dentro de cada grupo producido por este segundo algoritmo.

Los pasos que realiza Make Density Based Clusterer son los siguientes:

1. Calcula la $\in$-vecindad para todos los objetos en el espacio de datos.
2. Selecciona un núcleo objeto CO.
3. Para todos los objetos $co \in CO$, agrega los objetos al CO que están conectados con la densidad co. Continúa hasta que los más lejanos se hayan encontrado.
4. Se repiten los pasos 2 y 3 hasta que se hayan procesado todos los objetos del núcleo.

1.1.5. K-Means

Es un algoritmo basado en partición, ya que asigna un valor a k que representa el número de puntos iniciales que serán elegidos para representar los centros de los clusters iniciales; todas las instancias son asignadas al centro del grupo que esté más cercano, después se calcula el nuevo centro del grupo considerando los valores de todos los objetos que se asignaron a cada grupo. Este proceso es iterativo y continúa hasta que ya no hay cambios en los grupos [73].

Existen muchas variantes de este algoritmo que dependen de la elección de los centros iniciales de los grupos, el cálculo del centro (que generalmente

se calcula con la media de los objetos del grupo) y los criterios de parada [44].

Para este algoritmo es necesario conocer el número de grupos con anticipación, en caso de que no sea así, se puede solucionar probando diferentes opciones y ver cuál es la mejor. Una estrategia simple es partir de un mínimo determinado, por ejemplo $k = 1$, e ir incrementando el valor de k hasta un máximo fijo, se puede utilizar validación cruzada para encontrar el mejor valor, pero esto en muchas ocasiones no resulta factible ya que k-means es lento y la validación cruzada lo vuelve aún más lento. Otra opción es encontrar pocos clusters y determinar si vale la pena la división de ellos, por ejemplo iniciar con k=2 y considerar la división de cada grupo; en este caso se reduce el tiempo computacional de manera considerable, si la división de los dos grupos es irrevocable.

1.2. Medidas de Semejanza

Los grupos obtenidos, después de aplicar algún algoritmo de clustering, deben tener ciertas propiedades, entre las más importantes se encuentra la propiedad de que un objeto o tupla dentro de un clúster sea más similar a los objetos o tuplas del mismo cluster que a los que estén fuera de él [21]; es decir, cuando los patrones x y y sean similares la función de semejanza $f(x, y)$ presentará un valor grande [30].

Por tanto, una medida de similitud, $sim(t_i, t_l)$, es la semajanza entre dos objetos o tuplas, t_i, $t_l \in D$; dada una base de datos $D = \{t_1, t_2, ..., t_n\}$ de tuplas u objetos [21].

En ocasiones es necesario calcular la distancia entre los grupos (en lugar de los elementos), para ello existen varias alternativas que permiten hacer dicho cálculo [21], entre las cuales se encuentran las siguientes:

- Single link: La distancia más pequeña entre un elemento de un grupo y un elemento en otro grupo.
- Complete link: La distancia mayor entre un elemento de un grupo y un elemento en otro grupo.
- Promedio (average): Distancia media entre un elemento en un grupo y un elemento en otro grupo.

- Centroide: Si los clústers tienen un centroide representante, entonces la distancia centroide se define como la distancia entre los centroides.
- Medoid: Usando un medoid para representar cada clúster, la distancia entre el clúster puede ser definida por la distancia entre los medoides.

El centroide, representado en la ecuación 1.3, es el centro de la agrupación y no necesariamente tiene que ser un objeto real en el clúster; el medoide es el objeto situado en el centro del cluster y el radio, representado en la ecuación 1.4, es la raíz cuadrada del promedio de la distancia al cuadrado, de cualquier punto en el cluster al centroide [21].

$$Centroide \quad = \quad C_m = \frac{\sum_{i=1}^{N} (t_{mi})}{N} \tag{1.3}$$

$$Radio \quad = \quad R_m \quad = \quad \sqrt{\frac{\sum_{i=1}^{N} (t_{mi} \quad - \quad C_m)^2}{N}} \tag{1.4}$$

Otra forma de calcular la semejanza entre dos vectores de características u objetos, es midiendo la distancia entre ellos; para ello son empleadas comúnmente la distancia Euclidiana, distancia de Manhattan o City-block y la distancia de Minkowski, que calculan la distancia entre objetos descritos por atributos numéricos [30].

1.2.1. Distancia Euclidiana

La distancia Euclidiana, mostrada en la ecuación 1.5, es la medida más popular que se utiliza para el cálculo de la disimilitud [34], ya que calcula la distancia mínima posible entre dos vectores y toma su valor mínimo $d_0 = 0$ cuando dichos vectores coinciden [40].

$$d_e(x, y) = \sqrt{\sum_{i=1}^{l} (x_i - y_i)^2} \tag{1.5}$$

Dónde:

- $x, y \in X$

- X son objetos descritos por l atributos
- l es el número de atributos, y
- x_i, y_i son el i-ésimo atributo de x y y.

1.2.2. Distancia de Manhattan

Otra medida muy utilizada es la distancia de Manhattan, también conocida como City-block o de cuadras (Ecuación 1.6), ya que calcula la distancia entre dos objetos por bloques, haciendo la analogía de las cuadras de una ciudad [34].

$$d_{CB} = \sum_{i=1}^{l} |x_i - y_i| \tag{1.6}$$

Dónde:

- $x, y \in X$
- X son objetos descritos por l atributos
- l es el número de atributos, y
- x_i, y_i son el i-ésimo atributo de x y y.

Esta distancia y la Euclidiana son consideradas métricas, ya que satisfacen las siguientes propiedades:

- No negatividad: $d(x, y) \geq 0$ La distancia entre dos puntos es un número no negativo.
- Identidad de los indiscernibles: $d(x, x) = 0$: La distancia de un objeto a sí mismo es 0.
- Simetría: $d(x, y) = d(y, x)$ La distancia es una función simétrica.
- Desigualdad triangular: $d(x, y) \leq d(x, z) + d(z, y)$

1.2.3. Distancia de Minkowski

La distancia de Minlowski es una generalización de la distancia Euclidiana y la distancia de Manhattan [71]. La distancia de Minkowski de orden m, se define con la ecuación 1.7.

$$d_M = \sqrt[m]{\sum_{i=1}^{l} (|x_i - \quad y_i|)^m} \tag{1.7}$$

Dónde:

- $x, y \in X$
- X son objetos descritos por l atributos
- l es el número de atributos
- x_i, y_i son el i-ésimo atributo de x y y, y
- m es el orden de la distancia.

Cuando la distancia de Minkowski es de primer orden, es la misma que la métrica de Manhattan y cuando es de segundo orden es igual a la distancia Euclidiana [71].

1.2.4. Coeficiente de Similaridad de Gower

El coeficiente de similaridad de Gower es una medida de semejanza que permite establecer el grado de similitud entre individuos, cuando se requiere que sean medidos por sus características cualitativas, cuantitativas (continuas y discretas) y binarias; ya que admite la utilización simultánea de estas variables mixtas [15].

Este coeficiente se define con la ecuación 1.8 [18].

$$S_G(x, y) = \frac{\sum_{i=1}^{l} W_{xyi} S_{xyi}}{\sum_{i=1}^{l} W_{xyi}} \tag{1.8}$$

Dónde:

- $S_G(x, y)$ Es el coeficiente de similitud de Gower entre el individuo x y el individuo y
- S_{xyi} Es una medida parcial de similitud entre los individuos x y y al ser comparados con base en la variable $i - ésima$
- W_{xyi} Es un ponderador de la comparación entre los individuos x y y en la $i - ésima$ variable

Para las variables binarias $S_{xyi} = 1$ para las coincidencias y $S_{xyi} = 0$ para las divergencias. $W_{xyi} = 0$ para dobles ausencias y $W_{xyi} = 1$ en los demás casos.

Para las variables cualitativas $S_{xyi} = 1$ para las coincidencias y $S_{xyi} = 0$ para las divergencias. $W_{xyi} = 1$ en todos los casos.

Si la variable es cuantitativa, se calcula S_{xyi} con la ecuación 1.9.

$$S_{xyi} = 1 - \frac{|x_i - y_i|}{R_i} \tag{1.9}$$

Dónde x_i y y_i son los valores de la $i - ésima$ variable observados en los individuos x y y, respectivamente. R_i es el rango de la $i - ésima$ variable; $W_{xyi} = 1$ en todos los casos, sin importar el tipo de variable; sin embargo, $W_{xyi} = 0$ cuando falte al menos uno de los dos valores involucrados en la comparación.

A través de la equivalencia mostrada en la ecuación 1.10 se puede obtener la distancia de Gower, una vez obtenido el coeficiente de semejanza de Gower [15].

$$d_G(x, y)^2 = 1 - S_G(x, y) \tag{1.10}$$

Por lo tanto, la distancia de Gower $D_G(x, y)$ se obtiene con la ecuación 1.11.

$$d_G(x, y) = \sqrt{1 - S_G(x, y)} \tag{1.11}$$

1.3. Índices de Validación de Clusters

Después de haber procesado los datos con algún algoritmo de agrupamiento, es necesario evaluar las estructuras o particiones obtenidas y seleccionar la que mejor se ajuste a los datos [48], ya que no existe un algoritmo que sea el mejor en todas las situaciones. Esto se debe a que diferentes algoritmos e inclusive diferentes configuraciones de un mismo algoritmo, producen estructuras de agrupamiento diferentes, además de que existen algoritmos que no pueden determinar el número de grupos de manera natural en los datos, por lo que necesitan que se les proporcione un valor para un parámetro (generalmente conocido como k) que indique el número de grupos a formar. Sin embargo, el conocer este valor de k no es muy común, por lo que habitualmente se ejecuta el algoritmo varias veces con diferentes valores de este parámetro en cada ejecución.

La validación de los clusters se refiere a los procedimientos que evalúan los resultados que se obtuvieron del clustering o agrupamiento; dicha evaluación tiene que ser objetiva y de manera cuantitativa [37], generalmente se obtiene por medio de algún tipo de medida de disimilitud dentro de la agrupación [44].

Por tanto, un índice de validez de clúster o CVI (por sus siglas en inglés, Cluster Validity Index) mide la adecuación de las estructuras obtenidas del clustering a manera que se puedan interpretar objetivamente; dicha adecuación se refiere a que las estructuras proporcionen información verdadera sobre los datos [37].

Existen tres criterios que sirven para expresar la validez de la estructura de la agrupación:

- Criterio externo: Hace coincidir una estructura de agrupación con información a priori.
- Criterio interno: Evalúa la adecuación entre la estructura y los datos, utilizando sólo los datos en sí.
- Criterio relativo: Decide cuál de las dos estructuras es mejor en algún sentido, como ser más estable o más adecuado para los datos.

1.3.1. Índice de Davies-Bouldin

Dentro de los índices más utilizados para la comparación de CVI, se encuentra el índice de Davies Bouldin, que calcula la cohesión basado en la distancia desde los puntos de un cluster a su centroide y la separación basada en la distancia entre los centroides; un valor menor de ese índice indica una mejor partición o una mejor estructura de agrupamiento [48].

La fórmula de este índice se muestra en la ecuación 1.12.

$$DB(C) = \frac{1}{K} \sum_{C_k} \underset{C_l \in C \setminus C_k}{max} \left\{ \frac{S\left(C_k\right) + S\left(C_l\right)}{d_G\left(\overline{C_k}, \overline{C_l}\right)} \right\} \tag{1.12}$$

Dónde:

- K es el número de clústers
- C_k representa el clúster k
- $\overline{C_k}$ es el centroide del cluster k item d_e calcula la distancia entre dos puntos, y
- $S(C_k) = 1/|C_k| \sum_{x_i \in C_k} d_e(x_i, \overline{C_k})$, dónde X_i es el individuo i que pertenece al cluster k y $|C_k|$ es el número de individuos en el cluster k

1.4. Clasificación

Cualquier sistema de clasificación de patrones se basa en lo siguiente: dado un conjunto de objetos (que comúnmente se dividen en dos: conjunto de entrenamiento y conjunto de prueba o test) representados por pares <atributo, valor>, el problema consiste en encontrar una función $f(x)$ (llamada hipótesis) que clasifique dichos objetos.

Existen diversos algoritmos que permiten resolver esta tarea de clasificación. Los algoritmos utilizados para el desarrollo de este trabajo son descritos a continuación.

1.4.1. Árboles de Decisión

El aprendizaje de árboles de decisión está englobado como una metodología del aprendizaje supervisado [44]. Consiste en una estructura de árbol formada por nodos (internos y hojas) en el que cada nodo interno representa una elección entre varias alternativas, y cada nodo hoja representa una clasificación o una decisión [73]. Una estructura simple de árbol de decisión se puede observar en la Figura 1.3.

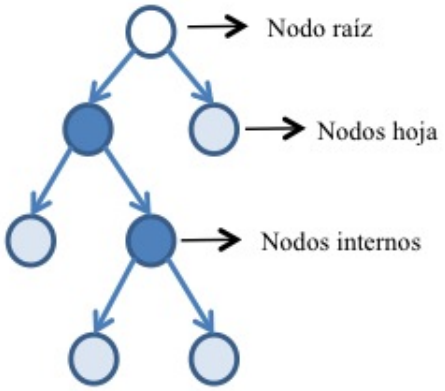


Figura 1.3: Representación de un árbol de decisión.

Un árbol de decisión puede interpretarse esencialmente como una serie de reglas compactadas para su representación en forma de árbol, donde cada camino de la raíz a una de las hojas del árbol es una regla.

Entre los algoritmos más conocidos de este paradigma se encuentran ID3 y C4.5 [55]. El procedimiento general para generar un árbol de decisión consiste en seleccionar del conjunto de entrenamiento un atributo como raíz del árbol y crear una rama con cada uno de los posibles valores de dicho atributo, en caso de que el atributo elegido sea de tipo categórico; o bien se crea una partición binaria del tipo $X \leq V$ para atributos de tipo numérico, donde V es un valor considerado como punto de corte para el resto de los valores del atributo en proceso. Para cada nuevo nodo formado en el árbol de decisión se realiza el mismo proceso, es decir, se selecciona otro atributo y se genera una nueva rama para cada posible valor del atributo seleccionado. Cada nodo que se desea expandir debe tener objetos con dos o más

clases diferentes, en caso contrario, se considera un nodo final y se convierte en nodo hoja, con la etiqueta de clase de los objetos que contiene dicho nodo.

En cada nodo del árbol de decisión se debe seleccionar un atributo para seguir dividiendo, y para encontrarlo se toma como referencia la Teoría de la Información, la cual está basada en la entropía que se encarga de medir la cantidad de información en un atributo. Entre más pequeño sea el valor de la entropía, menor será la incertidumbre y más útil será el atributo para la clasificación. Una de las medidas más utilizadas para realizar esta tarea en los algoritmos de árboles de decisión es la Proporción de Ganancia de Información (Gain Ratio).

La clasificación de un objeto nuevo del que se desconoce su clase se hace con la misma técnica, iniciando su recorrido en el nodo raíz y descendiendo por el árbol entre los nodos internos hasta llegar a una hoja. Cuando una hoja sea alcanzada, al atributo clase de ese objeto, cuyo valor se desconoce, se le asigna la etiqueta de dicha hoja.

1.4.2. Naïve Bayes

Clasificador estadístico que puede predecir la probabilidad de que un objeto pertenezca a una clase en particular [71]. La clasificación bayesiana se basa en el teorema de Bayes: "A partir de que ha ocurrido el suceso B (ha ocurrido un accidente) deducimos las probabilidades del suceso A (¿estaba lloviendo o hacía buen tiempo?)" .

La idea de usar el Teorema de Bayes en cualquier problema de aprendizaje automático (en especial los de clasificación) es que se pueden estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjunto de datos de entrenamiento para así escoger la hipótesis más probable. Lo que normalmente se busca encontrar en el aprendizaje es la mejor hipótesis dados los datos [21]. Típicamente, un objeto O es representado por un conjunto de atributos $(x_1, \ldots, x_n)$, donde x_i es el valor del atributo X_i. Se denota C la variable de clasificación, y c son los valores de C. Para estimar la hipótesis más probable (MAP, Maximum a Posteriori Hipotesis) se busca el mayor P(c—O), como se muestra en la Eq. 1.13.

$$h_{MAP} = \underset{c \in C}{argmax} \left(\frac{p(O|c)p(c)}{p(O)} \right) \tag{1.13}$$

1.4.3. Redes Neuronales

Las redes neuronales constituyen una forma diferente de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender patrones complejos y características dentro de los datos [5]. Se comportan de forma parecida al cerebro humano aprendiendo de la experiencia y del pasado, y aplicando tal conocimiento a la solución de problemas nuevos.

El elemento básico de un sistema neuronal biológico es la neurona y este sistema se encuentra formado por millones de éstas organizadas en capas. En el sistema neuronal artificial se puede establecer una estructura jerárquica similar a la existente en el cerebro, en este caso el elemento esencial será la neurona artificial, la cual se organiza en capas, y varias capas constituirán una red neuronal. Finalmente una red neuronal junto con las interfaces de entrada y salida constituirán el sistema global de proceso. Se puede definir una red neuronal artificial como un grafo dirigido con las siguientes propiedades:

- A cada nodo (neurona) i se le asocia un conjunto de variables de estado $(X_1, \ldots, X_n)$.
- A cada conexión (i, j) entre los nodos (neuronas) i y j se les asocia un peso w_{ij}.
- A cada nodo (neurona) i se le asocia un umbral θ_i.
- Para cada nodo i se define una función $f_i(x_1, \ldots, x_n, w_{ij}, \ldots, w_{in}, \theta_i)$. El valor de esta función proporciona el nuevo estado de la neurona.

La Figura 1.4 muestra una representación de una neurona artificial, donde se pueden apreciar los elementos mencionados anteriormente.

El entrenamiento de una neurona se lleva a cabo cuando recibe las entradas o estímulos de otras y los procesa para producir una salida que transmite a la siguiente capa de neuronas. La señal de salida tendrá una intensidad, fruto de la combinación de la intensidad de las señales de entrada y de los

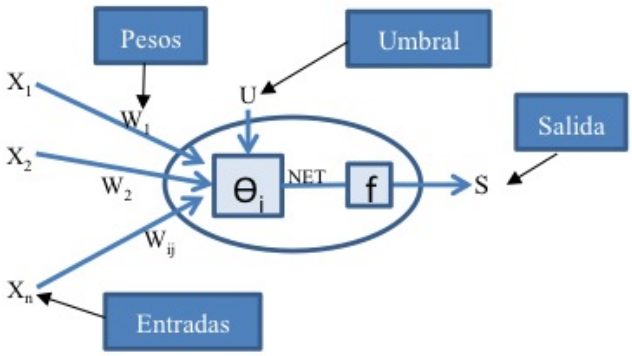


Figura 1.4: Representación de una neurona artificial.

pesos que las transmiten. Los pesos o dendritas tienen un valor distinto para cada par de neuronas que conectan pudiendo así fortalecer o debilitar la conexión o comunicación entre neuronas particulares. El objetivo de este método es ajustar los pesos durante el proceso de entrenamiento, de manera que la capa de salida sea capaz de clasificar correctamente los objetos de entrada.

1.4.4. K-Vecinos más cercanos

La idea básica en la que se fundamenta este método es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus k vecinos más cercanos [22].

Este método se fundamenta en una idea muy simple e intuitiva. Teniendo un conjunto de entrenamiento N formado por s objetos descritos por $(X_1, \ldots, X_n)$ atributos y un atributo a predecir, que será la clase $C = c_1, ..., c_m$, para clasificar un nuevo objeto, denotado por $o = (x_1, \ldots, x_n)$, se deben calcular todas las distancias de los objetos de entrenamiento con el nuevo objeto a clasificar. Una vez calculadas todas las distancias se seleccionan las k distancias más cortas, y el nuevo objeto se clasificará en la clase más común entre los objetos que corresponden a esas distancias. La Figura 1.5 muestra gráficamente un ejemplo, donde si $K = 3$, el objeto nuevo (punto negro) será clasificado a la clase triángulo.

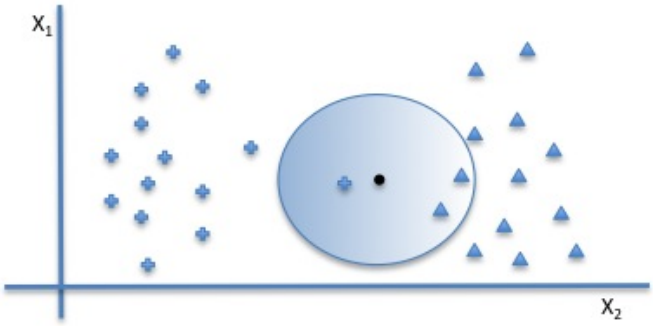


Figura 1.5: Representación del método K-Vecinos más Cercanos

1.4.5. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (SVM por sus siglas en inglés, Support Vector Machines) son un enfoque distintivo de patrones de clasificación y regresión [44]. La idea central de SVM es el ajuste de una función discriminante para que utilice óptimamente la información de separabilidad de los objetos. Este algoritmo asume que se posee una función lineal discriminante y dos clases linealmente separables de los valores objetivo +1 y -1.

SVM utiliza la siguiente metodología:

- Mapear puntos de entrenamiento a un espacio dimensional.
- Construir un hiperplano de margen máximo que separe los puntos en las clases respectivas.
- Clasificar un punto nuevo de acuerdo a su ubicación con respecto al hiperplano de separación.

Para el caso de dos clases, los x_k son objetos de entrenamiento, $k = 1, \ldots, n$ los cuales tienen un atributo z_k, el cual determina la clase, $z_k \in -1, 1$. Los vectores de soporte serán los objetos mas próximos de cada clase, al hiperplano de margen máximo. El conjunto de vectores de soporte definen a este hiperplano, los demás vectores (objetos) son irrelevantes. Una representación gráfica de el conjunto de vectores de soporte y el hiperplano de margen máximo se presenta en la Figura 1.6. Para obtener los vectores de soporte con clases no linealmente separables se realiza una transformación no lineal del espacio de entrada.

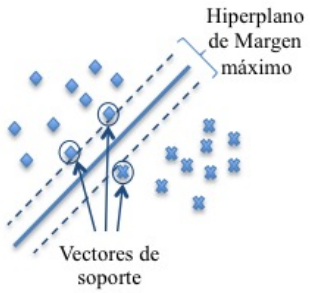


Figura 1.6: Representación del método SVM

1.5. Algoritmo LR-FIR

El algoritmo de Reglas Lingüisticas en FIR, LR-FIR (por sus siglas en inglés, Linguistic Rules in FIR) [13], extrae reglas de predicción del tipo IF-THEN, las cuales permiten representar las funciones lógicas de varios valores, sin limitarse a la lógica binaria. Este algoritmo parte de la metodología de Razonamiento Inductivo Difuso (Fuzzy Inductive Reasoning, FIR).

La metodología FIR, actualmente se ejecuta en una plataforma llamada Visual-FIR, la cual se encuentra disponible como un kit de herramientas de Matlab. Un modelo FIR, es un modelo cualitativo basado en la lógica difusa, se compone de una máscara óptima, que es el conjunto de variables relevantes involucradas y una base de reglas patrón, que es el conjunto de relaciones entrada/salida que representan el comportamiento. LR–FIR compacta, mediante un proceso iterativo, la Base de Reglas Patrón que se obtienen con FIR, en un conjunto reducido de reglas lingüísticas que sean interpretables, realistas, eficientes y que describan el comportamiento del sistema analizado.

El algoritmo LR-FIR, se compone de siete etapas, las cuales se describen a continuación:

1. **Eliminar los comportamientos poco representativos**: Este es un paso opcional de pre-procesamiento, el cual pretende filtrar las reglas patrón que tengan muy pocas instancias en el conjunto de entrenamiento.

2. **Compactación básica**: Es un paso iterativo que evalúa por cada variable o premisa, todas las reglas de la base de reglas patrón R. Un sub-

conjunto de reglas puede ser compactado en una regla simple, cuando todas las premisas y el consecuente comparten los mismos valores.

3. **Mejora de la compactación**: En esta etapa se amplía la Base de Conocimiento a casos que no han sido utilizados anteriormente para construir el modelo, ya que en el paso anterior sólo se presenta de manera compacta el conocimiento que estaba disponible. En este paso se tienen dos opciones:

   - APB, all possible beliefs (Todas las posibles creencias): A partir de la base de reglas compactada R', todas las premisas se revisan una vez más, en todas las reglas que tienen valores no negativos, y sus valores son sustituidos por -1. Posteriormente, para cada valor - 1 se hace una expansión a todo el conjunto completo de reglas en comparación con la base original de reglas.
   - RB, ratio of beliefs (Proporción de creencias): Esta opción es similar a la anterior, sólo que en ella la regla candidata es aceptada si y sólo si no se ha encontrado en las reglas en conflicto y en el conjunto extendido de reglas.

4. **Eliminar reglas duplicadas y conflictivas**: En este paso se deben eliminar todas las reglas que estén duplicadas y las reglas conflictivas que se hayan producido en los pasos anteriores; también son eliminadas aquellas reglas que tengan una menor calidad.

