INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
MAESTRÍA EN SISTEMAS COMPUTACIONALES

ANÁLISIS DE PATRONES EN LA PARTICIPACIÓN CIUDADANA EN PROCESOS ELECTORALES APLICANDO ALGORITMOS DE MINERÍA DE DATOS

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
ISC. JOSÉ LUIS GUTIÉRREZ VILLANUEVA

DIRECTOR DE TESIS:
MATI. LUIS ARMANDO CÁRDENAS FLORIDO

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, SEPTIEMBRE 2015.

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de La Paz

"2015, Año del Generalísimo José María Morelos y Pavón"

La Paz, B.C.S., **15/septiembre/2015**

DEPI/521/2015

Asunto: Autorización de impresión.

C.JOSÉ LUÍS GUTIÉRREZ VILLANUEVA,
ESTUDIANTE DE LA MAESTRÍA EN
SISTEMAS COMPUTACIONALES,
P R E S E N T E.

Con base en el dictamen de aprobación emitido por el Comité Tutorial de la Tesis denominada: **"ANÁLISIS DE PATRONES EN LA PARTICIPACIÓN CIUDADANA EN PROCESOS ELECTORALES APLICANDO ALGORITMOS DE MINERÍA DE DATOS"** mediante la opción de tesis (Proyectos de Investigación), entregado por usted para su análisis, le informamos que se **AUTORIZA** la impresión.

A T E N T A M E N T E
"Ciencia es Verdad, Técnica es Libertad"

M.A.T.I. LUÍS ARMANDO CÁRDENAS FLORIDO,
JEFE DE LA DIV. DE EST. DE POSGRADO E INV.

ESTADOS UNIDOS MEXICANOS
INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO
E INVESTIGACIÓN

C.c.p. Coordinación de la Maestría.
C.c.p. Depto. de Servicios Escolares.

LACF/fkso*

DICTAMEN DEL COMITÉ TUTORIAL

SUBDIRECCIÓN ACADÉMICA
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN.

La Paz, B.C.S., **15/SEPTIEMBRE/2015**

C. M.A.T.I. LUÍS ARMANDO CÁRDENAS FLORIDO,
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN,
P R E S E N T E.

Por medio del presente, enviamos a usted dictamen del Comité Tutorial de tesis para la obtención del grado de Maestro, con los siguientes datos generales:

No. de Control M13310017	Nombre JOSÉ LUÍS GUTIÉRREZ VILLANUEVA
Maestría en:	SISTEMAS COMPUTACIONALES
Título de la tesis: ANÁLISIS DE PATRONES EN LA PARTICIPACIÓN CIUDADANA EN PROCESOS ELECTORALES APLICANDO ALGORITMOS DE MINERÍA DE DATOS	
DICTAMEN: Se autoriza el trabajo de investigación, en virtud de que realizó las correcciones correspondientes conforme a las observaciones planteadas por este Comité Tutorial.	

A t e n t a m e n t e.

El Comité Tutorial

MSC. ILIANA CASTRO LIERA

MC. JESÚS ANTONIO CASTRO

MATI. LUIS ARMANDO CÁRDENAS FLORIDO

c.c.p. Coordinador de la Maestría.
c.c.p. Departamento de Servicios Escolares.
c.c.p. Estudiante.

Dedicatoria.

Dedico esta tesis principalmente a mis padres, quienes me dieron la vida, la educación y los valores necesarios para alcanzar mis metas. Dedico este trabajo especialmente a mi madre que me enseñó a tener siempre la fortaleza para salir adelante, para nunca rendirme y luchar por mis sueños. A mi padre, que me enseñó a hacer las cosas siempre con legalidad y honestidad y no desistir de mis sueños.

A mi hermana Alejandra, por su gran apoyo incondicional durante la realización de este documento y que siempre confió en mi para terminarlo exitosamente.

A mi novia Alicia, por su gran apoyo moral y emocional, por su gran paciencia en momentos de estrés y por la motivación en los días difíciles para seguir adelante.

A mi familia entera, quienes siempre dieron muestras de apoyo y me alentaron para poder concluir la tesis.

A mis amigos de toda la vida: Jaime Ángel y Andrés Alejandro, quienes siempre estuvieron al pendiente y brindando su apoyo emocional para terminar el documento de tesis.

Dedico esta tesis a todas y cada una de las personas que durante la redacción de la tesis fueron parte fundamental aportando cosas favorables para ello.

Agradecimientos.

Agradezco a Dios por permitirme terminar una meta más en mi vida profesional. A mis padres que siempre confiaron en mí y siempre me apoyan incondicionalmente en todas las metas que me propongo. A mi hermana Alejandra, a mi novia Alicia, a mis amigos Jaime Ángel y Andrés Alejandro y a toda mi familia por su gran apoyo motivacional para llevar a cabo esta meta de forma exitosa.