5. **Unificación de reglas**: Este paso se puede realizar antes o después de la filtración de reglas. Si se realiza antes, las reglas de baja calidad se unifican con las de mejor calidad, a manera que se conserve tanto como sea posible, el comportamiento del sistema. Si se realiza después, las reglas de baja calidad que tengan conflictos con reglas que tengan una mejor calidad, son eliminadas; por lo tanto el conjunto de reglas resultantes queda más sintetizada y clara, sin embargo, se pierde información. La Unificación de reglas es un proceso iterativo que evalúa, cada regla con respecto a las restantes para encontrar reglas candidatas similares a unificarse en una sola, hay cuatro opciones para realizar este paso:

   - Wise: Un subconjunto de reglas es unificado en una única regla, si y sólo si la calidad de la regla unificada es más alta que la mejor calidad de las reglas candidatas.

- Blind: Un subconjunto de reglas candidatas se unifica sin verificar la calidad de la regla unificada.
- Estas dos opciones se pueden combinar con repeticiones (la regla puede ser unificada con varias reglas) y sin repeticiones (una regla puede ser unificada una sola vez). Sin repeticiones son las opciones por defecto.

6. **Filtrado de Reglas**: En este paso se evalúa el conjunto de reglas obtenido, el usuario fija el valor mínimo aceptable de calidad, las reglas que tengan un valor menor en al menos una de las métricas son eliminadas.

7. **Evaluación de Reglas**: Cada regla se evalúa usando las métricas de sensitividad y especificidad, las cuales permiten una evaluación objetiva y realista; las dos métricas están en el intervalo de [0 – 1].

La sensitividad, expresada por la Fórmula 1.14, es la relación entre el número de datos dentro de la clase que la regla identifica, con el número total de datos en la clase; un valor alto de sensitividad implica una regla muy general, de lo contrario, un valor pequeño indica una regla muy específica. La especificidad, dada por la Fórmula 1.15, se define como la relación entre el número de datos fuera de la clase que la regla identifica, con el número total de datos fuera de la clase.

$$sensitividad = \frac{TP}{TP \quad + \quad FN} = \frac{TP}{Tot - in - class} = 1 - \frac{FN}{TP + FN} \quad (1.14)$$

$$especificidad = \frac{TN}{TN \quad + \quad FP} = \frac{TN}{Tot - out - of - class} = 1 - \frac{FP}{TN + FP} \quad (1.15)$$

Dónde:

- *TP ("verdadero positivo")* es el número de datos que la regla predice que está en la clase x, y que realmente pertenecen a la clase x.

- *FN ("falso negativo")* es el número de datos que la regla prevé que no está en la clase x, y que realmente pertenece a la clase x.

- *FP ("falso positivo")* es el número de datos que la regla predice que está en la clase x, y realmente no pertenecen a la clase de x.

- *TN ("verdaderos negativos")* es el número de datos que la regla predice que no está en la clase x, y que realmente no pertenece a la clase de x.

- *Tot –in - class*, es el número total de datos que están en la clase actual.

- *Tot-out- of-class*, es el número total de datos que no están en la clase actual.

1.6. Metodologías para proyectos de Minería de Datos

Las metodologías permiten aplicar el proceso de minería de datos de manera sistemática y no trivial; de forma que se pueda entender el proceso de descubrimiento de conocimiento; además proporcionan una guía para la planificación y ejecución de los proyectos[46].

Los proyectos de Minería de Datos se realizan bajo distintos escenarios [46], dependiendo de la manera en como inicie el proceso estos escenarios se pueden clasificar en:

- *Escenario donde se aplica la Minería de Datos en una situación organizacional.* Este escenario es comúnmente encontrado en el ámbito de las empresas y organizaciones, donde se busca encontrar patrones y relaciones que ayuden a la solución de un problema.

- *Escenario donde el proyecto inicia con un conjunto de datos.* En este escenario se busca encontrar relaciones interesantes, a través de la exploración de los datos, con la finalidad de que sean útiles en un dominio de aplicación.

En el primer escenario algunas empresas utilizan la metodología CRISP-DM, en el caso de que la empresa compre productos de la empresa SAS

puede recurrir a la metodología SEMMA y la metodología Catalyst ha estado adquiriendo popularidad [46]; sin embargo, en el segundo escenario, es comúnmente implementado el proceso KDD.

Un modelo de proceso es un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo, es común que a los modelos de proceso se les conozca como metodologías,además, en la literatura es frecuente que se manejen estos dos términos de manera indistinta; sin embargo, los modelos de procesos nos indican qué hacer; mientras que las metodologías nos dicen cómo hacerlo, no solo proporcionando las fases del proceso, sino que también definen las tareas que se tienen que realizar.

1.6.1. CRISP-DM

Cross-Industry Standard Process for Data Mining, fue creada en el año 2000 por las empresas SPSS, NCR y Daimer Chrysler [46]; es considerado un modelo de proceso jerárquico que consta de seis fases, las cuales se ilustran en la Figura 1.7, dichas fases se dividen en distintas tareas genéricas y éstas a su vez en tareas especializadas [41].

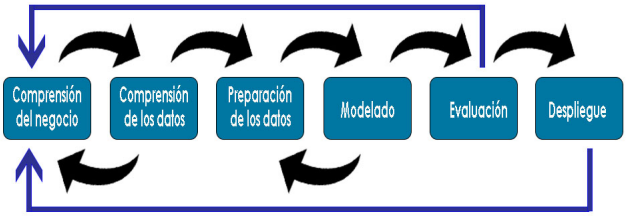


Figura 1.7: Fases de la metodología CRISP-DM

1. Comprensión del negocio: En esta fase se realizan las tareas de comprensión de los objetivos del negocio considerándo el contexto y los requerimientos, la evaluación de la situación, haciendo una indagación detallada, determinación de los objetivos de minería de datos y la generación de un plan de proyecto que incluya las herramientas, equipo y las técnicas.

2. Comprensión de los datos: Con la finalidad de familiarizarse con los datos y los objetivos del negocio, en esta fase se realizan las tareas de recopilación de los datos, descripción de los datos, exploración de los datos y verificación de la calidad de los datos.

3. Preparación de los datos: En esta fase se obtiene la vista minable, que es el conjunto de datos a utilizar en la fase siguiente; estos datos deben ser coherentes, con un formato único y libre de errores, para ello se realizan las siguientes tareas: selección de los datos, limpieza de dato, construcción de datos, integración de datos y formateo de los datos.

4. Modelado: Se aplican las técnicas de minería de datos a la vista minable, esta fase incluye las tareas de selección de la técnica de modelado y herramientas a utilizar, diseño de la evaluación y pruebas, construcción del modelo y evaluación del modelo.

5. Evaluación: Aquí se determina la utilidad de los modelos con base a las necesidades del negocio, por lo cuál esta fase contiene la evaluación de los resultados donde se evalúa en qué medida el modelo es apto o deficiente con respecto a los objetivos del negocio, revisón del proceso y determinación de los pasos próximos.

6. Despliegue: Se realiza la planificación del desarrollo, el plan de supervisión y mantenimiento, la generación de un informe final y la revisión del proyecto.

1.6.2. SEMMA

Esta metodología fue creada por SAS Institute, con la finalidad de que se trabajara con el Software de Minería de Datos de dicha empresa [46]; se define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos [53]. Su nombre proviene de sus cinco fases que la componen: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración)[46]; las cuales se ilustran en la Figura 1.8.

Esta metodología excluye el análisis y la comprensión del problema, y se enfoca en los aspectos técnicos.

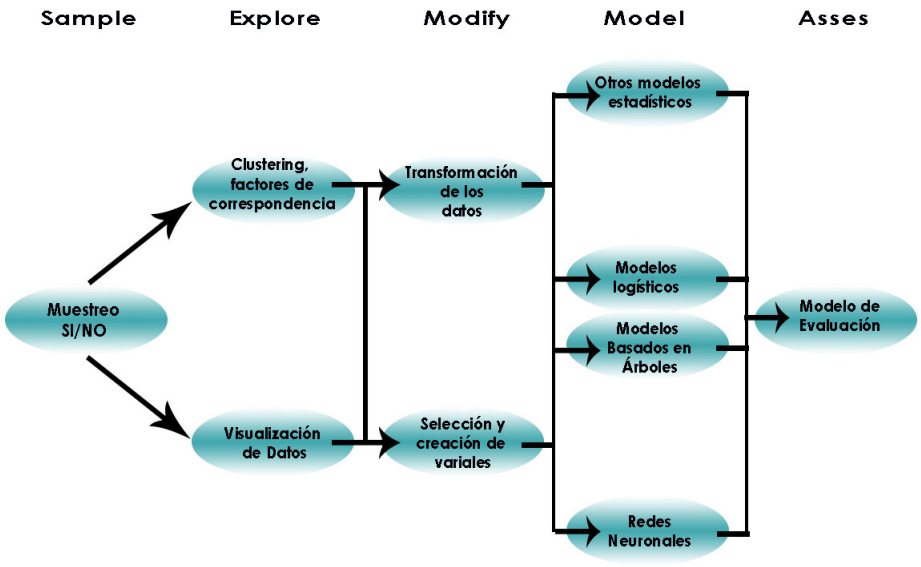


Figura 1.8: Fases de la metodología SEMMA

1.6.3. P^{3}TQ

La metodología Catalyst, es mejor conocida en la comunidad científica como P^{3}TQ por (Product, Place, Price, Time, Quantity) y fue creada en el año 2003 por Dorian Pyle [46]. Esta metodología está conformada por dos modelos: El Modelo de Negocio (MII) y el Modelo de Explotación de la Información (MIII), como se muestra en la Figura 1.9; el primero proporciona una lista de pasos guía para el desarrollo y la construcción de un modelo, en el cual se identifique un problema de negocio [53]; el segundo proporciona una guía de pasos para la construcción y ejecución de modelos de Minería de Datos, partiendo del Modelo de Negocio (MII)[46]. Estos pasos, en ambos modelos, son llamados *boxes*, la idea general es que después de realizar una acción se deben de evaluar los resultados para determinar el próximo paso (box).

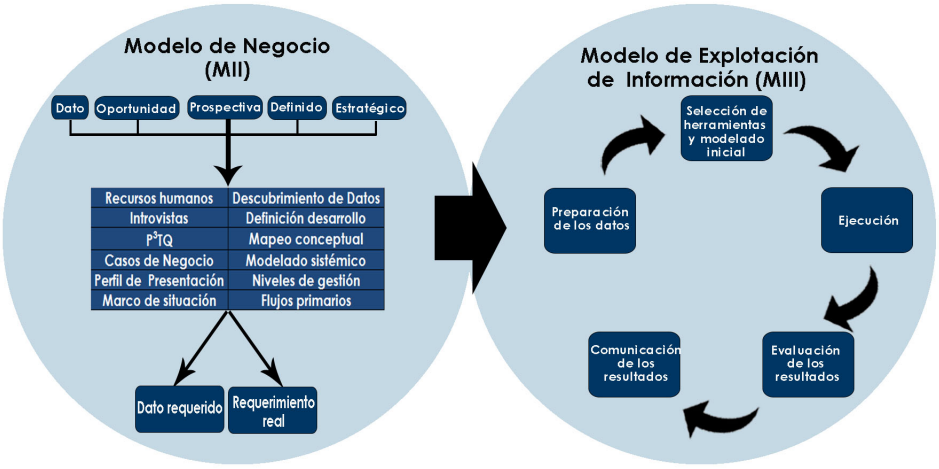


Figura 1.9: Fases de la metodología P^{3}TQ

1.6.4. KDD

KDD es el proceso de extracción de conocimiento en bases de datos, toma su nombre del inglés *Knowledge Discovery in Databases*, este proceso tiene un carácter iterativo ya que incluye ciclos entre sus fases para lograr retroalimentación y el refinamiento del conocimiento [75]; además se considera también interactivo, ya que generalmente un experto en el dominio ayuda en la preparación de los datos y en la validación del conocimiento extraído [35].

Es común que se le llame a la Minería de Datos para referirse al proceso completo del KDD; sin embargo, KDD es el proceso de búsqueda de información y pautas útiles en los datos, mientras que la Minería de Datos es el uso de algoritmos para extraer la información y los patrones obtenidos por el proceso del KDD, [21], en ese contexto la Minería de Datos es una de las fases del proceso KDD, el cual se conforma en cinco fases [35], como se muestra en la Figura 1.10, descritas a continuación:

1. **Integración y recopilación**: En esta fase se determinan las fuentes de información a utilizar y donde conseguirlas, ya que los datos pueden pertenecer a distintas organizaciones o a diferentes departamentos, o que simplemente no se hayan recolectado; en ocasiones, se tendrán que conseguir en bases de datos públicas (por ejemplo, censos, datos demográficos, datos climatlógicos) o bases de datos privadas (datos de

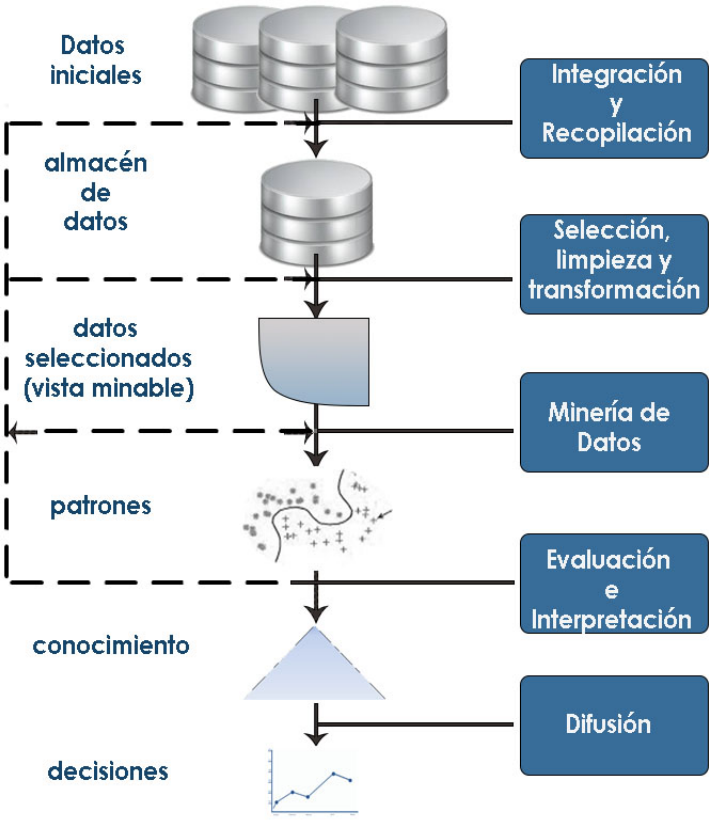


Figura 1.10: Fases del proceso KDD

compañías bancarias, de pagos, eléctricas). Debido a que cada fuente utiliza diferentes formatos de registro, diferentes claves primarias, diferentes tipos de error, etc. es necesario que se integren todos estos datos y posteriormente transformarlos a un formato común, generalmente, cuando se integran múltiples bases de datos, se utilizan los almacenes de datos (data warehousing). Un almacén de datos es un repositorio de información coleccionada desde varias fuentes, almacenada bajo un esquema unificado que reside en un único lugar, los datos se modelan con una estructura multidimensional, donde cada atributo le corresponde una dimensión. Estos almacenes de datos suelen ser muy útiles para las tareas de Mineria de datos, sin embargo, no son indispensables, ya

que en ocasiones se podrá trabajar con los datos originales o en otros formatos, sobre todo cuando el volumen de los datos no es muy grande.

2. **Selección, limpieza y transformación**: Se prepara la vista minable, que es el subconjunto de datos que se van a minar, ya que la calidad del conocimiento que se obtenga dependerá de la calidad de los datos minados; en esta fase se eliminan o corrigen datos que estén incorrectos, se determina que estrategia seguir con los datos incompletos, faltantes o perdidos (*missing values*) y además se consideran las variables a utilizar. Una cuestión a tratar en esta fase, es la presencia de valores que no se ajustan al comportamiento general de los datos (*outliers*), ya que algunos algoritmos los ignoran, otros los consideran como ruido yotros que son muy sensibles a ellos. Otras tareas que también se puede realizar en esta fase son el construir (crear automáticamente nuevos atributos aplicando alguna función u operación a los atributos originales), numerizar (reemplazar valores categóticos en enteros) o discretizar atributos (transformar valores númericos en atributos discretos o nominales), permitiendo así el uso de técnicas que requieran un tipo de dato en específico. Algunos autores fusionan las dos primeras fases en una sola, llamada *Preparación de los datos*.

3. **Minería de Datos**: En esta fase se seleccionan la tarea y los métodos de Minería de Datos que se van a aplicar para obtener el nuevo conocimiento, para ello se construye un modelo, que es la descripción de los patrones y las relaciones entre los datos, estos modelos se pueden utilizar para realizar predicciones, para explicar situaciones pasadas o para entender mejor los datos. Primero, se determina el tipo de tarea de minería apropiada, después se elije el tipo de modelo, posteriormente, se elije el algoritmo que cumpla con la tarea y el modelo seleccionado. Para la construcción del modelo es necesario revisar modelos alternos hasta encontrar el que sea de mayor utilidad, para eso se tendrá que regresar a pasos anteriores para hacer otra iteración probando la misma técnica con diferentes parámetros, otras técnicas, o incluso modificar la definición del problema.

4. **Evaluación e interpretación**: Se evaluán los resultados y se validan por los expertos, estos resultados deben ser precisos, comprensibles e interesantes (útiles y novedosos) y en muchas ocasiones será necesario considerar también el contexto donde el modelo se va a utilizar.

5. **Difusión**: Se hace uso del conocimiento obtenido, dándolo a conocer a todos los posibles usuarios; además es necesario revisarlo constantemente, es decir, deberá ser monitorizado para re-evaluarlo, re-entrenarlo y probablemente reconstruirlo, ya que en ocasiones los patrones pueden cambiar.

Capítulo 2

Estado del Arte

Este capítulo se encuentra dividido en dos secciones, en la primera se abordan los trabajos relacionados a la aplicación de técnicas de Minería de Datos en la solución de problemas sociales como el desempleo y el desempeño escolar, entre otros. La segunda parte comprende investigaciones realizadas sobre la migración utilizando técnicas estadísticas tradicionales, como históricos, tabuladores, porcentajes, promedios, etc. Esta división se hizo con el fin de no sólo revisar los trabajos de Minería de Datos aplicada a diferentes problemas de tipo social, sino también hacer una revisión sobre cómo se realiza, comúnmente, el análisis de la migración.

2.1. Trabajos de Minería de Datos en fenómenos sociales

Diferentes técnicas de Minería de Datos se han aplicado exitosamente a dar solución a diversos problemas sociales, permitiendo así llevar a cabo estudios que proporcionen conocimiento útil y novedoso del problema en cuestión. Algunos de estos trabajos son descritos a continuación.

2.1.1. Identificación de patrones característicos de la población carcelaria mediante Minería de Datos

Debido a un incremento, a partir del año 2001, en el número de casos de violencia, casos de sida y situaciones de promiscuidad en Argentina; este país

tuvo como consecuencia que para el año 2008 se tuviera una sobrepoblación en los centros penitenciarios. Esta situación alarmante llevó a la realización de un estudio por parte de la policía criminal [33], con la finalidad de poder crear una política eficiente que permitiera disminuir, a mediano plazo, el índice de estos delitos.

En este estudio se busca encontrar los patrones de comportamiento y conductas de la población carcelaria, a través de técnicas de Minería de Datos; específicamente el uso de técnicas de clustering, empleadas para formar grupos de esta población. Para este caso, se utilizaron $40,928$ registros con los atributos significativos de los reclusos tales como: edad, estado civil, nivel de instrucción, última situación laboral, lugar de residencia, capacitación laboral, delito cometido y reincidencia. Esta información fue obtenida de la base de datos del Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP), implementado en el año 2002 por la Dirección Nacional de Política Criminal.

Después de aplicar los algoritmos de clustering se detectaron cuatro grupos. El primero de ellos agrupa a internos que cometieron delitos contra la propiedad, que contaban con un trabajo parcial u oficio, pero que su salario escasamente cubría sus gastos básicos y la mayoría son reincidentes. El segundo grupo está conformado por aquellos que cometieron delitos contra personas, generalmente se trata de los reclusos sin oficio y sin profesión. El tercero agrupa a los jóvenes presos que cometieron robo, que no tienen estudios, trabajo ni profesión. El cuarto formado por los internos que en su mayoría cometieron delitos contra la integridad sexual y/o manejo de estupefacientes, por lo general su edad rebasa los 40 años, casados o en concubinato, con trabajo parcial o completo. Estos grupos fueron validados por fiscales penales, jueces, sociólogos y demás especialistas.

De acuerdo a los autores, el aplicar técnicas de clustering a información de tipo socioeconómica, ayuda a obtener esquemas que permitan comprender la conducta humana en problemas de tipo social, en este caso de tipo delictiva, y de esta manera entender qué variables están relacionadas para que una persona cometa un delito y comprender los patrones que lo llevan a ello. Sin embargo, los autores mencionan que usando los resultados obtenidos no les permite conocer a fondo la composición de los grupos de la población carcelaria, por lo cual proponen realizar un trabajo más profundo que per-

mita establecer las reglas de estos grupos o predecir si algún individuo es propenso a quebrantar la ley, para que de esa manera puedan realizar e implementar una política que a mediano plazo reduzca los índices de los delitos.

2.1.2. Una medida de similitud basada en las modas para la caracterización de una población estudiantil en edad extraescolar

Los autores de este trabajo afirman que una de las dificultades del sistema de educación es que existan casos donde el proceso de aprendizaje entre los alumnos sea lento, llevándolos a que tengan una actividad académica con lapsos donde interrumpan sus estudios. Es por eso que el encontrar estrategias y metodologías que apoyen el proceso de aprendizaje de los estudiantes se ha convertido en una de las principales necesidades de las Instituciones de Educación, como es el caso del Instituto de Educación Comfenalco Colombia, donde su población estudiantil está compuesta de individuos en edad extraescolar.

Para poder conocer las características y factores sociales, emotivos, fisiológicos, psicológicos, etc., que afectan el proceso de aprendizaje; el Instituto de Educación Comfenalco y la Universidad de Medellin realizaron un estudio que permitiera generar modelos de los estudiantes [20] a través de técnicas de Minería de Datos llevando a cabo el siguiente proceso:

1. Conformación de la Base de Datos: Diseñaron dos tipos de encuestas y se aplicaron a la población estudiantil del Instituto de Educación Comfenalco, una para los alumnos de educación formal y la otra para educación no formal; las variables utilizadas en dichas encuestas se seleccionaron con la finalidad de que permitieran identificar las necesidades y problemas de los individuos en edad extraescolar.

2. Aplicación de un algoritmo de agrupamiento: Utilizaron una medida de similitud de variables cualitativas para emplear un algoritmo jerárquico aglomerativo, donde cada vez que se fusionan dos grupos se calcula un vector de modas y un vector de frecuencias de las modas para el nuevo grupo; de esta manera las características más relevantes de cada grupo serán aquellas cuyas modas tengan las frecuencias más altas.

3. Selección de grupos: Obtuvieron tres grupos al aplicar el algoritmo jerárquico aglomerativo a la base de datos de educación formal, los cuales quedaron conformados de la siguiente manera: el primer grupo, etiquetado como cluster 0, integrado por 61 registros con 13 frecuencias altas; el segundo grupo, etiquetado como cluster 1, agrupó 98 registros con 12 frecuencias altas; y el último, etiquetado como cluster 2, compuesto por 36 registros con 4 frecuencias altas. También obtuvieron tres grupos, para el caso de la base de datos de educación no formal, los cuales se constituyen de la siguiente manera: el primer grupo, etiquetado como cluster 0, quedó integrado por 105 registros con 17 frecuencias altas; el segundo grupo, etiquetado como cluster 1, agrupó 17 registros con 15 frecuencias altas; y el último, etiquetado como cluster 2, compuesto por 32 registros con 7 frecuencias altas.

4. Identificación de las características relevantes de cada grupo: Para este punto, consideraron como relevantes a aquellas variables que sus modas se presentaran con una alta frecuencia. Para la población de educación formal son: nivel académico, rango de edad, estado civil, número de hijos, situación laboral, consumo de bebidas alcohólicas, etc.; para la población de educación no formal son: nivel académico, habilidad con la computadora, estrato social, comportamiento escolar, consumo de bebidas alcohólicas, etc.

Las características identificadas como las más relevantes en este estudio pertenecen a cada tipo de población (formal y no formal); sin embargo no se hace mención de los patrones representativos de cada uno de los tres clusters (de cada población) obtenidos del estudio realizado.

2.1.3. Diferenciación sociodemográfica del espacio urbano de la Ciudad de México

La segregación residencial (SR) es la aglomeración de los grupos sociales de una misma condición en el espacio y puede ser identificada de acuerdo a una condición socioeconómica, de etnicidad, migración, etc. [2].

El modelo tradicional de SR consistía en una estructura de ciudad basada en círculos concéntricos, donde en el centro se ubicaba la zona comercial e

industrial, rodeada del suburbio antiguo integrado por las mejores residencias; el siguiente nivel conformado por la zona de concentración *in situ*, con viviendas modestas y por último la zona periférica de asentamientos pobres. Este modelo dominó hasta la década de 1980 en las ciudades latinoamericanas, hasta que se comenzó a observar un cambio en el patrón de SR de compacto a disperso, ocasionando una transformación en las ciudades donde las élites se habían trasladado a la periferia, habían surgido nuevos subcentros urbanos y los nuevos grupos más pobres se habían establecido en una periferia lejana dentro del área rural.

Todos esos cambios dieron origen al Modelo de dispersión de la SR, donde el grado de dispersión de los distintos grupos que integran la ciudad, varía de ciudad en ciudad y de un período a otro y en algunos lugares la urbanización difusa es más marcada que en otras. A partir de este modelo surge el interés por encontrar dichos grupos y calcular los índices de SR, utilizando información censal a nivel manzana, para posteriormente realizar una clasificación geodemográfica basada en estilos de vida o niveles socioeconómicos; este análisis se estableció como una metodología llamada geodemografía o geomárketing.

La Metodología de Clasificación Geodemográfica consiste en clasificar zonas residenciales muy pequeñas de la ciudad, generalmente manzanas o segmentos de calles, mediante técnicas de clustering utilizando datos obtenidos de censos de población, encuestas y registros administrativos. Una vez que estos clusters son encontrados, se procede a su descripción en cuanto a los estilos de vida característicos de sus habitantes tales como la composición del hogar, densidad de la población, migración, etc.

Para poder hacer una diferenciación sociodemográfica del espacio urbano de la Ciudad de México, la Universidad Nacional Autónoma de México y la University College London realizaron un estudio [2] basado en la metodología de clasificación geodemográfica utilizando el Censo de Población y Vivienda de 2000 a nivel manzana para el espacio metropolitano. Estandarizaron las variables con el método z-scores, y seleccionaron las más adecuadas elaborando una matriz de correlaciones y un análisis de componentes principales (PCA); el agrupamiento de las manzanas la realizaron con la técnica k-means ponderado, la cual es la más utilizada para elaborar agrupamiento geodemográfico; parametrizaron el algoritmo con 100 iteraciones para llegar

a una solución óptima y el resultado final fueron seis clusters, etiquetados y descritos por los expertos de la siguiente manera:

- Cluster 1, Periferia urbano-rural marginal.- Presenta una estructura de población muy joven, con una alta relación de hijos por mujer, alto peso de migración interna, bajo porcentaje de instrucción media y superior, alto porcentaje de población sin seguro de salud y muy baja pertenencia al IMSS. La población reside principalmente en viviendas sin cocina exclusiva y sin acceso a red de drenaje, así como carentes de los bienes indicativos básicos, son construidas con materiales ligeros y predominantemente con un solo dormitorio, lo que deriva en una altísima tasa de ocupación por cuarto.

- Cluster 2, Empleados de oficina en unidades habitacionales.- Su estructura de población es relativamente joven, aunque con una relación de hijos por mujer cercana a la media. Los habitantes de este cluster viven en manzanas con una densidad de población altísima, lo cual denota un predominio de torres de viviendas o unidades habitacionales, cuenta con un alto componente de migrantes internos, el nivel de educación es alto en relación a la media, la tipología de viviendas que ocupan es en general de calidad, contando con boiler y computadora entre los bienes indicativos, el número de ocupantes por cuarto es bajo, y los hogares están principalmente encabezados por hombres.

- Cluster 3, Proletariado periférico.- Está caracterizado por una estructura de población joven y alta relación de hijos por mujer, con cierta presencia de migrantes internos, el nivel de instrucción predominante es la escuela primaria, y el número de años de escolaridad está por debajo de la media. Ocupa una tipología de viviendas por lo general ya pagadas, aunque de baja calidad constructiva, con drenaje pero no conectado a la red pública, y con cierta variabilidad de tamaños, predominan las viviendas de un solo cuarto, el número de ocupantes por cuarto y por vivienda es alto y la dotación de bienes indicativos está claramente por debajo de la media.

- Cluster 4, Élites urbanas.- Presenta una estructura de población muy envejecida, con baja relación de hijos por mujer, el nivel de instrucción superior es muy alto así como el promedio de años de escolaridad.

Las viviendas son de buena calidad y gran tamaño, generalmente ya pagadas, con presencia de todos los bienes indicativos, el porcentaje de ocupantes por vivienda y por cuarto es muy bajo, y el porcentaje de hogares encabezados por mujeres es alto.

- Cluster 5, Zonas mezcladas.- Es el cluster que mejor representa a la media de la ciudad, abarca zonas muy grandes y socioeconómicamente diversas, incluye a mucha población de todo tipo en su interior.

- Cluster 6, Clase media educada.- Presenta una estructura de edad joven pero, a diferencia de otros clusters, con una baja relación de hijos por mujer, tiene menor presencia de migrantes internos, el nivel educativo está por encima de la media de la ciudad. Predominan las viviendas ya pagadas, cuya calidad es mejor que la media, y de tamaño grande (de dos a cinco cuartos), el porcentaje de ocupantes por cuarto es bajo y las viviendas suelen contar con computadora y boiler. El porcentaje de hogares encabezados por mujeres es superior a la media.

K-means es uno de los algoritmos de clustering de partición más utilizados, sin embargo es necesario conocer previamente el número de grupos a formar para aplicarlo o de lo contrario aplicar algún otro algoritmo de agrupamiento sin restricción que ayude a determinar la cantidad de grupos y posteriormente aplicar el algoritmo de k-means, para finalmente comparar los resultados obtenidos y elegir los mejores. En este caso, podrían aplicar algún otro algoritmo que ratifique el número de grupos determinado por los expertos, para darle sustento científico.

2.1.4. Minería de Datos aplicada a los cambios en la estructura de la variable de desempleo. Caso de estudio: el estado Mérida.

Una de las formas de estudiar los cambios de la estructura en la fuerza laboral, es conociendo los cambios en la estructura del desempleo, la cual se mide a través de la tasa de desocupación. Esto llevó a realizar un proyecto que les permitiera conocer los cambios en la estructura de la variable del desempleo y en consecuencia cambios en la estructura de la fuerza laboral, usando métodos y técnicas de Minería de Datos [16].

Para llevar a cabo dicho estudio utilizaron las Encuestas de Hogares por Muestreo (EHM), realizada por el Instituto Nacional de Estadísticas (INE) del período comprendido entre el primer semestre del año 1995 y el segundo semestre del año 2005, específicamente los registros del estado de Mérida, Venezuela. Dicha encuesta se encuentra estructurada en tres tipos de registros: vivienda, hogar y personas, y es recolectada semestralmente. En este caso sólo utilizaron los registros de personas; y se dividieron en dos bloques de registros, uno desde 1995 hasta 1998 y el otro desde 1999 hasta 2005; esto debido a que algunas variables que aparecían en el primer bloque fueron eliminadas del segundo, además de que existen algunas variables que fueron agregadas en el segundo y que no aparecen en el primero.

El estudio comprendió dos fases: en la primera aplicaron la técnica de Análisis de Correspondencia Múltiple (ACM) para el agrupamiento y en la segunda aplicaron la técnica de Árboles de Decisión (AD) para la clasificación; con la finalidad de encontrar características o patrones que les indicaran cuáles son las variables que identifican la fuerza de trabajo y cuál es el comportamiento que siguen.

Para aplicar la técnica de ACM, utilizaron la herramienta computacional MINITAB versión 14, mediante la opción: Stat- multivariate - Multiple correspondence analysis; de esta manera se definieron tres clases que se utilizaron como variable de salida para los dos bloques de registros y con ellas realizaron la clasificación con los árboles de decisión.

La variable de salida estaba integrada de la siguiente manera: Ocupados, son todas aquellas personas que pertenecen a la población económicamente activa, que trabajan de manera remunerada o que no han trabajado pero tienen trabajo; Desocupados, son todas aquellas personas que pertenecen a la población económicamente activa que no tienen un trabajo pero que desean tenerlo y recientemente han hecho un tipo de esfuerzo por encontrarlo; e Inactivos, son todas aquellas personas que pertenecen a la población económicamente inactiva, que no tienen un trabajo y recientemente no han hecho algo para encontrar uno.

Para la construcción del modelo de clasificación aplicaron el algoritmo C4.5 en la herramienta WEKA 3.5.7.; para evaluar el árbol clasificador utilizaron la técnica de validación cruzada. Generaron un primer árbol, para el

primer bloque, donde se detectaron algunas inconsistencias en las respuestas de algunas variables, ya que se tenía no sabe/no declara; se consideraron estas respuestas como valores perdidos y se les asignó un valor usando la técnica de la mediana más cercana al punto usando el programa SPSS.

Este nuevo árbol mostró una estructura más coherente y todas las reglas que usaba para realizar la clasificación eran afines a lo que realmente se debería obtener. De acuerdo con la técnica de *gain ratio*, colocaron como raíz a una de las variables que consideraban más importantes dentro de la encuesta: *¿Qué hizo la semana pasada? (pp22).*

Para el bloque de registros 99_05, usaron el mismo modelo de evaluación y los mismos parámetros; pero se presentó el mismo problema con los valores no sabe / no declara; a pesar de que era buen clasificador, las reglas de decisión eran inconsistentes. Asignaron a las respuestas de no sabe/no declara, el valor de la mediana más cercana al punto. Obtuvieron un árbol de tamaño veintinueve con quince hojas, siendo el más pequeño al que llega el algoritmo; al estudiar las reglas de clasificación observaron que existía una consistencia en la clasificación de este árbol y además eran afines a lo que realmente debería de ser.

Realizaron tablas de contingencia, con el programa estadístico SPSS mediante el comando Analizar, estadística descriptiva, tablas cruzadas; en ellas se pudo observar que la tasa de desocupación bajó y que para el periodo comprendido en los años 1999_2005, las personas que deberían pertenecer por completo a la población ocupada se encontraban distribuidas en la población ocupada y unos pocos en la población inactiva; en el caso de la población desocupada, la mayoría pertenece a esta y unos pocos a los inactivos, sin embargo la población inactiva se encuentra ubicada totalmente dentro de esta población. Lo anterior indicó un cambio de la población activa a la población inactiva.

Este artículo muestra de manera detallada, la aplicación de técnicas de minería de datos a información de tipo socioeconómica, con lo que se permite encontrar las características que explican, en este caso, un comportamiento de tipo social como el desempleo.

2.1.5. Modelo clasificador para predecir el desempeño escolar terminal de un estudiante

Encontrar la manera de motivar a los estudiantes y darles las herramientas necesarias para que puedan tener una instrucción exitosa, ha sido uno de los principales problemas de la investigación en el área del aprendizaje. Existen diversos factores (socioeconómicos, psicológicos, personales, académicos, etc.) que determinan el desempeño escolar y que al detectar patrones en los mismos se podrían crear programas pedagógicos que apoyen a los estudiantes. Debido a los motivos anteriores, se llevó a cabo una investigación para obtener un modelo clasificador, con ayuda de la minería de datos, que permitiera predecir el desempeño escolar terminal de un estudiante; basándose en la información sobre sus características socioeconómicas y académicas en los primeros tres ciclos escolares de su carrera universitaria [61].

Para dicho estudio manipularon una muestra de 10520 registros, que tomaron de la información correspondiente a catorce generaciones del Instituto Tecnológico de Mérida, México. Para la fase del filtrado de datos, utilizaron los atributos de la base de datos referentes a: información personal, socioeconómica y académica de los primeros ciclos escolares.

Posteriormente para la fase de selección de datos, realizaron una selección de atributos considerando la disponibilidad y limpieza con que se encontraba la base de datos institucional, siendo 21 atributos los que eligieron, definieron el dominio de cada variable y determinaron la variable a observar (variable dependiente), que en este caso se trataba de si el estudiante en estudio va a egresar o no de la institución.

En la etapa de preprocesamiento extrajeron los datos mediante consultas, con el propósito de obtenerlos en un formato adecuado; utilizaron el programa de Sofware Open Source Weka para poder aplicar los algoritmos de minería de datos siguientes: Naïve Bayes, Naïve Bayes Tree y el C4.5 (o J48 en Weka), generándo un modelo matemático clasificador para el algoritmo Naïve Bayes y modelos de árbol para Naïve Bayes Tree y el C4.5. En el caso del algoritmo C4.5, utilizaron en 2 variantes: la primera prueba fue hecha podando el algoritmo (lo optimizaron para que generara pocas ramas), y la segunda fue dejándolo tal cual es (sin podar). Para evaluar el desempeño de los algoritmos clasificadores utilizaron el método de validación cruzada

con 10 particiones, tomaron el 90 % de los datos para entrenar el modelo y el restante 10 % para validar el mismo, repitiendo este proceso diez veces.

Al comparar los resultados de los distintos algoritmos, determinaron que los mejores algoritmos para clasificar, en el caso de estudio fueron: el algoritmo C4.5 y Naïve Bayes Tree, debido a que presentaron mayores porcentajes, principalmente en efectividad y precisión.

En este artículo, una restricción es que no se aplican técnicas de selección de variables, sino que únicamente consideran la disponibilidad y la limpieza en la base de datos para poder hacer dicha selección, lo que se puede prestar para la eliminación de variables importantes o la elección de atributos que no aporten mucha información para el caso de estudio. Otra limitante, es el hecho de que un estudiante necesariamente tiene que haber cursado el tercer semestre para poder predecir si puede egresar o no de la institución, esto ocasiona que a los alumnos que estén cursando en los primeros semestres no se les pueda predecir su desempeño escolar terminal y por ende no se les pueda ayudar con los factores de alarma, de manera oportuna.

2.2. Trabajos de migración con técnicas tradicionales

Esta sección presenta algunos trabajos donde se analiza el fenómeno de la migración haciendo uso de técnicas estadísticas tradicionales. Aunque en ellos se encontraron algunas características que describen este comportamiento, los resultados son muy pobres ya que están basados sólo en el análisis de variables independientes o de relaciones entre pares de variables.

2.2.1. Entre la convergencia y la exclusión. La deportación de mexicanos desde Estados Unidos de América

La deportación de mexicanos por el gobierno de Estados Unidos de América, es un problema que desde hace varios años afecta a ambos países, debido a sus interdependencias económicas y sociales. Para poder analizar las trayectorias históricas de dichas deportaciones, realizaron un estudio que analizara

el periodo comprendido entre 2007 y 2011 [32]. En ese estudio destacan la importancia de la migración de mexicanos hacia Estados Unidos de América y la denominan como un catalizador del cambio social, ya que está modificando el perfil demográfico y étnico de dicho país; y en México, la redefinición de temas como el concepto de nación, parámetros para medir la población mexicana, adecuación de la política pública, servicios sociales, derecho al voto, acceso a la información pública, entre otros.

Utilizan los datos del U.S. Census Bureau. The Hispanic Population: 2010 y el U.S. Census Bureau. The Foreign-Born Population in the United States: 2010, y obtienen unos porcentajes importantes como: el 10 % de los mexicanos viven en ese país, 3.8 % de la población de E.E.U.U. nacieron en México, del flujo de mexicanos que ingresó en los años 90 era del 30.7 % y aumentó al 34.5 % para el año 2000, además estiman que para el año 2050, cerca del 30 % de la población del país vecino, tendrán un perfil étnico hispano, en su mayoría mexicano.

Sobre el flujo migratorio, mencionan que ha coincidido con tres eventos de naturaleza distinta: la maduración institucional en materia de seguridad y asuntos migratorios, la recesión económica en Estados Unidos de América (2008-2010), y en consecuencia la reducción de los flujos laborales de mexicanos hacia el norte; con esto último, provocando que las deportaciones también sean menores, para poder mostrar esto, hacen una comparación de las estadísticas de detención y deportación de la Patrulla Fronteriza entre los años 2007 y 2011, y observan que los eventos disminuyeron de 876 mil a 340 mil. Observan también, la disminución de casos de repatriación, con los datos de la Encuesta sobre Migración de la Frontera Norte (EMIF), que van de 807 mil a 357 mil eventos.

Mencionan un cambio en el perfil de los mexicanos que intentan el cruce fronterizo, caracterizado ahora por la reducción del flujo que carece de algún documento para pasar, ya que en el año 2007 las personas que tenían algún documento (no de trabajo) para entrar a Estados Unidos de América, equivalía al 20.2 % del total del flujo, y en 2011, ascendió a un 47.7 %. Los tiempos de permanencia en E.E.U.U. son mucho más largos, esto lo observaron también con base en los datos de la EMIF, debido a que en el año 2007, el 6 % declaró residir en E.E.U.U. y entre los años 2008 a 2011, aumentó al 33 %. También mencionan que se modificó el perfil de las edades, ya que aho-

ra de las personas deportadas hay menos jóvenes y aumentaron las personas de edades mayores.

En este trabajo describen el perfil de los migrantes mexicanos, sin considerar el conjunto de rasgos que los caracterizan; ya que consideran una variable a la vez, como el documento para cruzar, la residencia y la edad, sin embargo no hacen alusión a la combinación de estas variables y omiten otras características que sean relevantes para la descripción completa de estos perfiles.

2.2.2. Rumbo al norte: Nuevos destinos de la emigración veracruzana

El estado de Veracruz había sido considerado hasta 1980 como una entidad en equilibrio migratorio, en los años 90 se empezó a notar un incremento de emigrantes y en el año 2000 adquirió la categoría de entidad de expulsión migratoria. Esto motivó a la realización de un estudio por parte del Colegio de la Frontera Norte que examinara la participación del estado de Veracruz en los flujos migratorios, principalmente aquellos que transitan en la frontera norte mexicana [6].

Para poder examinar dicho fenómeno, primero analizaron la situación migratoria de la entidad en la segunda mitad del siglo XX, posteriormente estudiaron la condición migratoria en el contexto nacional e identificaron los municipios con el mayor grado de intensidad migratoria, Finalmente, revisaron la participación de migrantes veracruzanos que transitan en la frontera norte del país. Para el primer punto, realizan cuadros con los datos de CONAPO con el número de inmigrantes, emigrantes y saldo migratorio en el estado de Veracruz, y con el volumen y las principales entidades de destino de los emigrantes veracruzanos, entre los años de 1955 y 1995, con ello observan que hasta el año de 1995 los destinos de los emigrantes veracruzanos eran los estados circunvecinos, el Distrito Federal y el estado de México. Encuentran que en el año 1992, en la zona sur del estado hubo 23,130 desempleados y lo atribuyen a la privatización o desaparición de empresas paraestatales como el Instituto Mexicano del Café (Inmecafe), Tabacalera Mexicana (Tabamex) y Azucarera, S.A, por mencionar algunas, ocasionando que los veracruzanos buscaran alternativas laborales en otros lugares, tanto en otras entidades del

país como en Estados Unidos.

Para estudiar la migración veracruzana en el contexto nacional, observan que en el año 2000 la entidad se coloca en el noveno lugar como entidad expulsora de población, además comparan el porcentaje de migrantes veracruzanos que captaron ciudades fronterizas, que era del 20.6 % entre los años de 1985 y 1990, con el porcentaje del periodo de 1995 al año 2000, el cuál aumentó al 41 %.

Los autores realizan una clasificación de las entidades federativas mexicanas según el grado de intensidad migratoria a Estados Unidos, el cual fue construido por CONAPO en el año 2002, utilizando la muestra del 10 por ciento del Censo General de Población y Vivienda de INEGI, del año 2000. De esta manera clasifican a las entidades como entidades de bajo o muy bajo índice de intensidad migratoria, medio, alto o muy alto; clasificando al estado de Veracruz en el grupo de bajo y muy bajo. Los estados con gran tradición migratoria, están colocados en el grupo de muy alto, mientas que los estados de Morelos e Hidalgo aparecen en el grupo de alto. A pesar de que Veracruz, presenta un índice bajo, a nivel regional y municipal observan que el 10 % de los municipios veracruzanos estaban participando significativamente en la emigración a Estados Unidos, para poder explorar las características de éstos municipios los agrupan de acuerdo a la región geográfica a la que pertenecen considerando los municipios con grado muy alto, alto y medio.

Para revisar la participación de migrantes veracruzanos que transitan en la frontera norte del país, utilizaron los datos de la Encuesta sobre Migración de la Frontera Norte (EMIF), identificaron las localidades de la frontera norte mexicana donde los veracruzanos permanecieron la mayor parte del tiempo durante su trayectoria y estancia migratoria, siendo la Ciudad de Matamoros una ciudad de arribo y Reynosa un destino laboral para los veracruzanos. Además, identificaron los municipios de residencia de los emigrantes veracruzanos que llegaron a las ciudades fronterizas y los que regresaban a Veracruz, creando un cuadro con diez municipios de residencia y ordenándolos de manera descendente según la participación porcentual durante la década de 1993 – 2003.

Con respecto a los migrantes que retornaban a Veracruz, que fueron captados por la encuesta EMIF, en la década antes mencionada, los estados de la

Unión Americana que aparecen como destinos de la emigración internacional veracruzana, destacan en los primeros lugares Texas, California y Carolina del Norte, seguidos por Arizona y Lousiana. Con participación intermitente Virginia y Georgia, y de manera menor Michigan, Oregon, Kansas y Minnesota.

En este artículo determinan cuatro clases según el índice de intensidad migratoria a Estados Unidos: a) bajo o muy bajo, b) medio, c) alto y d) muy alto; y posteriormente clasifican a las entidades en estos grupos; sin embargo no aplican alguna técnica que verifique que los grupos creados sean los más adecuados para clasificar a las entidades, ya que las entidades con índice bajo y con índice muy bajo podrían contener características que las discriminen entre sí, causando que formaran dos grupos en vez de uno.

2.2.3. Estados Unidos, lugar de destino para los migrantes chiapanecos

En años recientes, debido al incremento en la emigración hacia Estados Unidos, se ha visto la necesidad de encontrar información con sustento estadístico que describa la magnitud y las características de dicho fenómeno; esto llevó a que el Consejo Estatal de Población en el estado de Chiapas, realizara una investigación que examinara la evolución histórica de los patrones migratorios en Chiapas [38], considerando los antecedentes, los movimientos internos y la concentración espacial de la población en el periodo 1970 – 2005, y las migraciones interestatales entre los años de 1950 y 2005; además de dimensionar la magnitud de los chiapanecos involucrados en el proceso migratorio hacia Estados Unidos en los años de 1925 al 2003, y explorar sus características más importantes.

Para ello utilizaron el Censo General de Población y Vivienda de1970, 1990 y 2000, el Conteo de Población y Vivienda 1995 y 2005, la Encuesta Nacional de la Dinámica Demográfica 1997, el Módulo sobre Migración de la Encuesta Nacional de Empleo 2002, y la Encuesta sobre Migración de la Frontera Norte (EMIF) en el periodo de 1993 y 2003.

Crearon un cuadro comparativo con los datos del Censo General de Población y Vivienda de1970, 1990 y 2000 y el Conteo de Población y Vivienda

2005, en dicho cuadro registraban la población total y la distribución porcentual respecto a la población total, de las principales ciudades de la entidad; con ello determinan un patrón de concentración espacial en las ciudades de Tuxtla Gutiérrez, San Cristóbal de las Casas, Tapachula y Comitán de Juárez, denominando a estas ciudades como polos de atracción para los flujos migratorios intermunicipales. Con base en la Encuesta Nacional de la Dinámica Demográfica 1997, calculan que 97,426 chiapanecos cambiaron de municipio de residencia, entre los años de 1992 y 1997. Observaron que a partir de los años sesenta, el número de emigrantes era mayor que el número de inmigrantes, provocando esto un saldo migratorio negativo. De los migrantes que abandonaban la entidad, para el año de 1970, el 41 % vivían en el Distrito Federal y estado de México, en Tabasco el 20 %, en Veracruz el 16 % y Oaxaca el 7 %.

En la década de los 90's notaron un incremento en el número de emigrantes interestatales, relacionando este comportamiento con elementos contextuales de la historia política, económica, social y demográfica de la entidad, destacando lo siguiente: una fuerte presión en la tierra, ya que el 41.3 % dela población económicamente activa estaba dedicada a la agricultura, ocasionando que la tierra para uso agrícola no fuera suficiente; la caída del precio internacional del café, ya que en el año de 1989 se rompió la cláusula económica dela Organización Internacional del Café (OIC); los efectos del Tratado de Libre Comercio (TLCAN) sobre la producción del maíz, debido a que en 1992 se inició la entrada al país de este grano proveniente de Estados Unidos provocando el abandono de esta actividad por parte de los agricultores chiapanecos, quienes no pudieron competir con la industria estadounidense, llevando a la ocupación en el sector agrícola del 58.3 % en el año de 1990 al 41.3 % en el año de 2005. Existen otros elementos que también influyeron en la emigración de chiapanecos, como el levantamiento armado del Ejército Zapatista de Liberación Nacional (EZLN) en 1994; los desastres naturales, como el huracán Mitch en 1998; el incremento de la población en edad laboral, que pasó de 750 mil en 1970, a 2.3 millones en el 2005.

Con base en los datos de la Encuesta sobre Migración en la Frontera Norte de los años, 1993-1994 y 2002-2003, observaron un incremento en el número de chiapanecos que llegaron a las localidades fronterizas, el cuál aumentó de 6434 a 84,693, los cuales declaró el 44.14 % que sus razones que lo motivaron a realizar el desplazamiento era dirigirse a Estados Unidos, y el 39.94 % que

era para trabajar o buscar trabajo. Otros datos que encontraron fueron los siguientes: entre los años de 2000 al 2003, el 70 % señalaron a la localidad de Sásabe (ubicada en la frontera entre Arizona y Sonora) como punto de cruce; de los chiapanecos que ya habían elegido una destino final, se destacan el estado de California con el 39.1 %, seguido en orden de importancia Florida, Arizona, Oklahoma, Texas, Illinois, Colorado, Washington, Nuevo México y Oregon.

En este artículo nuevamente se abordan los patrones migratorios, describiéndolos a través de diferentes características en específico, las cuales las analizan de manera aislada sin considerar si existen relaciones entre ellas, esto es lo que lleva a que exista conocimiento útil oculto.

2.2.4. Migración y remesas en el sur del estado de México

Una de las principales fuentes de divisas para la economía mexicana, son el flujo de fondos provenientes de trabajadores que residen en el extranjero, los cuales han aumentado considerablemente en años recientes; esto llevó a que la Universidad Autónoma del Estado de México realizara una investigación [31] que les permitiera estimar la migración internacional mexiquense y las remesas; así como observar el destino final de las mismas, en los municipios de Tejupilco y Almoloya de Alquisiras en el estado de México.

En ese trabajo mencionan como patrones de emigración a Estados Unidos, de la entidad mexiquense, los siguientes:

- Los individuos que proceden de zonas de alta tradición migratoria, por ejemplo el sur del estado de México.
- Los individuos que proceden de zonas emergentes (zonas urbanas), por ejemplo los municipios de Nezahualcóyotl, Chimalhuacán y Ecatepec, entre otros.
- Los individuos de otras zonas, consideradas también emergentes, como las comunidades indígenas mazahuas y otomíes.
- La migración hacia Canadá y el resto del mundo, principalmente a Europa.

Con base en el Censo General de Población y Vivienda del año 2000, determinaron las características sociodemográficas de las personas que recibieron remesas, en los cinco años anteriores al censo, las cuales son las siguientes:

- En cuanto al género, 68.5 % eran mujeres y 31.1 % eran hombres.
- En cuanto a la edad, 55 % tenían entre 0 y 44 años de edad, y 44.1 % tenían 45 y más.
- En cuanto al parentesco, 55.7 % eran jefes (as) de hogar, 17.2 % eran esposos (as) y 14.5 % eran hijos (as).
- En cuanto al estado civil, 60.1 % eran casados (as) o unidos (as), 18 % solteros (as) y 21.9 % declaró otra relación de parentesco.
- En cuanto a la escolaridad, 22.6 % no había concluido la primaria o no tenía ninguna instrucción escolar, 41 % había concluido la primaria y el 36.4 % la secundaria.

Para analizar la migración y las remesas en los municipios de Tejupilco y Almoloya de Alquisiras, seleccionaron una muestra aleatoria de los datos del Conteo de Población y Vivienda de 1995. El tamaño de la muestra lo estimaron a partir del muestreo por proporciones, quedando para el caso de Tejupilco de 400 viviendas (200 rurales y 200 urbanas) y para Almoloya de Alquisiras, encuestaron a habitantes de 350 viviendas. Los instrumentos de medición que utilizaron fueron cuestionarios aplicados a personas que habían regresado al municipio o localidad a finales del año 2001 y principios del 2002, o personas mayores que tuvieran suficiente conocimiento sobre la migración de su familiar.

Para observar la situación sociodemográfica de estos municipios, consideraron el tamaño absoluto de la población, que fue de 84,897 personas, en el año 1995, a 95,032 en el año 2000, para el municipio de Tejupilco; en el caso del municipio de Almoloya de Alquisiras, aumentó de 13,667 habitantes en el año 1995, a 15,584 en el año 2000. Con base en la encuesta que aplicaron, pudieron percibir las características de la migración en estos municipios. En Tejupilco, 279 personas viajaron a Estados Unidos por motivos de trabajo, de los cuales 171 era su primera vez y 100, ya habían ido más de dos veces; para Almoloya de Alquisiras, fueron 121, de las cuales más del 90 % había emigrado más de dos veces. Respecto a los puertos fronterizos de cruce fueron:

Piedras Negras 42.04 %, Nuevo Laredo 18.18 %, Matamoros 5.68 %, Reynosa 3.79 % y Tijuana 9.09 %, para Tejupilco; Agua Prieta 43 %, Tijuana 22.53 %, Piedras Negras 12.68 %, para Almoloya de Alquisiras. En cuanto a la forma de cruce, el 79.31 de migrantes cruzaron sin papeles y con ayuda de un pollero, para Tejupilco y 79.05 % para Almoloya de Alquisiras. De los migrantes tejulpiquenses, el 88.93 % llegó con un amigo, y de los almoloyenses, el 86.9 %.

En este artículo describen los perfiles de migración, considerando únicamente dos características: la región de origen de los migrantes y su lugar destino; mencionan también las características de los migrantes resaltando datos específicos, como la edad, género, estado civil y escolaridad, por lo que la obtención de porcentajes es la más adecuada ya que sólo requieren enlistar dichas características; lo mismo sucede con el destino de las remesas.

2.2.5. La migración como respuesta de los campesinos ante la crisis del café: Estudio en tres municipios del estado de Puebla

Debido a la alarmante disminución del precio del café, los productores de este grano han sufrido una crisis en la que la cosecha sea incosteable para ellos y que a pesar de los apoyos recibidos, esto sea insuficiente; ocasionando así el abandono total y parcial de gran parte de los cafetales, llevando a que los productores busquen alternativas, como el caso de la migración, que les permita mejorar la calidad de vida de sus familias. Esto motivó a que la Universidad Autónoma Indígena de México realizara un estudio que analizara la relación entre la pobreza, la producción del café y la migración, así como el impacto en las familias [58].

En dicho estudio realizaron una encuesta a 49 campesinos de la Sierra Norte de Puebla. Distribuyeron la muestra de la siguiente manera: 25 productores del municipio de San Felipe Tepatlán, 18 de Amixtlán y 6 de Hueytamalco. Las técnicas que utilizaron para las variables de tipo cuantitativo fueron: análisis de varianza, prueba de t apareada y correlación de Pearson; y para las variables cualitativas fueron: prueba de Kruskal – Wallis y correlación de Spearman.

Obtuvieron algunas características de los productores de café entrevista-

dos, como la escolaridad y encontraron, mediante una correlación de Spearman, que existe una relación negativa entre el nivel de escolaridad y la edad del productor, lo que indica que los productores con mayor edad tienen menor nivel educativo. Además, mediante un análisis de varianza, encontraron que no existe diferencia significativa entre los municipios en relación con la edad de los productores.

En este artículo aplican diversas técnicas que permiten no sólo analizar de manera aislada las características de los productores sino que también les permite encontrar relaciones entre estas variables. Sin embargo, los autores sólo analizan un par de variables a la vez, por lo que al ser un número basto de variables, el procedimiento se volvería muy costoso dado que tendrían que aplicar varias veces estas técnicas.

Existen otros trabajos, que no mencionan las técnicas utilizadas, donde abordan el análisis de la migración basándose en encuestas propias, como en el trabajo ***De asalariado a empresario: la reinserción laboral de los migrantes internacionales en la región centro-occidente de México*** [49], donde aplican la encuesta a una muestra de 5532 ex migrantes internacionales y analizan las transferencias monetarias realizadas por los migrantes durante su estancia en Estados Unidos, sus inversiones, los empleos que crearon en su lugar de origen y su reinserción profesional al finalizar su ciclo migratorio. Otro es el trabajo ***Remesas e Inversión productiva en comunidades de alta migración a Estados Unidos. El caso de Teocaltiche, Jalisco*** [12], donde analizan el impacto económico de las remesas en la economía de Teocaltiche, Jalisco. Una forma comúnmente utilizada para analizar la migración, es haciendo uso de entrevistas como es el caso del trabajo ***Inversión social y productividad de los Migrantes mexicanos en los Estados Unidos*** [45], donde identifican a distintos tipos de migrantes con posibilidades de inversión para diseñar e implementar políticas públicas específicas y elaborar propuestas sobre la implementación de programas de estímulo a la inversión de los migrantes que activen la economía de los lugares de origen de los migrantes.

Capítulo 3

Análisis del factor demográfico de los migrantes

Este trabajo presenta el estudio del comportamiento de Migrantes en el estado de Hidalgo, a través del análisis de dos componentes principales. El primero, es el factor demográfico que muestra las codiciones de las viviendas de origen de aquellos que han sido reconocidos como migrantes en el Censo de Población y Vivienda 2010. Por otro lado, el factor social es analizado haciendo uso de la información obtenida de la Encuesta sobre Migración de la Frontera Norte de México (EMIF) levantada en el año 2012. Este capítulo muestra en particular el análisis del primer factor considerado en nuestro estudio. Las siguientes secciones describen los datos y el proceso KDD, seguido para llevar a cabo tal análisis.

3.1. Fuentes de información para el factor demográfico

Distintos organismos gubernamentales han desarrollado diversas fuentes de información que cuantifican y caracterizan el fenómeno complejo de la migración entre Estados Unidos de América y México [63]. Las primeras fuentes disponibles fueron:

- Encuesta Nacional de Migración a la Frontera Norte del País y a los Estados Unidos (ENEFNEU), efectuada por el Centro Nacional de Información y Estadísticas del Trabajo (CENIET), perteneciente a la

Secretaría del Trabajo y Previsión Social (STPS), en el periodo .de 1977 y 1979.

- Encuesta en la Frontera Norte a Trabajadores Indocumentados Devueltos por las Autoridades de los Estados Unidos (ETIDEU) y la Encuesta Nacional de Migración en Áreas Urbanas (ENMAU), desarrolladas por el Consejo Nacional de Población (CONAPO) en 1984 y 1986-1987.

Hoy en día, las fuentes que contienen módulos o preguntas referentes a la migración, son las siguientes:

- Encuesta Nacional de la Dinámica Demográfica (ENADID), llevada a cabo por el INEGI en 1992, 1997 y 2009, y por el Instituto Nacional de Salud Pública en 2006, como parte de un proyecto de cooperación interinstitucional entre la Secretaría de Salud, CONAPO y el INEGI.

- Conteos de Población y Vivienda de 1995 y 2005 y los Censos Generales de Población y Vivienda de 1990, 2000 y 2010, llevados a cabo por el INEGI.

- Encuesta Nacional de Empleo (ENE), a cargo del INEGI y de la Secretaría del Trabajo y Previsión Social (STPS). La ENE fue sustituida en 2005 por la Encuesta Nacional de Ocupación y Empleo (ENOE).

- Encuesta sobre Migración en la Frontera Norte de México (EMIF NORTE), llevada a cabo por el Consejo Nacional de Población (CONAPO), la Secretaría del Trabajo y Previsión Social (STPS) y el Colegio de la Frontera Norte (COLEF), con la participación del Instituto Nacional de Migración (INM), de la Secretaría de Relaciones Exteriores (SRE) y la Secretaría de Salud (SS).

Para el análisis de este factor se utilizaron los datos del Censo de Población y Vivienda 2010 de INEGI. Se eligió esta fuente debido a que las variables que contiene proporcionan información sobre las características de las viviendas que dejaron los migrantes en el estado. Además, describe las condiciones y calidad de vida de sus familias, con lo cual se pueden obtener los perfiles demográficos de la migración en el estado de Hidalgo.

3.1.1. Censo de Población y Vivienda 2010

El Instituto Nacional de Estadística y Geografía (INEGI) es el encargado de desarrollar y realizar los Censos y Conteos de Población y Vivienda en México, éstos son considerados como de las más completas fuentes de información estadística, ya que percibe la realidad nacional, permitiendo a diversos sectores sociales identificar el rezago social, los grupos vulnerables, necesidades primordiales de la población como la vivienda, educación, salud, servicios, etc., y con ellos elaborar programas que mejoren la calidad de vida de los habitantes. En complemento, proporciona (i) a las diferentes órdenes de gobierno e instituciones, los datos necesarios para la planeación, programación, toma de decisiones, seguimiento y evaluación de los planes y programas que elaboran; (ii) a los estudiantes e investigadores, les brinda estadísticas que les permitan conocer el perfil demográfico, económico y social de la población, como apoyo a la planeación de proyectos, estudios de la población y diagnósticos, etc.; (iii) a los empresarios, facilita información útil para la toma de decisiones referente a sus negocios; y (iv) suministra a la sociedad, datos básicos sobre el volumen y las características de su localidad, municipio, estado y, de manera general, del país [36].

En 1895, se realizó en nuestro país, el primer censo de carácter nacional, dando inicio a los censos contemporáneos, ya que fue seguido por el censo en 1900 y a partir de ahí, se han llevado a cabo cada 10 años (excepto por el de 1921, que se pospuso debido a la Revolución Mexicana). En 1995, INEGI realizó el primer Conteo de Población y Vivienda, con el fin de actualizar la estadística demográfica y socioeconómica del país, en periodos más cortos; y en el año 2005, realizó el segundo. En total, desde 1895 a 2010, se han realizado 13 censos y 2 conteos poblacionales, siendo el más reciente el Censo de Población y Vivienda 2010, realizado del 31 de mayo al 25 de junio de ese año. Este tuvo como objetivo principal, contar la población residente del país, actualizar la información sobre sus características demográficas y socioeconómicas, ubicar su distribución en el territorio nacional, enumerar a las viviendas y captar datos sobre las características de las mismas. Para llegar a este fin, el censo contiene preguntas sobre los habitantes como la edad, sexo, escolaridad, lugar de nacimiento, etc.; así como preguntas relacionadas con las viviendas, como el material con el que están construidas y los servicios con los que cuentan.

La periodicidad para dicho censo, como ya se mencionó, es decenal y se

realiza en los años terminados en cero, aplicándose en toda la República Mexicana; siendo la población objetivo, los residentes habituales del territorio nacional, los hogares censales, las viviendas particulares y las colectivas; para el caso del cuestionario ampliado, se considera además, al migrante internacional.

El cuestionario básico incluye los siguientes temas:

- Población
  - Sexo, Edad y Relación de parentesco.
  - Número de hijos nacidos vivos e Hijos fallecidos.
  - Lugar de nacimiento y Lugar de residencia en junio de 2005.
  - Condición de habla indígena, Lenguas indígenas y Condición de habla española.
  - Discapacidad, desde el enfoque de limitaciones en la actividad.
  - Condición de alfabetismo, Condición de asistencia escolar y Nivel y Grado de escolaridad.
  - Condición de actividad económica.
  - Derechohabiencia a servicios de salud.
  - Situación conyugal
  - Religión.
- Vivienda
  - Material en pisos.
  - Número de dormitorios y Número de cuartos.
  - Disponibilidad de energía eléctrica, Agua y Drenaje.
  - Disponibilidad de excusado y Admisión de agua en este servicio.
  - Disponibilidad de bienes y Tecnologías de Información y Comunicación (TIC).

El cuestionario ampliado contiene las mismas preguntas que el básico y se complementa con lo siguiente:

- Población

  - Condición de residencia de la madre, Condición de residencia del padre, y Presencia del cónyuge.
  - Hijos sobrevivientes, Fecha de nacimiento del último hijo, Sobrevivencia y Edad al morir.
  - Municipio de residencia en junio de 2005.
  - Comprensión de habla indígena y Autoadscripción étnica.
  - Causa de la limitación en la actividad.
  - Área de estudio.
  - Ocupación u oficio, Posición en el trabajo, Condición de prestaciones laborales y/o sociales, Horas trabajadas, Ingresos por trabajo, Sector de actividad económica, Lugar de trabajo y Condición de percepción de otros ingresos.
  - Uso de servicios de salud.

- Migración internacional en los últimos cinco años
  - Sexo y Edad
  - Fecha de emigración
  - Lugar de origen
  - País de destino
  - País de residencia actual
  - Fecha de retorno
  - Condición de residencia

- Alimentación
  - Condición de acceso a la alimentación en los últimos 3 meses.

- Vivienda
  - Material en paredes y Material en techos
  - Disponibilidad de cocina y Combustible para cocinar
  - Dotación de agua
  - Excusado de uso exclusivo
  - Forma de desechar la basura
  - Tenencia, Forma de adquisición y Equipamiento de la vivienda

3.2. Experimentos y resultados

En esta sección se detallan los experimentos realizados para el análisis del factor demográfico y se muestran los resultados obtenidos de cada uno de ellos.

3.2.1. Escenario de la investigación

Para esta investigación se siguieron las fases del proceso KDD, ya que como se mencionó en la Sección 1.6 es el que mejor se aplica en este tipo de estudios.

Para aplicar los algoritmos de Minería de Datos, se utilizó el Software Open Source Weka, versión 3.7.9, debido a que es un software de distribución libre [16].

Para la parte de clustering se aplicaron los algoritmos EM, SOM, Simple - K means y Make Density Based Clusterer, los resultados obtenidos fueron evaluados por el Índice de Davies Bouldin. Posteriormente, se eligió la mejor agrupación y se le aplicaron los siguientes algoritmos de clasificación: J48, LibSVM, MultilayerPreceptron, Naive Bayes e IBk (con valores de k de 3, 5 y 7); todos estos algoritmos fueron evaluados por validación cruzada con 10 particiones.

3.2.2. Integración y recopilación

En esta fase se recopilaron los datos del Cuestionario Ampliado del Censo de Población y Vivienda 2010, ya que contiene un módulo dedicado a la migración internacional en los últimos cinco años (de la fecha en que se realizó el censo), estos datos se encontraban en un formato compatible con el Software IBM SPSS Statistics, divididos en dos archivos (tablas): la Base Nacional de Migrantes y la Base Nacional de Viviendas; debido a que estas bases contaban con datos de toda la República Mexicana, se realizaron los filtros necesarios para que los datos fueran los específicos del estado de Hidalgo, donde se tuvo o se tiene un migrante en Estados Unidos de América. Para ello se consideraron las variables: MCONMIG de la pregunta: *Durante los últimos* 5 *años, esto es, de junio de* 2005 *a la fecha, ¿alguna persona que*

vive o vivía con ustedes se fue a vivir a otro país?, la variable MPDESOTR_C de la pregunta: *¿A qué país se fue?*, y la variable ENT que identifica a la entidad.

3.2.3. Limpieza y transformación

En esta fase, los datos fueron exportados a un servidor de MySQL, versión 5.0.51*a*, para poder darles un mejor tratamiento. Se relacionaron la Base Nacional de Migrantes con la Base Nacional de Viviendas, mediante la variable ID_VIV (identificador de la vivienda) y con una consulta de SQL se generó una sola tabla. Posteriormente, se excluyeron las variables que no contenían información o que tenían el mismo valor en todos los registros; a las preguntas que contenían respuestas de "No sabe", se les dio un formato que fuera congruente con el resto de las respuestas, a través de la creación de vistas en MySQL, por ejemplo, las variables de *PAREDES, TECHOS, PISOS, etc. tenían un valor de N para la respuesta de "No sabe", cuando generalmente el resto de las respuestas eran valores entre 1 y 10, por lo que se le asignó desde la vista de MySQL el valor de 0 cuando se hacía referencia a esta opción.* Finalmente, se obtuvo un conjunto de datos compuesto por 1567 registros y 45 variables, las cuales se describen en la Tabla 3.1.

Tabla 3.1: Descripción de las variables para el análisis del factor demográfico

Variable	Descripción
MUN	Clave del municipio o delegación
MSEXO	Sexo del Migrante
MEDAD	Edad que tenía el migrante cuando se fue la última vez
TAM_LOC	Tamaño de la localidad
CLAVIVP	Clase de Vivienda Particular, por ejemplo: Casa independiente, departamento en edificio, vivienda en vecindad, etc.
PAREDES	Material de las paredes, por ejemplo: material de desecho, lámina de carton, madera, adobe, ladrillo, etc.
TECHOS	Material del techo, por ejemplo: material de desecho, lámina metálica, palma o paja, teja, losa de concreto, etc

Variable	Descripción
PISOS	Material de los pisos, por ejemplo: tierra, cemento o firme, madera, mosaico, etc.
COCINA	Si la vivienda cuenta con cocina
CUADORM	Número de cuartos dormitorios que tiene la vivienda (sin contar pasillos)
TOTCUART	Total de cuartos de la vivienda (sin contar pasillos ni baños)
ELECTRI	Si la vivienda cuenta con luz eléctrica
DISAGU	Disponibilidad de agua en la vivienda, por ejemplo: agua entubada dentro de la vivienda, agua de pipa, agua de pozo, etc.
DOTAGUAD	Dotación de agua en la vivienda, por ejemplo: diario, cada tercer día, dos veces por semana, etc.
SERSAN	Si la vivienda cuenta con servicio sanitario (excusado, retrete, letrina, etc.)
USOEXC	Si el Servicio Sanitario lo comparten con otra vivienda
CONAGU	Si el servicio sanitario tiene descarga directa de agua, le echan agua con cubeta, etc.
DRENAJE	Si el drenaje se encuentra conectado a la red pública, fosa séptica, etc.
COMBUST	Combustible para cocinar, por ejemplo: gas, leña, carbón, electricidad, etc.
ELIBAS	Si la basura de la vivienda la recoge el camión o carrito de la basura, si la tiran en un basurero público, la queman, etc.
TENVIV	Si en la vivienda vive el dueño o propietario, la rentan o la ocupan en otra situación
FADQUI	Si el dueño o propietario de la vivienda la compró hecha, la mandó a construir, la construyó el mismo, etc.
ESTUFAG	Si la vivienda cuenta con estufa de gas
ESTUFAL	Si la vivienda cuenta con estufa de leña
TINACO	Si la vivienda cuenta con tinaco
BOILER	Si la vivienda cuenta con boiler
CISTERNA	Si la vivienda cuenta con cisterna
REGADERA	Si la vivienda cuenta con regadera
MEDLUZ	Si la vivienda cuenta con medidor de luz
RADIO	Si la vivienda cuenta con radio

Variable	Descripción
TELEVI	Si la vivienda cuenta con televisión
REFRIG	Si la vivienda cuenta con refrigerador
LAVADORA	Si la vivienda cuenta con lavadora
AUTOPROP	Si en la vivienda tienen automóvil o camioneta
COMPU	Si en la vivienda tienen computadora
TELEFONO	Si en la vivienda tienen teléfono
CELULAR	Si en la vivienda tienen teléfono celular
INTERNET	Si en la vivienda tienen internet
NUMPERS	Número de personas en la vivienda
TIPOHOG	Tipo de hogar censal, por ejemplo: familiar nuclear, familiar ampliado, familiar compuesto, etc.
MNUMPERS	Número de personas migrantes en la vivienda
COMIO1VEZ	Si por falta de dinero o recursos, alguna de las personas de la vivienda sólo comió una vez al día, en los últimos tres meses
NOCOMI1D	Si por falta de dinero o recursos, alguna de las personas de la vivienda dejó de comer todo un día, en los últimos tres meses
SINCOMER	Si por falta de dinero o recursos, alguna vez se quedaron sin comida, en los últimos tres meses
INGTRHOG	Ingresos mensuales por trabajo en el hogar

3.2.4. Aplicación de algoritmos de Minería de Datos

En este apartado se aplicaron diferentes algoritmos de agrupamiento, se eligió el mejor modelo de agrupamiento con base en los resultados del índice de validez de Davies - Bouldin y se le aplicaron diversos algoritmos de clasificación.

3.2.4.1. Agrupamiento

Para la parte de agrupamiento (clustering) se aplicaron 4 algoritmos, debido a que no se conocía a priori el número de clusters a formar. Se aplicó el algoritmo EM, del cual se obtuvieron 5 grupos; posteriormente se aplicó el algoritmo de Self Organizing Maps (SOM) dando como resultado 4 grupos. De acuerdo a estos resultados se decidió aplicar los algoritmos de Simple K-

means y Make Density Based Clusterer (MDBC) con valores para k de 3, 4 y 5 grupos.

A partir de los resultados obtenidos de los algoritmos antes mencionados, se extrajeron diversas gráficas de dispersión con el único propósito de mostrar la concentración de los individuos en los grupos formados.

Para mostrar las gráficas donde mejor se aprecia la separación de los grupos, se hizo una revisión manual de cada combinación *variable del cluster/variable de entrada* y se seleccionaron aquellas combinaciones en donde se observa la distribución de los grupos. Estas gráficas, son descritas a continuación, donde las variables aparecen en letra negrita y los valores de esas variables en letra cursiva.

La Figura 3.1 muestra la concentración de los grupos creados con el algoritmo EM con respecto a la variable **Boiler**. En el eje X se indica el número de **cluster** y en el eje Y *si tiene*(representado con el valor 3), *no tiene* (representado por el valor 4) o *no sabe* (representado por el valor 9). Como se puede observar el cluster 0 y el cluster 4 concentran a los objetos que *no tienen* boiler, el cluster 1 y el 3 a los que *si tienen*, y el cluster 2 concentra en la mayoría a los que *no tienen*, sin embargo tienen una cantidad importante de los que *si tienen* boiler.

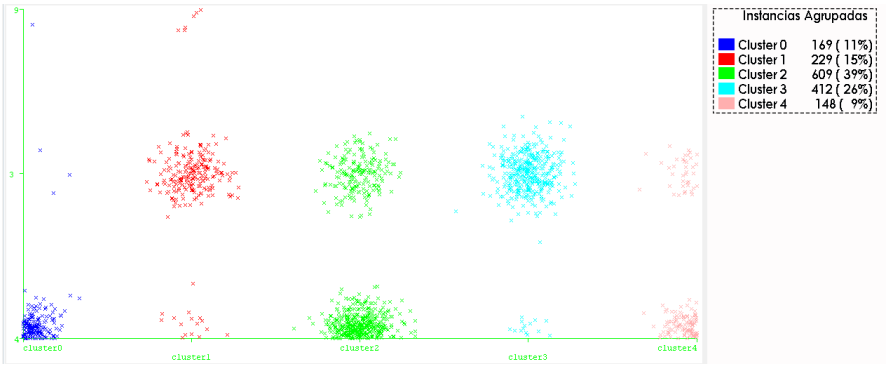


Figura 3.1: Diagrama de dispersión del resultado del algoritmo EM

Por su parte, la Figura 3.2 presenta la dispersión de los grupos creados con K-means, con respecto a la variable **Regadera**. En el eje X se indica el

número de **cluster** y en el eje *Y si tiene* (representado con el valor 3), *no tiene* (representado por el valor 4) o *no sabe* (representado por el valor 9). Como se puede observar el cluster 0 concentra a los objetos que *no tienen* regadera y los clusters 1 y 2 concentra a los que *si tienen*.

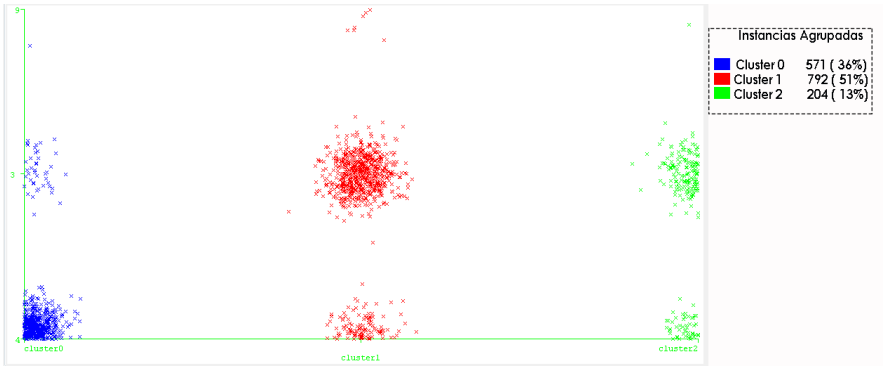


Figura 3.2: Diagrama de dispersión del resultado del algoritmo Simple K - Means con 3 grupos

En la Figura 3.3 se muestra la concentración de los grupos formados por MDBC con respecto a la variable **CONAGU, Tipo de descarga del servicio sanitario**. En el eje *X* se indica el número de **cluster** y en el eje *Y si tiene descarga directa de agua* (representado con el valor 5), *si le echan agua con cubeta* (representado por el valor 6), *si no se le puede echar agua* (representado por el valor 7), *si no tienen* (representado por el valor 0) o los *no especificado* (representado por el valor 9). Como se observa el cluster 0 se encuentra dividido entre los objetos que al servicio sanitario *tienen que echarle agua con cubeta* y en los que *no se les puede echar agua*, los clusters 1 y 2 concentra a los que *si tienen descarga directa de agua*, con algunos casos en los que al servicio sanitario *tienen que echarle agua con cubeta*.

Las gráficas anteriores muestran las variables con las que mejor se dividen los grupos, por tal motivo, éstas podrían ser consideradas para la caracterización de los grupos encontrados.

3.2.4.2. Aplicación del Índice de Validez

Para evaluar los resultados obtenidos de la parte de clustering, se utilizó el Índice de Validez de Clusters (CVI), en específico se utilizó el índice de va-

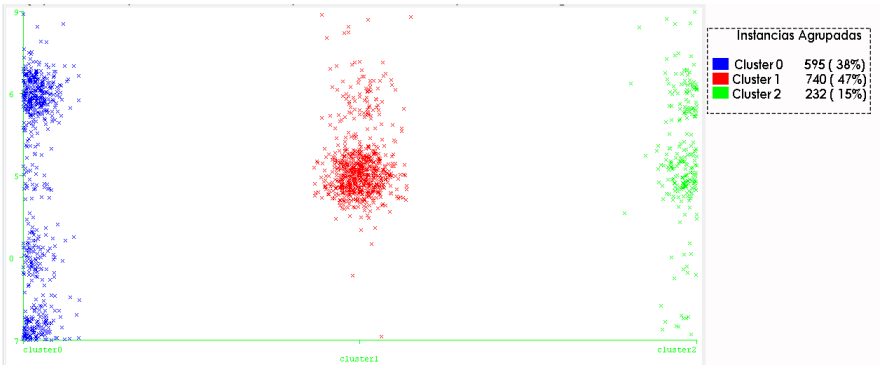


Figura 3.3: Diagrama de dispersión del resultado del algoritmo Make Density Based Clusterer con 3 grupos

lidez de Davies Bouldin. Este índice hace uso principalmente de la distancia euclidiana para estimar la cohesión entre los individuos de un grupo, su centroide y la separación entre centroides. Esta distancia es utilizada cuando las variables son de tipo numérico, sin embargo, nuestros datos son mezclados, es decir, contienen tanto atributos numéricos como categóricos. Por tal motivo y para poder procesar los datos de manera original, sin aplicarles alguna técnica de transformación, se decidió hacer una variación de este índice, cambiando la distancia Euclideana por la distancia de Gower, la cual permite hacer dicho cálculo cuando hay variables mezcladas (numéricas, categóricas, boleanas).

El índice de validez de Davies Bouldin se aplicó a los resultados obtenidos de los algoritmos EM, SOM, Simple K-means y Make Density Based Clusterer, los dos últimos con 3, 4 y 5 grupos. Los resultados de este índice para todos los algoritmos se muestran en la Tabla 3.2. EL valor N/A indica que para el algoritmo implicado, no se aplicó el índice de Davies Bouldin, ya que como se mencionó anteriormente, en los algoritmos EM y SOM el número de grupos no se configura de manera previa, sino que se obtiene al ejecutar dicho algoritmo.

En este índice, entre más bajo sea el valor obtenido, se considera un mejor agrupamiento, es decir, que los grupos que se obtuvieron son más consistentes y con mayor cohesión. Por lo tanto, el mejor modelo de agrupamiento es el que se obtuvo a través del algoritmo Make Density Based Clusterer (MDBC) de tres grupos con un valor de 2.137. Sin embargo, el valor obtenido con el

Tabla 3.2: Índices de Validez Davies - Bouldin en los diferentes algoritmos

Algoritmo	3 grupos	4 grupos	5 grupos
SOM	N/A	2.948	N/A
EM	N/A	N/A	4.358
Simple K-means	2.162	2.145	4.996
MDBC	2.137	2.228	5.064

algoritmo de Simple K-means con cuatro grupos fue también muy bajo 2.145, en comparación con los otros valores, por tal motivo se decidió analizar más a fondo la conformación de los grupos, antes de elegir la agrupación a caracterizar.

Para comparar ambos resultados, con tres y cuatro grupos, se revisó el porcentaje de instancias que fueron colocadas dentro de su grupo correspondiente, en Simple K - means y en Make Density Based Clusterer. Para los resultados de 4 grupos, mostrados en la Tabla 3.3, el 87.43 % (1370) de las instancias que formaban un grupo en el algoritmo de Simple K-means, quedaron agrupadas de la misma manera con el algoritmo de Make Density Based Clusterer. Con respecto a los resultados de 3 grupos, mostrados en la Tabla 3.4, se obtuvo que el 100 % de las instancias quedaron agrupadas de la misma manera por ambos algoritmos. Con esto se confirma que efectivamente el mejor resultado para este agrupamiento, es el obtenido de aplicar el algoritmo de Make Density Based Clusterer con 3 grupos.

Tabla 3.3: Comparación de las instancias que conforman los 4 grupos

% de instancias	No. de instancias	Grupos de Kmeans	No. de instancias del grupo	Grupos MDBC
89.36 %	529	C0	592	C0
8.61 %	51	C1		
1.01 %	6	C2		
1.01 %	6	C3		
91.37 %	646	C1	707	C1
1.41 %	10	C0		
4.10 %	29	C2		
3.11 %	22	C3		
69.77 %	120	C2	172	C2
11.63 %	20	C0		
15.70 %	27	C1		
2.91 %	5	C3		
78.13 %	75	C3	96	C3
2.08 %	2	C0		
16.67 %	16	C1		
3.13 %	3	C2		

Tabla 3.4: Comparación de las instancias que conforman los 3 grupos

% de instancias	No. de instancias	Grupos de Kmeans	No. de instancias del grupo	Grupos MDBC
100 %	571	C0	571	C0
0 %	0	C1		
0 %	0	C2		
100 %	792	C1	792	C1
0 %	0	C0		
0 %	0	C2		
100 %	204	C2	204	C2
0 %	0	C0		
0 %	0	C1		

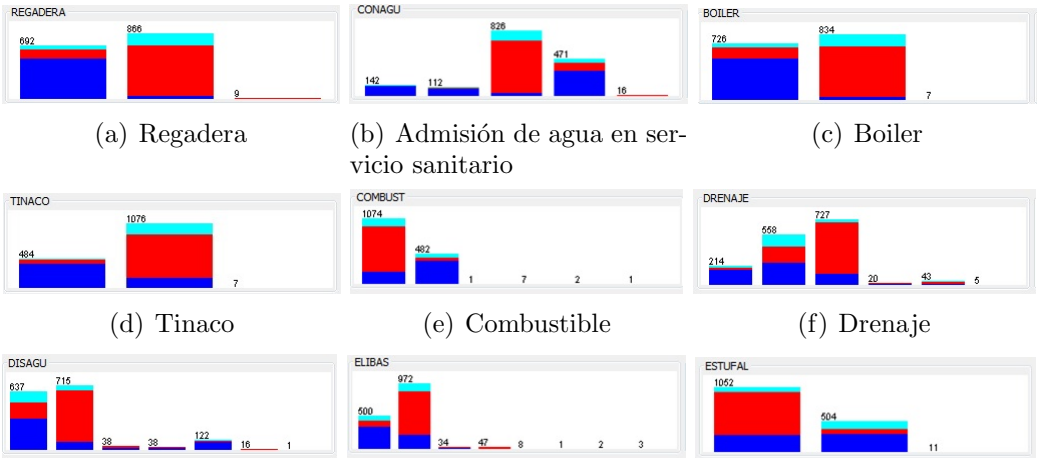
3.2.4.3. Cálculo del peso informacional de las variables

Se aplicó el método de InfoGainAttributeEval al mejor agrupamiento, con el fin de encontrar las variables con mayor peso informacional. Este método permite evaluar el valor de un atributo mediante la medición de la ganancia de información con respecto a la clase, utiliza el método Ranker que evalúa de manera individual a los atributos y los enlista de acuerdo a esa evaluación. En la Tabla 3.5 se muestran los valores obtenidos para cada variable del conjunto de datos utilizado, a través de este método.

Tabla 3.5: Ganancia Informacional

Ganancia Informacional	Variable	Ganancia Informacional	Variable
0.50444	CONAGU	0.47268	REGADERA
0.45498	BOILER	0.39688	MUN
0.31275	TINACO	0.29088	COMBUST
0.29058	DRENAJE	0.25818	DISAGU
0.19849	ELIBAS	0.18095	ESTUFAL
0.15824	TAM_LOC	0.14809	TOTCUART
0.14654	LAVADORA	0.14478	INGTRHOG
0.1365	PISOS	0.13438	CUADORM
0.12298	CELULAR	0.11901	REFRIG
0.11564	ESTUFAG	0.10955	TECHOS
0.10277	COMPU	0.09502	AUTOPROP
0.09329	DOTAGUAD	0.08068	TELEFONO
0.07614	USOEXC	0.07529	SERSAN
0.06795	TELEVI	0.05495	INTERNET
0.05157	MEDLUZ	0.04994	PAREDES
0.04427	MNUMPERS	0.04147	TIPOHOG
0.03987	ELECTRI	0.03658	FADQUI
0.03375	NUMPERS	0.03211	CLAVIVP
0.02688	COCINA	0.02026	MSEXO
0.02022	CISTERNA	0.01736	RADIO
0.0168	TENVIV	0.01342	MEDAD
0.00651	SINCOMER	0.0051	NOCOMI1D
0.00352	COMIO1VEZ		

En la Figura 3.4 se ilustra el comportamiento de las variables encontradas como más representativas de este conjunto de datos. Como se puede observar, se aprecia la distribución de los objetos en los tres grupos; el cluster 0 está representado con el color azul, el cluster 1 con el color rojo y el cluster 2 con el color turquesa.



(a) Regadera (b) Admisión de agua en servicio sanitario (c) Boiler

(d) Tinaco (e) Combustible (f) Drenaje

(g) Disponibilidad de agua (h) Eliminación de la basura (i) Estufa de Leña

Figura 3.4: Variables representativas del factor demográfico

En la sección 3.2.4.1, se mostró la gráfica de dispersión obtenida con el algoritmo Make Density Based Clusterer con 3 grupos, donde se apreciaba la separación de los grupos con respecto a la variable *CONAGU*. Hay que destacar que esta variable resultó ser la de mayor ganancia después de que se aplicó el calculo del peso informacional.

3.2.5. Evaluación e interpretación de los perfiles

Teniendo en cuenta la ganancia informacional de las variables, aplicada en la sección 3.2.4.3, se han analizado los resultados del agrupamiento para expresarlo en un conocimiento procesable y fácilmente comprensible por los expertos sociales. Esto ayudará a desarrollar acciones sociales y tomar decisiones dirigidas a mejorar la calidad de vida de los migrantes y sus familias, además, para disminuir o tratar de evitar la migración en el estado de Hidalgo. La caracterización de la partición obtenida, se describe a continuación, donde la etiqueta del cluster se encuentra con letra cursiva y negrita, las variables en letra cursiva y los valores de las variables con negrita.

- **Cluster 0:** ***Viviendas en exclusión y pobreza.*** Este grupo se caracteriza por tener viviendas que **no cuentan** con servicios básicos como *regadera, calentador de agua (boiler), tinaco o lavadora*; realizan la *descarga del servicio sanitario* **echándole agua con una cubeta** y el *drenaje* está conectado a una **fosa séptica**, cuentan con *la disponibilidad de agua* **entubada fuera de la vivienda** pero dentro del terreno, el *combustible para cocinar* que más utilizan es la **leña**, dado que *tienen estufa de leña* o carbón, y la *eliminación de la basura* la realizan a través de la **quema**. Estas viviendas se encuentran en su mayoría ubicadas en los *municipios* de **Pisaflores, Tlahuiltepa, Omitlán de Juárez, Tepehuacán de Guerrero, Tenango de Doria, Chapulhuacán, La Misión, Tlanchinol, Eloxochitlán, Lolotla, Nicolás Flores. Pacula, Molango de Escamilla, Tecozautla, Juárez Hidalgo, Calnali, Cardonal y Tianguistengo.**

- **Cluster 1:** ***Viviendas con servicios medios y desarrollo.*** Este grupo se encuentra caracterizado por tener viviendas que **cuentan** con servicios como *regadera, calentador de agua (boiler), tinaco y lavadora*; el *servicio sanitario* cuenta con **descarga directa de agua** y el *drenaje* está conectado a la **red pública**, cuentan con *disponibilidad de agua* **entubada dentro de la vivienda**, el *combustible para cocinar* que más utilizan es el **gas de cilindro** o de tanque estacionario, la *basura* es recolectada por el **camión de la basura** o por un carrito de basura y al menos una persona de la vivienda **cuenta** con *celular*. Estas viviendas se encuentran ubicadas en su mayoría en los *municipios* de **Tulancingo de Bravo, Francisco I. Madero, Tasquillo, Zimapán, Pachuca de Soto, Acatlán, Mixquiahuala de Juárez, Atotonilco el Grande, Tezontepec de Aldama, Progreso de Obregón, San Agustín Metzquititlán, Santiago Tulantepec de Lugo Guerrero, Cuautepec de Hinojosa, Metepec, Tlanalapa, Tolcayuca e Ixmiquilpan.**

- **Cluster 2:** ***Viviendas en marginación.*** Este grupo se caracteriza por tener viviendas que **cuentan** con servicios como *regadera, calentador de agua (boiler), tinaco y lavadora*; el *servicio sanitario* cuenta con **descarga directa** de agua y el *drenaje* está conectado a una **fosa séptica**, cuentan con *disponibilidad de agua* **entubada fuera de la vivienda** pero dentro del terreno, el *combustible para cocinar* que más utilizan es el **gas de cilindro** o de tanque estacionario, aunque tam-

> bién **cuentan** con *estufa de leña* o carbón, la *basura* es recolectada por el **camión de la basura** o por un carrito de basura. Estas viviendas se encuentran ubicadas en su mayoría en los *municipios* de **Huasca de Ocampo, Agua Blanca de Iturbide, Jacala de Ledezma, Alfajayucan, Chapantongo, Mineral del Chico, Huehuetla y Nopala de Villagrán,**

La Figura 3.5 muestra el mapa del estado de Hidalgo, donde se puede observar la ubicación de los tres grupos caracterizados, mencionados anteriormente. En este mapa se puede apreciar que el cluster *Viviendas en exclusión y pobreza* domina en la región norte; mientras tanto, en las regiones del este y del sureste del estado se ubica el cluster *Viviendas con servicios medios y en desarrollo.* Por último, la mayoría de las instancias del cluster *Viviendas en marginación* se encuentran distribuidas en el suroeste del estado de Hidalgo. Cabe mencionar que los resultados de la agrupación descritos aquí, se han analizado y validado por expertos en materia social en el fenómeno de migración. Su opinión refleja que los resultados son consistentes de acuerdo a su percepción de la migración en el estado de Hidalgo.

3.2.5.1. Clasificación

Una vez que se obtuvo el mejor agrupamiento y fue caracterizado, se aplicaron algoritmos de clasificación, con el objetivo de crear modelos y así identificar el mejor algoritmo que permita predecir los casos de migración en el estado de Hidalgo, con respecto al factor demográfico. Los algoritmos aplicados fueron árboles de decisión, máquinas de vectores de soporte, redes neuronales artificiales, Naive Bayes y KNN (con valores de 3, 5 y 7), llamados en Weka como J48, LibSVM, MultilayerPerceptron, NaiveBayes e IBk, respectivamente. Estos algoritmos se configuraron para que fueran evaluados por validación cruzada con 10 particiones. Los porcentajes de instancias clasificadas correctamente por estos algoritmos se encuentran ilustradas en la Figura 3.6.

Existen algunas métricas para calcular la exactitud de este tipo de modelos; entre la cuales se encuentra la precisión, la cobertura, el F-measure, los falsos positivos y los verdaderos positivos [19]. A excepción de los falsos positivos, estas medidas entre más cercano sea su valor a 1, indican un mejor modelo; en el caso de los FP, un modelo es mejor cuando esta medida es más

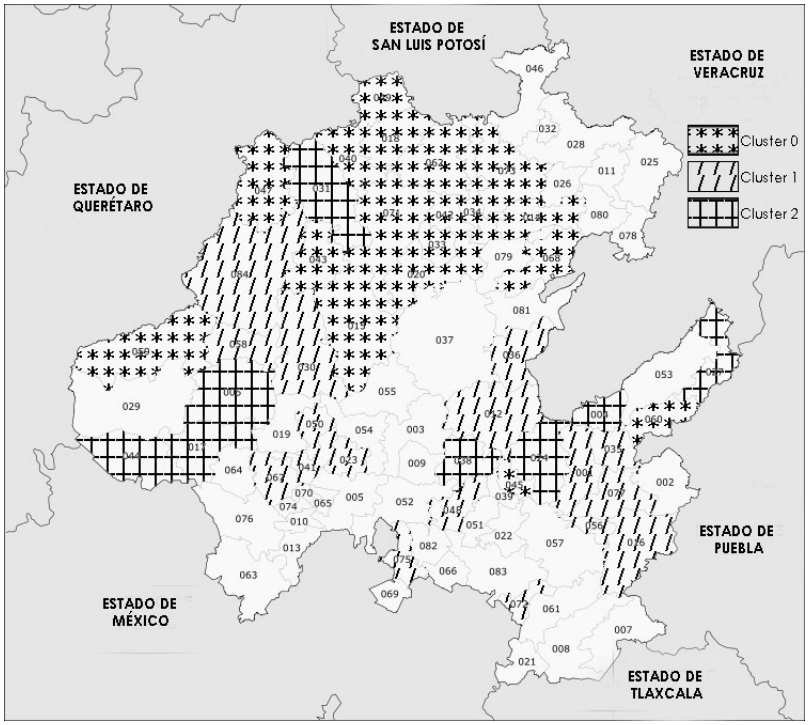


Figura 3.5: Distribución territorial en el estado de Hidalgo de los grupos encontrados

cercana a 0.

- *Precisión*: Mide el número de instancias correctamente clasificadas, con respecto al total de las instancias predichas; lo cual esta dado por la fórmula 3.1.

$$Precision = \frac{VP}{VP + FP} \tag{3.1}$$

- *Recall(Cobertura)*: Calcula la proporción de instancias correctamente clasificadas, con respecto al total de las instancias reales, como se muestra en la fórmula 3.2.

$$Cobertura = \frac{VP}{VP + VN} \tag{3.2}$$

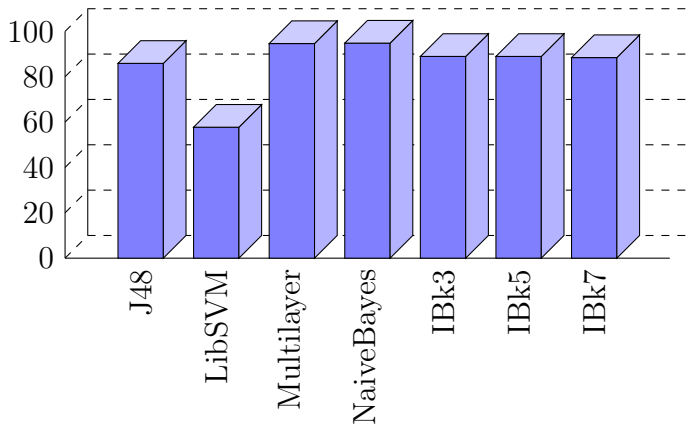


Figura 3.6: Porcentaje de instancias clasificadas correctamente

- *F-measure*: Determina un valor único, el cual permite caracterizar la eficacia de un algoritmo; este valor es calculado por la fórmula 3.3.

$$F = \frac{2Precision * Cobertura}{Precision + Cobertura} \tag{3.3}$$

- *True Positive Rate (Verdaderos Positivos, VP)*: Es la cantidad de instancias clasificadas en la clase x, que han sido clasificados como dicha clase.

- *False Positive Rate(Falsos Positivos, FP)*: Es la cantidad de instancias clasificadas en la clase x, pero que pertenecen a otra clase.

En la Tabla 3.6 se muestran las métricas antes mencionadas obtenidas como resultado de la aplicación de los diferentes algoritmos de clasificación. Como se puede observar, los algoritmos de Naive Bayes y Multilayer Perceptron son los que tienen las mejores métricas; sin embargo, Naive Bayes, tiene medidas más altas en precisión, F-measure, cobertura, verdaderos positivos y un valor menor en falsos positivos; por esta razón, se recomienda utilizar este proceso cuando se quieran predecir los casos de migración en el estado, desde el punto de vista del factor demográfico.

Tabla 3.6: Comparación entre los resultados de los diferentes algoritmos de clasificación

Algoritmo	Precisión	Recall	F-measure	TP Rate	FP Rate
J48	0.857	0.859	0.858	0.859	0.083
LibSVM	0.62	0.578	0.536	0.578	0.349
MultilayerPerceptron	0.947	0.946	0.947	0.946	0.028
NaiveBayes	0.951	0.948	0.949	0.948	0.022
IBk3	0.887	0.89	0.884	0.89	0.076
IBk5	0.889	0.89	0.882	0.89	0.08
IBk7	0.886	0.884	0.873	0.884	0.087

3.2.5.2. Obtención de reglas usando Weka

Con la finalidad de hacer más robusta la caracterización de los grupos, se aplicaron los algoritmos de reglas DTNB, FURIA, JRip, PART y Ridor, utilizando el software Weka; ya que considerar únicamente las variables de mayor peso y haciendo la caracterización manual, esta podría resultar subjetiva.

Se hizo un análisis de los resultados obtenidos de los algoritmos de reglas aplicados, sin embargo, se pudo observar que éstos no fueron satisfactorios, ya que los modelos resultantes contenían reglas confusas, por ejemplo, una regla indicaba la pertenencia a un grupo y posteriormente, la misma regla indicaba la pertenencia a otro grupo; además los modelos generados se componían de una gran cantidad de reglas, lo cual no es muy útil al momento de describir los grupos. Esto se puede observar en la Tabla 3.7.

Tabla 3.7: Resultados de la aplicación de Reglas con Weka

Algoritmo	% Instancias Clasificadas Correctamente	No. Reglas
DTNB	89.41 %	1492
FURIA	88.51 %	62
JRip	84.30 %	22
PART	85.83 %	52
Ridor	83.34 %	35

A pesar que el porcentaje de instancias clasificadas correctamente es alto, el número de reglas es también elevado, como en el caso del algoritmo DTNB

con el 89.41 % de instancias clasificadas correctamente, pero con 1492 reglas que describen a los grupos.

3.2.5.3. Aplicación del Algoritmo LR-FIR

Los resultados de la sección 3.2.5.2 llevaron a utilizar otra herramienta que permitiera obtener un conjunto de reglas más compacto, y que a su vez fueran comprensibles y fácilmente interpretables; el algoritmo elegido fue LR-FIR. Se utilizó el conjunto de datos completo (1567 registros) como entrenamiento (training), ya que no se pretendía probar el modelo, sino encontrar las reglas que describieran los grupos de migrantes de manera robusta y objetiva.

Se realizaron ocho experimentos con diferentes configuraciones, sin embargo, por la naturaleza del algoritmo algunas variables tuvieron que ser discretizadas para poder ejecutarlo correctamente; por ejemplo, la variable CONAGU se discretizó en 5 clases: 0 que significa *No tiene servicio sanitario* en el rango [0 - 0.5], el 5 que indica que *el servicio sanitario tiene descarga directa de agua* en el rango [0.5 - 5.5], el 6 que*al servicio sanitario le echan agua con cubeta* en el rango [5.5 - 6.5], el 7 que *al servicio sanitario no se le puede echar agua*[6.5 - 7.5] y el 9 que *no se especificó* en el rango [7.5 - 9.5]. Las variables de REGADERA, BOILER Y ESTUFAL, se discretizaron en dos clases, el 3 que indicaba *Si tiene* en el rango [0 - 3.5] y el 4 que indicaba *No tiene* en el rango [3.5 - 9.5]; la variable TINACO también fué discretizada en dos clases, el 1 que correspondía a *Si tiene* en el rango de [0 - 1.5] y el 2 que indicaba *No tiene* en el rango [1.5 - 9.5]; y la variable numérica CUADORM se encuentra en el rango [2-10].

Posteriormente, se realizaron cuatro experimentos, utilizando la *Mejora de la compactación* con la opción 1(APB) y posteriormente, con la opción 2(RB); en algunos experimentos se aplicó primero el Filtrado de Reglas y después la Unificación de Reglas y en los otros se aplicaron a la inversa.

Los parametros del algoritmo utilizados para estos experimentos fueron los siguientes: *Compactation: Minimal Ratio de 0.75, Desicion Unification Method: Wise Unification with Repetitions, Filtering: Percentage of filtering bad rules de 0.1, Otherwise rule: Desactivada.*

Para el primer experimento, se aplicó la opción 1 de Mejora de la compactación, se aplicó el filtrado de reglas y posteriormente la unificación de reglas; las reglas obtenidas son mostradas en la Tabla 3.8, donde en la primer columna aparece el número de regla, en la segunda columna la descripción de la regla, en la tercer columna la métrica de especificidad y en la cuarta columna la métrica de sensitividad.

Tabla 3.8: Experimento 1: Extracción de reglas lingüísticas para el factor demográfico de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI tiene boiler ES *No* Y la descarga del servicio sanitario ES *no se le puede echar agua* Y tienen regadera ES *No* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	1	0.16
2	SI tiene boiler ES *No* Y la descarga del servicio sanitario ES *con cubeta* Y tienen estufa de leña ES *Si* Y tienen regadera ES *No* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	0.99	0.14
	Métricas para el Cluster *Viviendas en exclusión y pobreza*	**0.99**	**0.98**
3	SI la descarga del servicio sanitario ES *directa* Y el número de cuartos que utilizan para dormir ES *de 2 a 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas con servicios medios y desarrollo***	0.88	0.61
	Métricas para el Cluster *Viviendas con servicios medios y desarrollo*	**0.88**	**0.61**
4	SI la descarga del servicio sanitario ES *directa* Y el número de cuartos que utilizan para dormir ES *de 2 a 3* Y tienen estufa de leña ES *Si* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.96	0.24
5	SI la descarga del servicio sanitario ES *directa* Y el número de cuartos que utilizan para dormir ES *más de 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.82	0.15
	Métricas para el Cluster *Viviendas en marginación*	**0.69**	**0.66**
	MÉTRICAS PARA EL MODELO	**0.89**	**0.76**

En el segundo experimento, mostrado en la Tabla 3.9, también se aplicó la opción 1 de Mejora de la compactación, pero a diferencia del primer experimento, se aplicó de manera inicial el primer paso de unificación de reglas, después el filtrado de reglas y posteriormente el segundo paso de la unificación de reglas.

El tercer experimento, mostrado en la Tabla 3.10, es similar al primero sólo que se utilizó la opción 2 de Mejora de la compactación.

En el cuarto experimento, mostrado en la Tabla 3.11, también se aplicó la opción 2 de Mejora de la compactación, aunque primero se aplicó el paso 1 de unificación de reglas, después el filtrado de reglas y el paso 2 de la unificación.

Tabla 3.9: Experimento 2: Extracción de reglas lingüísticas para el factor demográfico de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI la descarga del servicio sanitario ES *con cubeta* o *no se le puede echar agua* Y tienen regadera ES *No* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	0.97	0.52
2	SI tienen boiler ES *No* Y la descarga del servicio sanitario ES *con cubeta* Y tienen estufa de leña ES *Si* Y tienen regadera ES *No* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	0.99	0.14
	Métricas para el Cluster *Viviendas en exclusión y pobreza*	**0.97**	**0.96**
3	SI la descarga del servicio sanitario ES *descarga directa de agua* Y el número de cuartos que utilizan para dormir ES *más de 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster Es ***Viviendas con servicios medios y desarrollo***	0.96	0.32
4	SI la descarga del servicio sanitario ES *descarga directa de agua* Y el número de cuartos que utilizan para dormir ES *2* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster Es ***Viviendas con servicios medios y desarrollo***	0.9	0.29
	Métricas para el Cluster *Viviendas con servicios medios y desarrollo*	**0.92**	**0.58**
5	SI la descarga del servicio sanitario ES *No tiene* o *directa o con cubeta* Y el número de cuartos que utilizan para dormir es *más de 3* Y tienen estufa de leña ES *Si* Y tienen regadera ES *No* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	1	0.11
6	SI la descarga del servicio sanitario ES *directa o con cubeta* Y el número de cuartos que utilizan para dormir es *más de 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.82	0.19
7	SI la descarga del servicio sanitario ES *directa o con cubeta* Y el número de cuartos que utilizan para dormir es *de 2 a 3* Y tienen estufa de leña ES *Si* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.95	0.35
	Métricas para el Cluster *Viviendas en marginación*	**0.71**	**0.76**
	MÉTRICAS PARA EL MODELO	**0.91**	**0.75**

Las métricas de especificidad y sensitividad de cada modelo obtenido, se pueden apreciar en la Tabla 3.12, en donde la primer columna se refiere al número de experimento, la segunda columna muestra la métrica de especificidad y la tercer columna la métrica de sensitividad.

Se realizaron dos experimentos más, activando el parámetro de Otherwise rule; sin embargo las reglas obtenidas no fueron útiles para describir al grupo con *Otherwise*; por lo cual esos resultados no fueron incluídos.

Como se puede observar en la Tabla 3.12, las mejores métricas son las obtenidas en el experimento dos, ya que tiene un valor de especificidad de 0.91 y un valor de sensitividad de 0.75. Sin embargo, las del experimento cuatro son muy similares con 0.9 de especificidad y 0.75 de sensitividad. Al revisar las métricas de cada regla y por cluster, nos dimos cuenta de que son mejores las del experimento cuatro, además las reglas obtenidas son fáciles de comprender y muy intuitivas; por lo tanto se eligieron como el mejor modelo.

Tabla 3.10: Experimento 3: Extracción de reglas lingüísticas para el factor demográfico de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI la descarga del servicio sanitario ES *no se le puede echar agua* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	1	0.17
2	SI tienen boiler ES *No* Y SI la descarga del servicio sanitario ES *con cubeta* Y tienen estufa de leña ES *Si* Y tienen regadera ES *No* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	0.99	0.14
	Métricas para el Cluster *Viviendas en exclusión y pobreza*	**0.99**	**0.98**
3	SI la descarga del servicio sanitario ES *directa* Y el número de cuartos que utilizan para dormir ES *de 2 a 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas con servicios medios y desarrollo***	0.88	0.61
	Métricas para el Cluster *Viviendas con servicios medios y desarrollo*	**0.88**	**0.61**
4	SI la descarga del servicio sanitario ES *directa* Y el número de cuartos que utilizan para dormir ES *de 2 a 3* Y tienen estufa de leña ES *Si* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.96	0.24
5	SI la descarga del servicio sanitario ES *directa* Y el número de cuartos que utilizan para dormir ES *más de 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.82	0.15
	Métricas para el Cluster *Viviendas en marginación*	**0.69**	**0.66**
	MÉTRICAS PARA EL MODELO	**0.89**	**0.76**

Para el cluster ***Viviendas en exclusión y pobreza***, solo existe una regla que dice que si en el servicio sanitario se le echa agua con cubeta o no se le puede echar agua y que no tienen tinaco, entonces pertenecen a este cluster. La regla dos dice que si al servicio sanitario se le echa agua con cubeta o que si tiene descarga directa de agua, y si utilizan de 2 a 3 cuartos para dormir, y si no tienen estufa de leña, pero si tienen regadera y tinaco, entonces pertenecen al cluster ***Viviendas con servicios medios y desarrollo.***

La regla tres es similar a la regla dos, pero las hace diferentes la variable que indica el número de cuartos que utilizan para dormir, ya que en la regla dos es de 2 a 3 cuartos y en la regla tres es más de 3, por lo tanto ese valor en combinación con las demás variables hace que los individuos que cumplan con esa regla pertenezcan al cluster ***Viviendas en marginación.***

La regla cuatro también es similar a la regla dos, sólo que en la variable que indica si tienen estufa de leña el valor es Si, por lo tanto, ese valor combinado con el resto de las variables hacen que los individuos pertenezcan al cluster ***Viviendas en marginación.***

Cabe señalar que las reglas del experimento cuatro coinciden con la des-

Tabla 3.11: Experimento 4: Extracción de reglas lingüísticas para el factor demográfico de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI la descarga del servicio sanitario ES *con cubeta o no se le puede echar agua* Y tienen tinaco ES *No*, ENTONCES Cluster ES ***Viviendas en exclusión y pobreza***	0.97	0.97
	Métricas para el Cluster *Viviendas en exclusión y pobreza*	**0.97**	**0.97**
2	SI la descarga del servicio sanitario ES *directa o con cubeta* Y el número de cuartos que utilizan para dormir ES *de 2 a 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas con servicios medios y desarrollo***	0.9	0.6
	Métricas para el Cluster *Viviendas con servicios medios y desarrollo*	**0.9**	**0.6**
3	SI la descarga del servicio sanitario ES *directa o con cubeta* Y el número de cuartos que utilizan para dormir ES *más de 3* Y tienen estufa de leña ES *No* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.82	0.19
4	SI la descarga del servicio sanitario ES *directa o con cubeta* Y el número de cuartos que utilizan para dormir ES *de 2 a 3* Y tienen estufa de leña ES *Si* Y tienen regadera ES *Si* Y tienen tinaco ES *Si*, ENTONCES Cluster ES ***Viviendas en marginación***	0.95	0.35
	Métricas para el Cluster *Viviendas en marginación*	**0.73**	**0.68**
	MÉTRICAS PARA EL MODELO	**0.9**	**0.75**

Tabla 3.12: Especificidad y Sensitividad de los modelos generados en la extracción de las reglas lingüísticas

No. de Exp.	Esp.	Sen.
1	0.89	0.76
2	0.91	0.75
3	0.89	0.76
4	0.90	0.75

cripción manual que se realizó en la sección 3.2.5, donde se consideraron las variables con mayor ganancia informacional y se describe cada uno de los grupos.

Capítulo 4

Análisis del factor social de los migrantes

Este capítulo muestra la metodología KDD aplicada para el análisis del factor social que describe a los migrantes del estado de Hidalgo. Se describen los datos utilizados, así como los resultados obtenidos en cada paso.

4.1. Fuentes de información para el factor social

Las fuentes de información mencionadas en el capítulo anterior, en la sección 3.1, contienen preguntas relacionadas a la migración y a las condiciones de las viviendas de los migrantes; sin embargo, no poseen información propia del migrante. En esta investigación además de analizar el factor demográfico de los migrantes, se busca analizar el factor social, por ello se utilizaron los datos de la Encuesta sobre Migración en la Frontera Norte de México.

4.1.1. Encuesta sobre Migración en la Frontera Norte de México

A principios de los años 90's, el Colegio de la Frontera Norte (COLEF), realizó el estudio "Migración Interna e Internacional en México", como un encargo de la STPS y CONAPO; de dicho proyecto nace el compromiso de las tres instituciones participantes, en realizar la Encuesta sobre Migración en la Frontera Norte de México (EMIF NORTE), la cual se aplicó por primera

vez en el año de 1993. A partir de 1999, se ha aplicado de manera continua, incorporándose posteriormente el Instituto Nacional de Migración (INM), en el 2004 la Secretaría de Relaciones Exteriores (SRE) y a partir del 2011, la Secretaría de Salud (SS) [63].

Esta encuesta comprende un procedimiento de muestreo de poblaciones móviles, convirtiendo a las localidades fronterizas como observatorios naturales de los desplazamientos internacionales. En México, esta encuesta es única en su tipo, ya que su metodología y marco muestral están diseñados para captar personas en movimiento y de esa manera obtener características más detalladas del fenómeno bajo estudio, como sus aspectos sociodemográficos, trayectoria migratoria y laboral, motivos de la emigración, condiciones en las que se da su desplazamiento, el origen y el destino de la migración, los riesgos del cruce fronterizo, etc.; lo cual no es posible obtener con cuestionarios diseñados para aplicar en hogares. Además, la información es directa del migrante, ya que es entrevistado en el momento de su desplazamiento.

La EMIF - Norte capta los desplazamientos de la población que se dan en dirección de norte a sur y viceversa, dando como resultado de esa movilidad la medición de tres grandes flujos migratorios: los migrantes procedentes del sur, con destino a la frontera norte y con destino a Estados Unidos; los migrantes procedentes del norte, con procedencia de la frontera norte o con procedencia de Estados Unidos (que abarca a los migrantes que ingresan por vía terrestre o por vía aérea); y los migrantes devueltos por las autoridades migratorias de Estados Unidos. Para cada flujo migratorio, se describen las poblaciones objetivo a continuación:

a) Migrantes Procedentes del Sur: Son migrantes de 15 años o más, nacidos y residentes en México, que no viven en la ciudad de la entrevista, la razón de su desplazamiento se debe a motivos laborales, cambio de residencia, etc. y no tienen fecha precisa de regreso a su lugar de residencia habitual. De acuerdo a su destino final, se dividen de la siguiente manera:

- **Migrantes Procedentes del Sur con destino a la Frontera Norte**: Son aquellos que tienen como destino final alguna ciudad de los estados de la Frontera Norte de México.
- **Migrantes Procedentes del Sur con destino a Estados Unidos**: Son aquellos migrantes que tienen como destino final alguna

ciudad de los Estados Unidos.

b) **Migrantes Procedentes del Norte**: Personas de 15 años o más, nacidas en México y residentes en México o Estados Unidos, que no viven en la ciudad de la entrevista y que proceden de la Frontera Norte o de Estados Unidos, la razón de su desplazamiento se debe a motivos laborales, cambio de residencia, etc., siempre y cuando su estancia haya sido superior a un mes. De acuerdo a la ciudad donde hayan permanecido la mayor parte del tiempo, se dividen en:

   - **Migrantes Procedentes de la Frontera Norte**: Son aquellas personas procedentes del Norte, residentes en México, que viajan al interior del país y que permanecieron la mayor parte de su estancia en una ciudad de la Frontera Norte de México, a pesar de que hayan cruzado a Estados Unidos.
   - **Migrantes Procedentes de Estados Unidos**: Son aquellas personas procedentes del Norte que viajan al interior del país por vía terrestre o aérea. En el caso de que viajen por vía terrestre, su estancia mayor haya sido en una ciudad estadounidense, aunque hayan estado en alguna ciudad de la Frontera Norte de México.

c) **Migrantes devueltos por las autoridades migratorias de Estados Unidos**: Son individuos con 15 años o más, nacidas en México y residentes en México o Estados Unidos, devueltas por las autoridades migratorias de Estados Unidos a las autoridades migratorias de México en alguno de los puntos establecidos a lo largo de la línea fronteriza.

La EMIF - Norte se aplica en once ciudades y cuatro aeropuertos, las ciudades están distribuidas en tres regiones de muestreo: la Región Este, que incluye a Matamoros, Reynosa, Nuevo Laredo, Piedras Negras y Ciudad Acuña; la Región Centro, conformada por Ciudad Juárez, Agua Prieta, Nogales y Altar; y la Región Oeste, que abarca a Mexicali y Tijuana; en cuanto a los aeropuertos, consideraron por su gran número de vuelos a Estados Unidos los siguientes: Aeropuerto Internacional Lic. Benito Juárez de la Ciudad de México, Aeropuerto Internacional Miguel Hidalgo de Guadalajara, Jalisco, Aeropuerto Internacional del Bajío, en León, Guanajuato, Aeropuerto Internacional Gral. Francisco J. Mújica de Morelia, Michoacán.

Para el análisis del factor social de los migrantes del estado de Hidalgo, se utilizaron los datos de la Encuesta sobre Migración en la Frontera Norte de México (EMIF) del año 2012, en específico los datos de los migrantes provenientes del sur con destino a Estados Unidos; ya que estos datos contienen información propia del migrante, su situación laboral antes de migrar, su condición migratoria, etc. Estas características permiten describir el perfil de los migrantes del estado de Hidalgo.

4.2. Experimentos y resultados

En esta sección se describen los experimentos y los resultados obtenidos, en cada una de las fases realizadas de la metodología KDD.

4.2.1. Escenario de la investigación

Para el análisis del factor social, se adoptó el mismo escenario descrito en la sección 3.2.1.

4.2.2. Integración y recopilación

En esta fase se recopilaron los datos necesarios para el análisis del factor social, para ello se tomaron los datos de la EMIF- Norte del año 2012, correspondientes al flujo migratorio *Provenientes del sur*. Se filtraron los datos para que fueran los correspondientes a los migrantes nacidos en el estado de Hidalgo y que el lugar de destino fuera Estados Unidos, considerando las siguientes variables: P9_PAI y P9_EST, que se refieren a las preguntas de *¿En qué país nació usted?* y *¿En qué estado nació usted?*, respectivamente; también se consideró la variable LUG_DES, que corresponde al lugar de destino.

4.2.3. Limpieza y transformación

En esta fase los datos se exportaron a un servidor de MySQL, versión 5.0.51a. Se excluyeron las variables que no contenían información o que tenían el mismo valor en todos los registros; a las preguntas que contenían respuestas de "No Sabe", se les dio un formato que fuera congruente con el resto de las respuestas, a través de la creación de vistas en MySQL, por ejemplo, a las variables de *P12_5 Horas diarias que trabajó en promedio*, *P12_6 Días*

a la semana que trabajó, etc. tenían un valor de 98, cuando el resto de las respuestas eran valores con máximos entre 18 y 7 respectivamente, por lo que se les asignó desde la vista de MySQL el valor de 0 cuando se refería a "No sabe". Las variables relacionadas a tiempo y valores monetarios fueron procesadas para que su descripción correspondiera a la misma unidad; por ejemplo, como las variables P11_3_1C *Tiempo en que buscó trabajo - CANTIDAD* y P11_3_1T *Tiempo en que buscó trabajo - TIEMPO*, se combinaron en la variable P11_3_1C indicando que el tiempo era en meses; otro ejemplo fue el caso de las variables P12_7C *Salario en ese trabajo - CANTIDAD*, P12_7U *Salario en ese trabajo - UNIDAD* y P12_7T *Salario en ese trabajo - TIEMPO* donde se convirtieron los valores monetarios a pesos y los valores de tiempo a semanas, obteniendo así una sola variable P12_7C que indica el salario en pesos y a la semana.

Finalmente se obtuvieron 327 registros y 127 variables, las cuales se describen en la Tabla 4.1.

Tabla 4.1: Descripción de las variables para el análisis del factor social

Variable	Descripción
SEXO	Sexo del migrante
EDAD	Años cumplidos del migrante
RAZON	Razón por la cual se debe su visita a la zona fronteriza, por ejemplo: estudio, turismo, en tránsito hacia EU.
TRABAJA	Si tiene trabajo en su lugar de procedencia
P1	Si viene solo o acompañado en este viaje
P1_1	Número de personas que lo acompañan
P1_2	Número de acompañantes menores a 15 años
P1_3	Número de acompañantes que son sus padres, hermanos, hijos o esposa
P1_4	Número de acompañantes nacidos en México
P1_5	Número de acompañantes mujeres
P2	Si habla algún dialecto o lengua indígena
P3	Si sabe leer y escribir
P4N	Nivel de escuela
P4_1	Si ese último nivel que aprobó, fue en México, Estados Unidos o en otro país
P5	Si habla inglés

Variable	Descripción
P5_1	Nivel de inglés, por ejemplo: bien, muy bien, no habla
P6	Estado civil, por ejemplo: Soltero, casado, en unión libre
P7	Si en su casa es el jefe del hogar
P7_1	Relación con el jefe del hogar
P8	Número total de personas que viven en su casa
P8_1	Número total de personas que viven en su casa, menores de 15 años
P8_2	Número total de personas que viven en su casa, que trabajan
P8_3	Número total de personas que viven en su casa, que aportan un ingreso económico al hogar
P9_MUN	Municipio de nacimiento
P9_1	Si el lugar donde vive es el mismo donde nació
P10_PAI	País donde vive
P10_EST	Estado donde vive
P10_MUN	Municipio donde vive
P10_LOC	Localidad donde vive
P11	Si ha trabajado en el lugar donde vive o en algún lugar cercano
P11_1	Razón por la que no ha trabajado en ese lugar
P11_2	Durante los 30 días anteriores al inicio de este viaje, si trabajó en alguno de esos lugares
P11_3	Motivo por el que no trabajó, por ejemplo: Si tenía trabajo pero no trabajó, estaba buscando trabajo, era estudiante, etc.
P11_3_1C	Tiempo en que buscó trabajo
P11_4	Razón principal por la que dejó su último trabajo, por ejemplo: bajos ingresos, lo despidieron, por buscar trabajo en EU, etc.
P12_1	Nombre del oficio o profesión que desempeñó en ese trabajo
P12_2	Puesto o posición que tenía en ese trabajo, por ejemplo: trabajador a sueldo fijo, trabajador por obra, patrón, etc.
P12_3	Si firmó contrato de trabajo con el patrón o empresa, al momento de ser contratado
P12_4	Si tenía alguna prestación o beneficio, en ese trabajo

Variable	Descripción
P12_4_1	Prestación que tenía: Servicios de salud, servicios de salud y otras prestaciones, sin servicios de salud pero con otras prestaciones
P12_5	Horas diarias que trabajó en promedio
P12_6	Días a la semana que trabajó
P12_7C	Salario en ese trabajo
P12_8	Si recibió algún curso de capacitación en la empresa donde trabajó
P12_9	Número de personas que laboraban en el establecimiento donde trabajó
P12_10	Giro del establecimiento, negocio, fábrica o empresa donde trabajó
P12_11	Lugar donde realizó sus actividades laborales, por ejemplo: local establecido, puesto semifijo, puesto móvil, etc.
P12_12	Manera en la que aprendió el oficio al que se dedicó, por ejemplo: en la escuela, recibió curso de capacitación, por medio del desempeño, etc.
P13	Si visitó a un médico o un centro de salud como preparativo para el viaje
P13_1_1	Si, antes de emprender el viaje, alguna autoridad o institución le proporcionó alguna información sobre la prevención de accidentes
P13_1_2	Si, antes de emprender el viaje, alguna autoridad o institución le proporcionó alguna información sobre la prevención de adicciones
P13_1_3	Si, antes de emprender el viaje, alguna autoridad o institución le proporcionó alguna información sobre el manejo de enfermedades
P13_1_4	Si, antes de emprender el viaje, alguna autoridad o institución le proporcionó alguna información sobre la deshidratación
P13_1_5	Si, antes de emprender el viaje, alguna autoridad o institución le proporcionó alguna información sobre el manejo de cartilla de salud
P13_2	Si durante el viaje a la frontera recibió atención de salud ya sea por enfermedad, lesión o accidente

Variable	Descripción
P13_3	Causa por la que no recibió atención en salud durante el viaje
P13_4	Si durante el viaje sufrió una lesión tal como: cortada, herida, hueso roto, etc.
P13_5	Causa de la lesión, por ejemplo: choque de vehículos, atropellamiento, etc.
P14	Número de veces que ha llegado a ésta u otra ciudad fronteriza por la misma razón que en este viaje, en los últimos 12 meses
P15	Si es la primera vez que está en la ciudad fronteriza, donde se le encuestó, por la misma razón
P15_2	Si la última vez que estuvo en esa ciudad fronteriza, tenía familiares o amigos
P15_2_1	Si la última vez que estuvo en esa ciudad fronteriza, le proporcionaron préstamo monetario
P15_2_2	Si la última vez que estuvo en esa ciudad fronteriza, le proporcionaron alojamiento y/o alimentos
P15_2_3	Si la última vez que estuvo en esa ciudad fronteriza, le proporcionaron ayuda para conseguir trabajo
P15_2_4	Si la última vez que estuvo en esa ciudad fronteriza, ellos lo emplearon
P15_2_5	Si la última vez que estuvo en esa ciudad fronteriza, le proporcionaron ayuda para cruzar a Estados Unidos
P15_2_6	Si la última vez que estuvo en esa ciudad fronteriza, le proporcionaron otra ayuda
P15_3C	Tiempo que permaneció, la última vez, en esa ciudad fronteriza
P16C	Tiempo que piensa permanecer, en esta ocasión, en la ciudad fronteriza
P16_1	Si en esta ocasión tiene familiares o amigos en la ciudad fronteriza
P16_2	Lugar dónde piensa pasar la noche, el día de la entrevista
P16_3	Si va a trabajar o buscar trabajo en la ciudad
P16_3_1	Sector de la economía donde piensa trabajar en la ciudad fronteriza, por ejemplo: actividades agropecuarias, construcción, industria, comercio, etc.
P17_2	Ciudad mexicana por la que piensa cruzar

Variable	Descripción
P17_3	Razón por la que eligió esa ciudad para cruzar
P18	Si contrató a alguna persona (coyote, pollero, guía, lanchero, patero) para que lo ayude a cruzar la frontera
P18_1	Lugar donde hará el contacto con la persona que lo ayudará a cruzar la frontera
P19	Razón por la cual cruzará al otro lado
P20	Si en su destino final tiene un trabajo ya asegurado
P20_1	Sector de la economía donde piensa trabajar en EU, por ejemplo: actividades agropecuarias, construcción, industria, comercio, etc.
P21	Si se dirige a alguna ciudad en especial
P21_1C	Ciudad a la que se dirige
P21_1O	Condado al que se dirige
P21_1E	Estado al que se dirige
P21_2_1	Si en esa ciudad vive su esposa (o) o pareja
P21_2_2	Si en esa ciudad viven sus hijos (as)
P21_2_3	Si en esa ciudad vive su padre y/o madre
P21_2_4	Si en esa ciudad viven sus hermanos (as)
P21_2_5	Si en esa ciudad viven otros familiares
P21_3	Si en su destino final, tiene un lugar fijo donde llegar
P22C	Tiempo que piensa quedarse en Estados Unidos
P23	Si tiene documentos para cruzar
P23_1	Documento que tiene para cruzar, por ejemplo: Visa de negocios, visa de estudiante, etc.
P23_2C	Antigüedad del documento para cruzar
P24	Si tiene documentos para trabajar
P24_1	Documento que tiene para trabajar, por ejemplo: permiso temporal, tarjeta verde, etc.
P24_2C	Antigüedad del documento para trabajar
P25	Si ha realizado algún trámite oficial con la finalidad de obtener documentos para: entrar a EU, trabajar en EU, quedarse a vivir allá, etc.
P26	Total de veces que ha cruzado a Estados Unidos para trabajar o buscar trabajo
P26_2	Si en otra ocasión, usó algún tipo de documento para cruzar a Estados Unidos
P28	Ciudad mexicana por la cual cruzó, en esa última vez

Variable	Descripción
P29	Si esa última vez contrató a alguna persona (coyote, pollero, guía, lanchero, patero) para que lo ayudara a cruzar la frontera
P30	Si esa última vez usó algún tipo de documento para cruzar a Estados Unidos
P31	Si en esa última ocasión que cruzó a Estados Unidos, llevaba algún documento para trabajar
P32C	Tiempo que permaneció en Estados Unidos, en esa última ocasión
P33	Estado en el que estuvo la mayor parte del tiempo, esa última vez
P33_1	Si tenía familiares o amigos en ese estado, en esa ocasión
P34	Si en esa ocasión, trabajó en Estados Unidos
P34_1	Cantidad de trabajos diferentes que tuvo en Estados Unidos
P35_1	Nombre del oficio o profesión que desempeñó en ese trabajo
P35_2	Cantidad de horas diarias que trabajó en promedio
P35_3	Cantidad de días a la semana que trabajó
P35_4	Giro del establecimiento, negocio, fábrica o empresa donde trabajó en EU
P36	Razón por la cuál regresó a México
P37	Si considera su estado de salud: Muy bueno, bueno, regular, malo o muy malo
P38_1	Si tiene derecho a los servicios médicos del Seguro Popular
P38_2	Si tiene derecho a los servicios médicos del IMSS
P38_3	Si tiene derecho a los servicios médicos del ISSSTE
P38_4	Si tiene derecho a los servicios médicos de Oportunidades
P38_5	Si tiene derecho a los servicios médicos del otra institución
P40_1	Si algún médico o profesional de la salud le ha dicho que tiene hipertensión (Presión Alta)
P40_2	Si algún médico o profesional de la salud le ha dicho que tiene diabetes

Variable	Descripción
P40_3	Si algún médico o profesional de la salud le ha dicho que usted tiene colesterol (Problemas de grasas en sangre)
P41_1	Si conoce el programa de salud: Ventanillas de salud, que el gobierno mexicano tiene para los migrantes
P41_2	Si conoce el programa de salud: Vete Sano, Regresa Sano, que el gobierno mexicano tiene para los migrantes
P41_3	Si conoce el programa de salud: Salud del Migrante, que el gobierno mexicano tiene para los migrantes
P41_4	Si conoce el programa de salud: Asistencia a repatriados, que el gobierno mexicano tiene para los migrantes

4.2.4. Aplicación de algoritmos de Minería de Datos

En este apartado se aplicaron algoritmos de clustering, se validaron los resultados obtenidos con el Índice de Validez de Clusters de Davies - Bouldin y se eligió el mejor modelo de agrupamiento; finalmente, se obtuvo el peso informacional de las variables que componen a este conjunto de datos.

4.2.4.1. Agrupamiento

Para este punto, como en el capítulo anterior, se comenzó aplicando el algoritmo EM con el cual se obtuvieron 3 grupos, posteriormente se aplicó el algoritmo SOM y se obtuvieron 4 grupos; por lo tanto se decidió aplicar los algoritmos de Simple K-means y MDBC, con una configuración de 2, 3 y 4 grupos.

4.2.4.2. Aplicación del Índice de Validez

Todos los resultados obtenidos fueron evaluados por el Índice de Validez de Clusters de Davies - Bouldin, tal como se especificó en el capítulo anterior. En la Tabla 4.2 se muestran los valores de este índice, para las particiones producidas de cada algoritmo; el valor N/A indica que no aplica para el algoritmo en cuestión con esa configuración.

El valor más bajo, es el mostrado con Simple K - means con 4 grupos con 1.886, significando así el mejor agrupamiento para este conjunto de datos. Al analizar más profundamente la conformación de los grupos y comparando las

Tabla 4.2: Índices de Validez Davies - Bouldin en los diferentes algoritmos

Algoritmo	2 grupos	3 grupos	4 grupos
SOM	N/A	N/A	2.641
EM	N/A	2.271	N/A
Simple K-means	2.155	1.944	1.886
MDBC	2.155	1.957	1.913

particiones de Simple K - means y MDBC con 3 grupos, mostrado en la Tabla 4.3 resultó que el 98.78 % (323) de las instancias que formaban un grupo con Simple K-means permanecían en su grupo correspondiente, al agruparse con MDBC; mientras que al comparar dichos algoritmos con los resultados de la configuración de 4 grupos, mostrados en la Tabla 4.4, permanecían el 98.47 % (322) de las instancias agrupadas de la misma manera. Como puede observarse, la diferencia entre ambas medidas es mínima. Además, se consultó a los expertos en migración, qúienes opinaron que los 4 grupos obtenidos con Simple K - means, son los que mejor se adecúan a la realidad, confirmando lo obtenido con el índice de validez.

Tabla 4.3: Comparación de las instancias que conforman los 3 grupos

% de instancias	No. de instancias	Grupos de Kmeans	No. de instancias del grupo	Grupos MDBC
90.48 %	38	C0	42	C0
4.76 %	2	C1		
4.76 %	2	C2		
100 %	83	C1	83	C1
0 %	0	C0		
0 %	0	C2		
100 %	202	C2	202	C2
0 %	0	C0		
0 %	0	C1		

4.2.4.3. Cálculo del peso informacional de las variables

Se aplicó el método de InfoGainAttributeEval para obtener los atributos con mayor peso informacional, los cuales fueran considerados para caracterizar a los grupos formados por K-means. Estos resultados se muestran en la Tabla 4.5.

Tabla 4.4: Comparación de las instancias que conforman los 4 grupos

% de instancias	No. de instancias	Grupos de Kmeans	No. de instancias del grupo	Grupos MDBC
90 %	18	C0	20	C0
5 %	1	C1		
0 %	0	C2		
5 %	1	C3		
85.71 %	18	C1	21	C1
0 %	0	C0		
14.29 %	3	C2		
0 %	0	C3		
100 %	202	C2	202	C2
0 %	0	C0		
0 %	0	C1		
0 %	0	C3		
100 %	84	C3	84	C3
0 %	0	C0		
0 %	0	C1		
0 %	0	C2		

Tabla 4.5: Ganancia Informacional

Ganancia	Variable	Ganancia	Variable
0.96124	P35_1	0.95874	P35_4
0.93767	P33	0.92562	P28
0.92433	P36	0.91071	P26_2
0.90091	P30	0.8928	P34_1
0.8928	P34	0.8928	P35_3
0.8928	P35_2	0.89232	P29
0.88609	P33_1	0.88609	P31
0.88496	P32C	0.88496	P26
0.62965	P12_1	0.60952	P21_1C
0.60381	P12_10	0.56967	P11_1
0.54583	P11_3	0.53741	P11_3_1C
0.53342	P12_2	0.53156	P12_9
0.53129	P11_4	0.5191	P12_3
0.51675	P12_11	0.50998	P11_2
0.50402	P12_12	0.50243	P12_8
0.50106	P12_4_1	0.49932	P12_4
0.49757	P12_5	0.49757	P12_6
0.49757	P12_7C	0.4627	P11
0.42354	P21_1O	0.40023	P9_MUN
0.2374	P21_1E	0.17784	P17_2
0.16112	P5_1	0.15264	P20
0.1448	P20_1	0.14008	P21_3

Ganancia	Variable	Ganancia	Variable
0.13038	P16_2	0.12218	P10_MUN
0.11871	P5	0.117	P19
0.11297	P18_1	0.08523	P21_2_2
0.08363	P23_1	0.07656	P24
0.07656	P24_1	0.07572	P4N
0.07504	P25	0.07187	P21_2_3
0.06968	P17_3	0.06763	P21_2_4
0.06672	P21_2_5	0.06396	P21_2_1
0.06309	SEXO	0.06055	P24_2C
0.05865	P7_1 0.05624	P21	
0.05345	P6	0.05183	P16_1
0.04966	RAZON	0.04919	EDAD
0.04612	P23	0.04482	P10_EST
0.03975	P15_2_2	0.03752	P16_3
0.03752	P16_3_1	0.03621	P7
0.03527	P18	0.03471	P40_2
0.03133	P15_2_3	0.03133	P15_2_5
0.03133	P15_2	0.03133	P15_2_1
0.03133	P15_2_6	0.03133	P15_2_4
0.03085	P38_4	0.02996	P15
0.02966	P1_1	0.02966	P1_4
0.0289	P37	0.02558	P10_LOC
0.02556	P38_3	0.02427	P41_2
0.02252	P3	0.02252	P4_1
0.02084	P41_4	0.02055	P10_PAI
0.02055	P9_1	0.02028	P1
0.01965	P38_1	0.0164	P13
0.0139	P41_3	0.01351	P13_1_2
0.01239	P2	0.0122	P13_1_1
0.00969	P41_1	0.00951	P40_1
0.00898	P13_1_4	0.00892	P38_5
0.00801	P13_5	0.00513	P13_1_5
0.00429	P13_3	0.00414	P38_2
0.00222	P13_2	0.0019	P13_4
0.0019	P13_1_3	0.00165	TRABAJA
0.00165	P40_3	0	P14
0	P22C	0	P1_2

Ganancia	Variable	Ganancia	Variable
0	P23_2C	0	P8_1
0	P8_3	0	P1_3
0	P1_5	0	P16C
0	P8_2	0	P8
0	P15_3C		

La variable con mayor ganancia informacional es la P35_1 que se refiere a *Nombre del oficio o profesión que desempeñó en su último trabajo en EU*, con una ganancia de 0.96124, otra variable con una ganancia alta es la P35_4 con 0.95874 que se trata del *giro del establecimiento, negocio, fábrica o empresa donde trabajó en EU*; estas variables pueden ser consideradas para caracterizar a los grupos de migrantes; sin embargo, existe una gran cantidad de variables que tienen una ganancia informacional menor a 0.1 e incluso algunas con ganancia de 0, lo que indica que se podría prescindir de estas variables en este estudio.

4.2.5. Evaluación e interpretación de los perfiles

Para poder comprender los resultados obtenidos del agrupamiento, se consideró la ganancia informacional y se expresó el conocimiento obtenido de la siguiente manera:

- **Cluster 0:** ***Migrantes reincidentes repatriados.*** Son personas que ya habían cruzado hacia EU, pero que fueron regresados a México por las autoridades de EU porque habían cruzado de manera ilegal. Han cruzado de 2 a 6 veces para buscar trabajo en EU, en la última ocasión, contrataron a un coyote o pollero para que los ayudaran a cruzar por la ciudad de El Sásabe-Altar, además contaban con el apoyo de familiares o amigos que los esperaban en el lugar donde llegaron, permaneciendo la mayor parte del tiempo en el estado de California. Su estancia en EU fue de 2 a 3 años y la ciudad a donde se dirigen actualmente es Los Ángeles, Ca.
  El último grado de escolaridad de estas personas es el nivel Secundaria y nunca han trabajado en su lugar de origen, por lo que toda su vida laboral ha sido desarrollada en EU, donde tuvieron alrededor de 2 trabajos diferentes y trabajaban de 5 a 8 horas diarias, de 5 a 7 días a la semana, desempeñándose en actividades agrícolas.

- **Cluster 1:** ***Migrantes con perfil de estudiante.*** Son personas que el último grado de escuela que aprobaron fue en el nivel Preparatoria y antes de realizar el viaje eran estudiantes, razón por la cual nunca trabajaron en su lugar de origen. Es la primera vez que intentan cruzar la frontera para trabajar o buscar trabajo y no se dirigen a alguna ciudad en especial de EU.

- **Cluster 2:** ***Migrantes sin experiencia migratoria.*** Son personas que nunca han cruzado la frontera para trabajar o buscar trabajo, pero que si han trabajado en su lugar de origen y no se dirigen a alguna ciudad en especial de EU. Antes de realizar el viaje trabajaban en una finca agrícola o en el campo, donde laboraban de 6 a 50 personas, pero no tenían contrato, prestaciones, capacitación. Tenían un sueldo fijo, entre 650 y 1300 pesos a la semana y trabajaban de 7 a 9 horas diarias, de 5 a 7 días a la semana. Aprendieron su oficio a través del desempeño de su trabajo y el último año de escuela que aprobó fue en el nivel Primaria.

- **Cluster 3:** ***Migrantes reincidentes de visita en México.***Son personas que ya habían cruzado hacia EU, pero que regresaron para visitar a familiares y amigos. Han cruzado de 2 a 4 veces para buscar trabajo en EU, en la última ocasión, cruzaron de manera ilegal, por lo que contrataron a un coyote o pollero para que los ayudaran a cruzar por la ciudad de El Sásabe-Altar, además contaban con el apoyo de familiares o amigos que los esperaban en el lugar donde llegaron, permaneciendo la mayor parte del tiempo en el estado de California. Su estancia en EU fue de 6 meses a 3 años, donde tuvieron alrededor de 2 trabajos diferentes y trabajaban de 5 a 10 horas diarias, de 5 a 7 días a la semana, desempeñándose en actividades agrícolas. Se dirigen principalmente a las ciudades de Phoenix y Maricopa, en el estado de Arizona.
  Estas personas si han trabajado en su lugar de origen. Antes de realizar el viaje trabajaban en una finca agrícola o en el campo, donde laboraban de 2 a 5 personas, pero no tenían contrato, prestaciones, capacitación. Su salario era máximo de 650 pesos a la semana y trabajaban de 7 a 9 horas diarias, de 5 a 7 días a la semana. Aprendieron su oficio a través del desempeño de su trabajo y el último año de escuela que aprobó fue en el nivel Secundaria.

4.2.5.1. Clasificación

Una vez que se obtuvo el mejor agrupamiento y se caracterizó, se seleccionaron los atributos que tuvieran una ganancia informacional mayor a 0.1, teniendo así un nuevo conjunto de datos con 39 variables para aplicar algoritmos de clasificación. Estos algoritmos fueron los señalados en la sección 3.2.5.1 y de igual forma, fueron configurados para que se evaluaran por validación cruzada con 10 particiones. Los porcentajes de instancias clasificadas correctamente por estos algoritmos se encuentran ilustradas en la Figura 4.1.

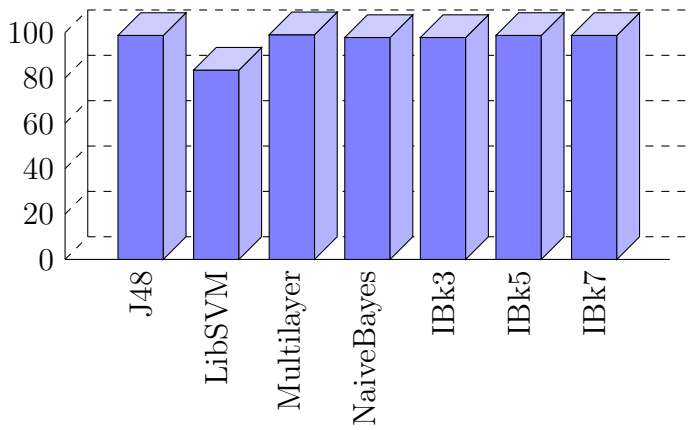


Figura 4.1: Porcentaje de instancias clasificadas correctamente

En la Tabla 4.6 se muestran las métricas que miden la exactitud de los modelos, las cuales fueron obtenidas al aplicar los algoritmos de clasificación, mencionados anteriormente.

Los algoritmos J48 e IBk 5 y 7 obtuvieron los mismos valores en todas las métricas; teniendo valores elevados en precisión, cobertura, F-measure y verdaderos positivos, y un valor bajo en los falsos positivos; sin embargo, el algoritmo de Redes Neuronales denominado *Multilayer perceptron* obtuvo el mejor resultado, en todas las medidas con 0.991 en precisión y 0.990 de F-measure. Por esta razón, se recomienda utilizar este algoritmo para futuras predicciones.

Tabla 4.6: Comparación entre los resultados de los diferentes algoritmos de clasificación

Algoritmo	Precisión	Recall	F-measure	TP Rate	FP Rate
J48	0.988	0.988	0.987	0.988	0.016
LibSVM	0.844	0.835	0.808	0.835	0.212
MultilayerPerceptron	0.991	0.991	0.990	0.991	0.010
NaiveBayes	0.979	0.011	0.979	0.979	0.979
IBk3	0.979	0.979	0.978	0.979	0.021
IBk5	0.988	0.988	0.987	0.988	0.016
IBk7	0.988	0.988	0.987	0.988	0.016

4.2.5.2. Aplicación del algoritmo LR-FIR en el factor social

Como en el capítulo anterior se aplicó exitosamente el algoritmo LR-FIR para obtener reglas lingüísticas que describieran los grupos encontrados, se decidió aplicar también dicho algoritmo. Primero, se utilizó el conjunto de datos completo (327 registros) como entrenamiento (training), con 127 variables; se hicieron cuatro experimentos con máscaras con complejidad 4 y 5; sin embargo las métricas de las reglas no mostraban resultados satisfactorios; por lo tanto se decidió utilizar las 39 variables utilizadas en la sección 4.2.5.1. Posteriormente, se realizaron seis experimentos más, con máscaras con complejidad de 3, 4 y 5; sin embargo, los resultados seguían teniendo métricas bajas.

Finalmente, se realizaron tres experimentos, los cuales se describen más adelante, con máscaras con complejidad de 5 y 6, con el conjunto de 327 registros y 39 variables; se discretizaron algunas variables; por ejemplo: P11 *Ha trabajado en el lugar donde vive o en algún lugar cercano* en dos clases: 1 que significaba *Si* en el rango [0 a 1.5] y el 2 que indicaba *No* en el rango [1.5 a 2.5]; la variable P11_2 *Si trabajó en el lugar donde vive los 30 días anteriores al viaje*, se discretizó en tres clases: -2 que indicaba *No, porque no ha trabajado en el lugar donde vive* en el rango [-2.5 a 0], 1 de *Si* en el rango [0 a 1.5] y 2 de *No* en el rango [1.5 a 2.5]; las variables P34 *Si la última vez que cruzó, trabajó en EU.* y P29 *Si la última vez que cruzó, contrató a alguien para que lo ayudara a cruzar*, se discretizaron en tres clases: -2 que indicaba *Ninguna vez ha cruzado* en el rango [-2.5 a 0], el 1 de *Si* en el rango [0 a 1.5] y 2 de *No* en el rango [1.5 a 2.5].

Para ejecutar el algoritmo se utilizaron los mismos parámetros que en la sección 3.2.5.3.

Para el primer experimento se utilizó una máscara de complejidad 6, con la opción uno de mejora de la compactación, filtrado de reglas y unificación de reglas, paso uno y dos; y se obtuvieron las reglas mostradas en la tabla 4.7.

Tabla 4.7: Experimento 1: Extracción de reglas lingüísticas para el factor social de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje y la última vez que cruzó a EU trabajó como trabajador de actividades agrícolas en un negocio dedicado a la agricultura, ENTONCES Cluster ES ***Migrantes reincidentes repatriados***	1	1
	Métricas para el Cluster *Migrantes reincidentes repatriados*	**1**	**1**
2	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes con perfil de estudiante***	1	0.84
3	SI ha trabajado en el lugar donde vive o en un lugar cercano y no trabajó los 30 días anteriores al viaje y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes con perfil de estudiante***	0.99	0.11
	Métricas para el Cluster *Migrantes con perfil de estudiante*	**0.98**	**1**
4	SI ha trabajado en el lugar donde vive o en un lugar cercano y si trabajó los 30 días anteriores al viaje y ninguna vez ha cruzado, ENTONCES Cluster ES ***Migrantes sin experiencia migratoria***	1	0.98
	Métricas para el Cluster *Migrantes sin experiencia migratoria*	**1**	**0.98**
5	SI ha trabajado en el lugar donde vive o en un lugar cercano y si trabajó los 30 días anteriores al viaje y la última vez que cruzó a EU trabajó como trabajador de actividades agrícolas en un negocio dedicado a la agricultura, ENTONCES Cluster ES ***Migrantes reincidentes de visita en México***	1	1
	Métricas para el Cluster *Migrantes reincidentes de visita en México*	**1**	**1**
	MÉTRICAS PARA EL MODELO	**1**	**0.99**

Para el segundo experimento se utilizó una máscara de complejidad 5, con la opción uno de mejora de la compactación, filtrado de reglas y unificación de reglas, paso uno y dos; y se obtuvieron las reglas mostradas en la tabla 4.8.

En el experimento tres, se utilizó la misma máscara que en el experimento dos, con la opción uno de mejora de la compactación; pero primero se aplicó el primer paso de unificación de reglas, después el filtrado de reglas y posteriormente el paso dos de unificación de reglas; las reglas que se obtuvieron se muestran en la tabla 4.9.

Como se puede observar las reglas obtenidas en el experimento dos y en el experimento tres, son exactamente las mismas, incluso con las mismas métricas de especificidad y sensitividad, sólo cambia el orden en el que aparecen

Tabla 4.8: Experimento 2: Extracción de reglas lingüísticas para el factor social de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje, y la última vez que cruzó a EU trabajó como trabajador de actividades agrícolas, ENTONCES Cluster ES ***Migrantes reincidentes repatriados***	1	0.33
2	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje, y la última vez que cruzó a EU trabajó como trabajador en la elaboración de alimentos, bebidas y productos de tabaco, ENTONCES Cluster ES ***Migrantes reincidentes repatriados***	1	0.11
	Métricas para el Cluster *Migrantes reincidentes repatriados*	**1**	**1**
3	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje, y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes con perfil de estudiante***	1	0.84
4	SI ha trabajado en el lugar donde vive o en un lugar cercano y no trabajó los 30 días anteriores al viaje, y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes con perfil de estudiante***	0.99	0.11
	Métricas para el Cluster *Migrantes con perfil de estudiante*	**0.98**	**1**
5	SI ha trabajado en el lugar donde vive o en un lugar cercano y si trabajó los 30 días anteriores al viaje y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes sin experiencia migratoria***	1	0.98
	Métricas para el Cluster *Migrantes sin experiencia migratoria*	**1**	**0.98**
6	SI ha trabajado en el lugar donde vive o en un lugar cercano y si trabajó los 30 días anteriores al viaje y la última vez que cruzó a EU trabajó como trabajador de actividades agrícolas, ENTONCES Cluster ES ***Migrantes reincidentes de visita en México***	1	1
	Métricas para el Cluster *Migrantes reincidentes de visita en México*	**1**	**1**
	MÉTRICAS PARA EL MODELO	**1**	**0.99**

las reglas para el cluster ***Migrantes con perfil de estudiante***. Con respecto al experimento uno, se diferencía por que hay una sola regla para el cluster ***Migrantes reincidentes repatriados***; sin embargo las reglas para los clusters ***Migrantes con perfil de estudiante***, ***Migrantes sin experiencia migratoria*** y ***Migrantes reincidentes de visita en México*** son las mismas que en los otros dos experimentos.

En el primer experimento, la única regla para el cluster ***Migrantes reincidentes repatriados***, trata de las personas que no han trabajado en el lugar donde viven, pero que ya han cruzado a EU y han trabajado allá en actividades agrícolas; en el caso de los experimento dos y tres, hay dos reglas para el cluster ***Migrantes reincidentes repatriados*** que tratan también de personas que no han trabajado en el lugar donde viven, que ya han cruzado a EU y han trabajado allá, pero la primer regla trata de trabajadores agrícolas y la segunda de trabajadores en la elaboración de alimentos, bebidas y productos del tabaco.

Considerando que las métricas de las reglas para los clusters ***Migrantes***

Tabla 4.9: Experimento 3: Extracción de reglas lingüísticas para el factor social de los migrantes

No. de regla	Reglas de LR-FIR	Esp.	Sen.
1	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje, y la última vez que cruzó a EU trabajó como trabajador de actividades agrícolas, ENTONCES Cluster ES ***Migrantes reincidentes repatriados***	1	0.33
2	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje, y la última vez que cruzó a EU trabajó como trabajador en la elaboración de alimentos, bebidas y productos de tabaco, ENTONCES Cluster ES ***Migrantes reincidentes repatriados***	1	0.11
	Métricas para el Cluster *Migrantes reincidentes repatriados*	**1**	**1**
3	SI ha trabajado en el lugar donde vive o en un lugar cercano y no trabajó los 30 días anteriores al viaje, y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes con perfil de estudiante***	0.99	0.11
4	SI no ha trabajado en el lugar donde vive o en un lugar cercano y por esa razón no trabajó los 30 días anteriores al viaje, y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes con perfil de estudiante***	1	0.84
	Métricas para el Cluster *Migrantes con perfil de estudiante*	**0.98**	**1**
5	SI ha trabajado en el lugar donde vive o en un lugar cercano y si trabajó los 30 días anteriores al viaje y ninguna vez ha cruzado a EU, ENTONCES Cluster ES ***Migrantes sin experiencia migratoria***	1	0.98
	Métricas para el Cluster *Migrantes sin experiencia migratoria*	**1**	**0.98**
6	SI ha trabajado en el lugar donde vive o en un lugar cercano y si trabajó los 30 días anteriores al viaje y la última vez que cruzó a EU trabajó como trabajador de actividades agrícolas, ENTONCES Cluster ES ***Migrantes reincidentes de visita en México***	1	1
	Métricas para el Cluster *Migrantes reincidentes de visita en México*	**1**	**1**
	MÉTRICAS PARA EL MODELO	**1**	**0.99**

con perfil de estudiante, ***Migrantes sin experiencia migratoria*** y ***Migrantes reincidentes de visita en México***, son iguales en los tres experimentos, se analizaron las métricas para el cluster ***Migrantes reincidentes repatriados***, tanto de las reglas individualmente como las del cluster y se puede decir que los resultados del experimento uno son los más adecuados, ya que como métricas de la regla tiene especificidad de 1 y sensitividad de 1, y de métricas del cluster también.

Tomando en cuenta lo anterior, las reglas 1 y 5, del experimento uno, tratan de personas que ya han cruzado a EU y trabajado allá en actividades agrícolas, pero se diferencía en que el cluster ***Migrantes reincidentes repatriados*** agrupa a las personas que no han trabajado en el lugar donde vive y que por ello no trabajaron los 30 días anteriores al viaje (regla 1); y el cluster ***Migrantes reincidentes de visita en México*** contiene a las personas que si han trabajado en el lugar donde viven y si trabajaron los 30 días anteriores al viaje (regla 5).

Las reglas 2, 3 y 4, tratan de personas que ninguna vez han cruzado a EU;

pero en el cluster ***Migrantes con perfil de estudiante*** están las personas que no han trabajado en el lugar donde viven y por eso no trabajaron los 30 días anteriores al viaje (regla 2) y las personas que si han trabajado en el lugar donde viven, pero que no trabajaron los 30 días anteriores al viaje (regla 3); y en el caso del cluster ***Migrantes sin experiencia migratoria***, se encuentran las personas que si han trabajado en el lugar donde viven y que también trabajaron los 30 días anteriores al viaje.

El conjunto de reglas obtenido con el experimento uno coincide con la descripción de los grupos que se habia hecho de manera manual en la sección 4.2.5, donde se analizó el comportamiento de las variables con mayor ganancia informacional para hacer dicha descripción.

Conclusiones

La Minería de Datos se ha aplicado exitosamente en muchos campos, en donde ha sido necesario procesar grandes volúmenes de datos para resolver problemas de descripción y predicción. Una de las áreas donde la minería de datos puede contribuir para obtener nuevo conocimiento que permita la toma de mejores decisiones, es en la solución de problemas sociales. En México, la migración hacia los Estados Unidos de América es uno de los principales problemas sociales, ya que los migrantes tienen que dejar a sus familias y sus viviendas. En esta investigación se presenta un estudio que analiza los factores demográfico y social de los migrantes; para ello se analizaron los datos obtenidos del Censo de Población y Vivienda 2010 de INEGI y la Encuesta sobre Migración en la Frontera Norte de México (EMIF), aplicada en el año 2012.

Para analizar el factor demográfico se utilizó el conjunto de datos de INEGI, ya que este contenía información acerca de las viviendas donde tenían al menos un miembro migrante en el periodo del año 2005 al año 2010. El proceso que llevado a cabo fue el siguiente:

- Se aplicaron diferentes algoritmos de clustering y se definió como mejor agrupación a los resultados obtenidos por el algoritmo de Make Density Based Clusterer con tres grupos.

- Estos grupos fueron etiquetados con ayuda de expertos en migración, de la siguiente manera:

    - *Viviendas en exclusión y pobreza*
    - *Viviendas con servicios medios y desarrollo*
    - *Viviendas en Marginación*

- Se encontraron las variables más representativas, con base en la ganancia informacional, se analizó el comportamiento de estas variables en cada grupo y se realizó una descripción manual de ese comportamiento.
- Posteriormente, se aplicaron diversos algoritmos de Clasificación, teniendo buenos resultados los algoritmos Perceptron Multicapa y Naive Bayes; sin embargo, las mejores métricas fueron las obtenidas por el algoritmo Naive Bayes, por lo cual se recomienda utilizar este algoritmo para cuando se quieran predecir, desde el punto de vista del factor demográfico, los casos de migración en el estado.
- Para verificar que los resultados obtenidos en la caracterización manual de los grupos encontrados sean objetivos, se aplicó el algoritmo LR-FIR, en diferentes experimentos, encontrando un conjunto de reglas lingüísticas que coinciden con la descripción anterior.

En el caso del factor social, se analizaron los datos de la EMIF del 2012, con los datos de los migrantes provenientes del sur con destino a EU (personas que van del interior de la República Mexicana a la frontera norte, con la intención de cruzar a EU). Los pasos realizados fueron los siguientes:

- Después de aplicar varios algoritmos de clustering, se estableció como mejor partición la obtenida por el algoritmo Simple K-means con 4 grupos.
- Estos grupos fueron evaluados y etiquetados en colaboración con los expertos sociales de la siguiente manera:
  - *Migrantes reincidentes repatriados*
  - *Migrantes con perfil de estudiante*
  - *Migrantes sin experiencia migratoria*
  - *Migrantes reincidentes de visita en México*
- Se calculó la ganancia informacional de las variables para obtener aquellas que fueran más representativas y que diferenciaran a los grupos, se analizó su comportamiento y se describió a cada uno de los grupos.
- Se aplicaron los mismos algoritmos de clasificación utilizados en el factor demográfico y se recomienda utilizar el algoritmo de Multilayer Perceptron, para cuando se quieran predecir a los habitantes del estado que

puedan llegar a convertirse en migrantes, considerando su factor social; ya que tuvo el mejor resultado (con respecto al de los demas algoritmos aplicados), muy cercano a los valores ideales.

- Se obtuvo un modelo de estos grupos, que consiste en un conjunto de reglas lingüísticas, utilizando el algoritmo LR-FIR; cabe mencionar que estas reglas concuerdan con la descripción que se había hecho de los grupos, considerando el comportamiento de las variables más representativas.

Los algoritmos de Weka de clustering y de clasificación, fueron eficientes para el análisis de los factores demográfico y social; sin embargo, el algoritmo de LR-FIR, fue el más adecuado para obtener las reglas que definieran a los grupos generados, ya que las reglas obtenidas por Weka fueron numerosas y confusas al momento de querer utilizarlas para describir los grupos.

La Minería de Datos se puede aplicar exitósamente en el análisis de problemas sociales, tal como se muestra en este trabajo. Los resultados presentados en esta investigación pueden ser usados por el gobierno del estado de Hidalgo para la creación de programas sociales dirigidos a proporcionar asistencia a las familias de los migrantes que se quedaron en el estado o a prevenir la migración en la población. Además los resultados obtenidos son muy valorados por los expertos del fenómeno social, debido a que el conocimiento extraído podría realizarse mediante un análisis manual; sin embargo, el proceso para lograrlo crecería de manera exponencial y por lo tanto el tiempo empleado para alcanzar resultados similares estaría fuera de control.

Trabajos Futuros

Como trabajos futuros se propone aplicar la herramienta CR-FIR, incluida en la herramienta Visual-FIR, para la selección de variables a los dos conjuntos de datos utilizados, con el fin de considerar no sólo la ganancia informacional, sino también la relevancia relativa de las variables.

Una vez que se cuente con los conjuntos de datos con las variables más relevantes, se seguirá todo el procedimiento presentado en este trabajo y se compararán los resultados con los obtenidos en esta tesis.

Se plantea buscar otros conjuntos de datos que permitan analizar la migración en el estado.

Contribución

El aporte de este trabajo es el descubrimiento de conocimiento en un problema de índole social a través de la aplicación eficiente de diversas técnicas de Minería de Datos, tales como: agrupamiento, validación de particiones resultantes y caracterización de las mismas, clasificación y obtención de reglas comprensibles para la interpretación de este conocimiento.

Producto Obtenido

Franco, A.; Franco, K.D.; Castro F.A. y García, L.H.: "Data Mining for Discovering Patterns in Migration". En: *Nature-Inspired Computation and Machine Learning*. Springer International Publishing. pp. 285 – 295 (2014)

SEP
SECRETARIA DE EDUCACION PUBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de Tuxtla Gutiérrez

The Mexican Society for Artificial Intelligence (SMIA) and the Instituto Tecnológico de Tuxtla Gutiérrez (ITTG)

award this certificate to

Anilu Franco-Arcega, Kristell D. Franco Sanchez, Felix A. Castro Espinoza and Luis H. García Islas

for presentation of the paper entitled

Data mining for discovering patterns in Migration

at the 13TH Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Chiapas, México, November 17-21, 2014

M. E. H. José Luis Méndez Navarro
Director del ITTG

Dr. Alexander Gelbukh
Presidente de SMIA

M. E. P. Jaime Valls Esponda
Rector de la UNACH

SECRETARIA DE EDUCACION PUBLICA
INSTITUTO TECNOLOGICO de Tuxtla Gutiérrez
DIRECCION

RSGC 596
Fecha de Inicio: 2009.09.22
Fecha de Recertificación: 2012.07.27
Fecha de Terminación: 2015.07.27

Bibliografía

[1] Aghabozorgi, S.; Mahroeian, H.; Dutt, A.; Ying Wah, T. y Herawan. T. "An Approachable Analytical Study on Big Educational Data Mining". En: *Computational Science and Its Applications–ICCSA 2014*. Springer International Publishing. pp. 721–737 (2014).

[2] Aguilar, A. M. y Mateos, P.: "Diferenciación sociodemográfica del espacio urbano de la Ciudad de México". *Revista Eure*, vol. 37, no. 110, pp. 5–30 (2011).

[3] Aguilar S.; Benítez J.L. y Tafolla R.:*Problemas sociales, económicos y políticos de México*. Editorial Universidad Nacional Autónoma de México (2006).

[4] Alba F.: *Migración Internacional, Consolidación de los patrones emergentes*. México: UNAM-COLMEX-FPNU-INEGI (2000).

[5] Anderson J.A.: *Redes Neuronales* Editorial Alfaomega (2007).

[6] Anguiano, M.: "Rumbo al norte: nuevos destinos de la emigración veracruzana". *Migraciones Internacionales*, vol. 3, no.1, pp. 82 – 110 (2005).

[7] Ansari, A. y Ghalamkari, S.: "Segmenting Online Customers Based on their Lifetime Value and RFM Model by Data Mining Techniques". En: *International Journal of Information Science and Management*, Special Issue, pp. 69–82 (2014).

[8] Banks, D.; House, L.; McMorris, F. R.; Arabie, P. y Gaul, W.: *Classification, Clustering, and Data Mining Applications. Procedings of the Meeting of the International Federation of Classification Societies (IFCS)*. Institute of Technology, Chicago. Springer. Illinois (2004).

[9] Baker, R.S.: "Educational Data Mining: An Advance for Intelligent Systems in Education". *IEEE Intelligent Systems*, pp. 78–82 (2014).

[10] Bosch E. y Ferrer V.: "La violencia de género: De cuestión privada a problema social". *Psychosocial Intervention*, vol. 9, no. 2, pp.7–19 (2000).

[11] Boris, M.: *Clustering for Data Mining: A data recovery approach*. Chapman & Hall (2005).

[12] Canales, A. I.; Montiel, I.:"Remesas e inversión productiva en comunidades de alta migración a Estados Unidos. El caso de Teocaltiche, Jalisco". *Migraciones Internacionales*, vol. 2, no. 003, Enero-Julio, pp. 142–172 (2004).

[13] Castro, F; Nebot, A. y Mugica, F.:"On the extraction of decision support rules from fuzzy predictive models". *Applied Soft Computing*, vol. 11, no. 4, pp. 3463–3475 (2011).

[14] Chakravarthy, S. y Raman, A.: "Educational Data Mining on Learning Management Systems using Experience API". En: *Fourth International Conference on Communication Systems and Network Technologies*, pp. 424–427 (2014).

[15] Chávez, D.; Miranda, I.; Varela, M. y Fernández, L.: "Utilización del análisis de cluster con variables mixtas en la selección de genotipos de maíz (Zea mays)". *Revista Investigación Operacional*, vol. 30, no. 3, pp. 209–216 (2010).

[16] Colmenares, G.; Amaru-yawa, R.: "Minería de Datos aplicada a los cambios en la estructura de la variable desempleo. Caso de estudio: El estado Mérida". *Universidad de Los Andes*. Mérida (2007).

[17] CONAPO: Índice de intensidad migratoria México-Estados Unidos (2000).

[18] Correa Londoño, G.; Lavalett Oñate, L.L.; Galindo Villardón, M.P. y Afanador Kafuri, L.: "Uso de métodos multivariantes para la agrupación de aislamientos de Colletotrichum spp. con base en características morfologicas y culturales". *Revista Facultad Nacional de Agronomía - Medellín*, vol. 30, no. 1, pp. 3671–3690 (2007).

[19] Corso, C. L.: "Aplicación de algoritmos de clasificación supervisada usando Weka". En: *Congreso Información y Comunicación para la Sociedad del Conocimiento*. Universidad Tecnológica Nacional, Facultad Regional Córdoba. Argentina (2009).

[20] Díaz, I. B.; Montoya, D.M.: "Una medida de similitud basada en las modas para la caracterización de una población estudiantil en edad extraescolar". *Revista de Ingenierías Universidad de Medellín*, vol. 4, no. 007, pp. 101-109 (2005).

[21] Dunham M.H.: *Data Mining. Introductory and Advanced Topics*. New Jersey: Prentice Hall - Pearson Education (2003).

[22] Duda, R.O.; Hart, P.E. y Stork D.G.: *Pattern Classification*. Wiley (2001).

[23] Ellison S. L.; Patton D. R.; Simard L. y McConnachie A. W.: "Galaxy Pairs in the Sloan Digital Sky Survey. I. Star Formation, Active Galactic Nucleus Fraction, and the Luminosity/Mass-Metallicity Relation". *The Astronomical Journal*, vol. 135, no. 5, pp.1877 (2008).

[24] Fayyad U., Piatetsky G., Smyth P. y Uthurusamy R.: *Advances in Knowledge Discovery and Data Mining*. MIT Press (1996).

[25] Franco L.M.: *Migración y Remesas en la ciudad de Ixmiquilpan*. Universidad Autónoma del Estado de Hidalgo (2012).

[26] Garrocho C.: *Distribución espacial de la población ZMCM 1950-1990 en Estudios Demográficos y urbanos*. El Colegio de México (1995).

[27] Geach, J. E.: "Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys". *Monthly Notices of the Royal Astronomical Society*, vol. 419, no. 3, pp. 2633–2645 (2012).

[28] Gregory D. y Urry J.: *Social Relations and Spatial Structures*. Basingstoke: Macmillan (1985).

[29] Godara, S.; Yadav, R: "Performance analysis of clustering algorithms for character." *International Journal of Advanced Computer and Mathematical Sciences*, vol.4, no. 1, pp. 119–123 (2013)

[30] Gomez Chova, L.: "Pattern Recognition Methods for Crop Classification from Hyperspectral Remote Sensing Images". *Dissertation.com*, Florida, USA (2004).

[31] González, J. G.: "Migración y remesas en el sur del Estado de México". *Papeles de Población*. pp. 223–252 (2006).

[32] Guillén, T.: "Entre la convergencia y la exclusión. La deportación de mexicanos desde Estados Unidos de América". *Realidad, Datos y Espacio: Revista Internacional de Estadística y Geografía*. pp. 164–179 (2012).

[33] Gutiérrez, P.; Merlino, H.; Rancan, C.; Procopio, C.; Rodriguez, D.; Britos, P.V. y Garcia, R.: "Identificación de patrones característicos de la población carcelaria mediante minería de datos". En: *Proceedings X Workshop de Investigadores en Ciencias de la Computación*, pp. 461–465 (2008).

[34] Han, J.; Kamber, M. y Pei, J.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers (2012).

[35] Hernández, J.; Ramírez, M. y Ferri, C.: *Introducción a la Minería de Datos*. Prentice Hall (2005).

[36] INEGI. Instituto Nacional de Estadística y Geografía. Recuperado de: http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2010/presentacion.aspx el 2 de Julio de 2014.

[37] Jain, A.K.; Dubes, R.C.: *Algorithms for Clustering Data*. Prentice Hall. New Jersey, USA (1988).

[38] Jáuregui, J.A.; Ávila, M.J.: "Estados Unidos, lugar destino para los migrantes chiapanecos". *Migraciones Internacionales*. pp. 5–38 (2007).

[39] Kamani, I. A.; Samarasinghe, C.; Kodituwakku, S. y Yapa, R.D.: "Understanding the Internet Usage Habits of the Students of University of the Visual and Performing Arts through Data Mining". *International Journal of Soft Computing and Engineering*, vol. 4, no. 2, pp. 45–48 (2014).

[40] Konstantinos, S.T.:*Pattern Recognition*. American Press. USA (2003).

[41] Lagos, C. V.: *Creación de perfiles de deudores de crédito universitario, para mejoramiento de campañas de cobranza, usando Minería de Datos*. Universidad Austral de Chile. (2011).

[42] Lewis J.: *Strategies for Survival: Migration and Fair Trade-Organic Coffee Production in Oaxaca, Mexico*. University of California. San Diego. USA (2005).

[43] López S.: "De lo global a lo local: cambios de cultivos y estrategias de sobrevivencia ante la crisis del mercado internacional del café. El caso de la Sierra Otomí-Tepehua en el estado de Hidalgo". *Problemas del desarrollo*, vol. 33, no. 131,X-XII, pp. 131-162 (2002).

[44] Marques de Sá, J.P.: *Pattern Recognition, Concepts, Methods and Applications*. Springer editors. Portugal (2001).

[45] Moctezuma L, M.: "Inversión social y productividad de los migrantes mexicanos en Estados Unidos". *Migración y Desarrollo*. pp. 85–126 (2003).

[46] Moine, J. M.; Haedo, A. S.; Gordillo, S.: "Estudio comparativo de metodologías para minería de datos". En *XIII Workshop de Investigadores en Ciencias de la Computación*. pp. 1–4 (2011).

[47] Olaru, C.:"Business Intelligence in Telecommunications Industry". *International Journal of Economic Practices and Theories*, vol. 4, no. 1, pp. 89–100 (2014).

[48] Olatz, A; Gurrutxaga, I; Muguerza, J.; Pérez, J.M.; Perona, I.: "An extensive comparative study of cluster validity indices". *Pattern Recognition*, vol. 46, pp. 243–256 (2013).

[49] Papail, J.: "De asalariado a empresario: la reinserción laboral de los migrantes internacionales en la región centrooccidente de México". *Migraciones Internacionales*, vol.1, no.3, Julio-Diciembre, pp. 79–102 (2002).

[50] Pascual, D.; Pla, F.; Sánchez, S.: *Algoritmos de agrupamiento*. Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I. (2007)

[51] Patiño J.C.: "Impacto de las políticas migratorias en las familias Mazahuas". *Convergencia 9*, pp. 217–252 (2002).

[52] Pejic-Bach, M.; Jukovic, S.; Dumicic, K. y Sarlija, N.: "Business client segmentation in banking using self-organizing maps". *South East European Journal of Economics and Business*, pp. 1–10 (2013).

[53] Pollo Cattaneo, F.; Amatriain, H.; Rodriguez, D.; Pytel, P.; Ciccolella, E.; Vegega, C.; Dearriba, M.; Rodiguez Aubert, M.; Bose, F.; Giordano, L.; Britos, P.; García, R.: "Ingeniería de Proyectos de Explotación de Información", En *XII Workshop de Investigadores en Ciencias de la Computación*, pp. 172–176 (2010).

[54] Poonguzhali, S. y Sheshasaayee, A. "A review on data mining techniques for Digital Mammographic Analysis". *International Journal of Data Mining Techniques and Applications*, vol. 3, pp. 374–381 (2014).

[55] Quinlan, J.: "C4.5 Programs for Machine Learning". Morgan Kaufmann Publishers (1993).

[56] Quezada M.F.: "Apuntes sobre la Migración Indígena: Cruzando la frontera norte hacia los Estados Unidos". *Aquí estamos, revista de ex becarios indígenas del IFP- México*, vol. 2, no. 3, pp. 7–13 (2005).

[57] Quezada M.F. y Franco L.M.:*Distribución geográfica de la migración internacional y las remesas en el Estado de Hidalgo*. Universidad Autónoma del Estado de Hidalgo (2012).

[58] Ramírez, B. y González,A.: "La Migración como respuesta de los campesinos ante la crisis del café: Estudio en tres municipios del Estado de Puebla". *Revista de Sociedad, Cultura y Desarrollo Sustentable*, vol. 2, no. 2, pp.319–341 (2006).

[59] Ren, D.; Zheng, D.; Huang, G.; Zhang, S. y Wei, Z.: "Parallel Set Determination and K-means Clustering for Data Mining on Telecommunication Networks". En: *IEEE International Conference on High Performance Computing and Communications and International Conference on Embedded and Ubiquitous Computing*, pp. 1553–1557 (2013).

[60] Revathi, S.; Nalini, T.: "Performance Comparison of Various Clustering Algorithm". *International Journal of Advanced Research in Computer Science and Software Engineering*. vol. 3, no. 2, Febrero, pp. 62–72 (2013).

[61] Sabido, M.; Maldonado, R. M.: "Modelo Clasificador para Predecir el Desempeño Escolar Terminal de un Estudiante". En: *International Institute of Informatics and Systemics. Décima Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI 2011*.Orlando, Florida. USA(2011).

[62] Programas Sociales, *Secretaría de Desarrollo Social*, Marzo 2014. Disponible en: http://www.sedesol.gob.mx/en/SEDESOL/Programas_Sociales

[63] (SEGOB), Secretaria de Gobernación. *Encuesta sobre Migración en la Frontera Norte de México, 2011*. Secretaría de Gobernación/Consejo Nacional de Población/Instituto Nacional de Migración/Unidad de Política Migratoria-Centro de Estudios Migratorios/Secretaría de Relaciones Exteriores/Secretaría del Trabajo y Previsión Social/El Colegio de la Frontera N. (2013).

[64] Serrano T.: *Y se fue. . . Los municipios hidalguenses de muy alto grado de intensidad migratoria internacional*. Universidad Autónoma del Estado de Hidalgo (2006).

[65] Serrano T.: *Migración Internacional y Pobreza en el Estado de Hidalgo*. Universidad Autónoma del Estado de Hidalgo (2006).

[66] Serrano T. y Quezada M.F.: *Indocumentado, sabe a mentira tu verdad. Los municipios hidalguenses de alto grado de intensidad migratoria internacional*. Universidad Autónoma del Estado de Hidalgo. Amalgama Arte Editorial (2007).

[67] Shaikh, T.A.: "Predictive Data Mining for Medical Diagnosis: An Overview of Segment and Arrhythmia Diseases". *International Journal of Emerging Trends in Engineering and Development*, vol. 4, no. 3, pp. 625–634 (2014).

[68] Smita y Sharma, P.: "Use of data mining in various field: a survey paper". *Journal of computer engineering*, vol. 16, no. 3, pp. 18–21 (2014).

[69] Thangarasu, G. y Dominic, P.D.: "Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques". En: *International Conference on Computer and Information Sciences*, pp. 1–5 (2014).

[70] Valcarcel, V.: "Data mining y el descubrimiento del conocimiento". *Ind. Data*, vol. 7, no. 2, Julio - Diciembre, pp. 83–86 (2004).

[71] Webb, A. R.; Copsey, K. D.: *Statistical Pattern Recognition*. Wiley (2011)

[72] Welti C.: *Demografía I*. México D.F.: PROLAP (1997).

[73] Witten, I.H. y Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers (2005).

[74] Yu-ting, G. y Kai, S.: "The study on the Bank Customer Model Based on the Improved Data Mining". *Advances in information Sciences and Service Sciences*, vol. 5, no. 7, pp. 955–962 (2013).

[75] Zytkow, W. K.: *Handbook of Data Mining and Knowledge Discovery*. Oxford University Express. USA. (2002).

[76] Zuñiga, E.; Arroyo, J.; Escobar, A. y Verduzco, G.: *Migración México - Estados Unidos, Implicaciones y retos para ambos países*. Consejo Nacional de Población (2006).