Agradezco al Instituto Estatal Electoral de Baja California Sur por permitir llevar a cabo la investigación, por los apoyos económicos y permisos necesarios para el correcto desempeño de la tesis. Agradezco especialmente al Maestro Jesús Alberto Muñetón Galaviz, Consejero Electoral. A la Licenciada Rebeca Barrera Amador, Consejera Presidente. Al Maestro Cesar Adonai Taylor Maldonado, Director de la Unidad de Computo y Servicios Informáticos (UCSI). Al Ingeniero Mario Yee Castro, Coordinador de la UCSI.

Agradezco al Instituto Tecnológico de La Paz por permitirme realizar mis estudios de Maestría en Sistemas Computacionales, en especial, a mis maestros que compartieron su conocimiento durante el desarrollo del posgrado. A mi director de tesis M.A.T.I. Luis Armando Cárdenas Florido quien fue parte fundamental en la redacción de ella, al comité tutorial quien evaluó y autorizó la presente tesis del cual formaron parte la M.S.C. Iliana Castro Liera y el M.C. Jesús Antonio Castro.

Agradezco al Consejo Nacional de Ciencia y Tecnología por su apoyo económico con el otorgamiento de una beca para el estudio del posgrado.

Resumen.

En Baja California Sur se aplica un sistema democrático para el proceso de elección a gobernador constitucional. Las elecciones se llevan a cabo con la participación de los ciudadanos y del Instituto Estatal Electoral de Baja California Sur (IEEBCS). Esto genera información muy importante: los resultados electorales y la participación ciudadana.

Los resultados electorales son el conteo final de todos los votos agrupados por candidato. La participación ciudadana es el porcentaje del total de personas que votaron entre el total de personas que podían votar en la elección.

Una de las métricas para la evaluación del proceso electoral es la participación ciudadana. Por lo tanto, el IEEBCS dentro de sus planes estratégicos para asegurar unas elecciones exitosas ejecuta una promoción del voto antes y durante el proceso electoral para motivar a los ciudadanos a votar el día de la elección. Sin embargo esta promoción del voto se realiza de forma muy general.

En el presente trabajo se diseñó un modelo de minería de datos para el análisis de la participación ciudadana en procesos electorales de Baja California Sur. El cual permite determinar las variables y patrones de comportamiento presentados en las elecciones. Se realizó la construcción de una bodega de datos alimentada por datos de la participación ciudadana y de fuentes externas con información demográfica de la entidad. Para el análisis de la información se aplicaron técnicas de Minería de Datos para la obtención de los patrones de comportamiento. Las técnicas utilizadas son: el método de Árbol de Decisión (dentro del algoritmo de Clasificación) y el método de K-Means (dentro de los algoritmos de Clúster o Agrupamiento).La investigación se llevó a cabo mediante la aplicación de la metodología CRISP-DM, ya que es una de las metodologías de minería de datos más completa y de las más utilizadas.

El análisis de la participación ciudadana permite llevar a cabo una planificación estratégica y focalizada a las diferentes regiones demográficas del estado, además de lograr una promoción más eficiente y un mayor porcentaje de participación.

Abstract.

For the purposes of electing the State Governor, the Mexico's state Baja California Sur applies a democratic system. Regular citizens, with the support of the State electoral institute (the spanish name is Instituto Estatal Electoral de Baja California Sur, with the acronym IEEBCS), organizes the elections. This creates a very important group of information: the election results and the citizen participation rate.

The election results consist in the final count of all votes, grouped by the candidates involved in the election. The citizen participation rate is the percentage calculated by comparing the total count of citizens who voted with the total of citizens who could legally vote.

One of the usual metric for evaluating the success of an election process, is the citizen participation rate. Therefore, the IEEBCS, in his strategic plans, always assures to have a very strong promotion campaign, before and during the electoral process, to motivate the citizens and encourage them to vote. However, this vote promotion is usually conducted in a very general way across the State.

In this present work, with the objective of analyzing the historic citizen participation rate in recent elections, a data mining model was designed for this specific purpose. This particular model, allows to determine all the variables and behavior patterns occurs during an election process. Also, a data warehouse was built with two main data inputs, one consists in the historic citizen participation rate from recent elections and the other one, a database with diverse demographic data from Baja California Sur. For data analysis, there were used various data mining techniques, such as: Decision Tree (within the Classification algorithm) and K-Means (in specific, the Cluster or Grouping algorithms). Also, CRISP-DM methodology was used for this investigation, mainly because is one of the most complete and used data mining methodology.

The analysis of citizen participation rate allows to plan better strategies, and most importantly, in a more focalized way across the many regions of Baja California Sur, and therefore, achieving a wider vote promotion and consequently, and increased citizen participation rate.

1. Introducción 5
1.1. Contexto 5
1.2. Antecedentes 6
1.3. Descripción del problema 7
1.4. Objetivo general 8
1.5. Objetivos específicos 8
1.6. Alcances y limitaciones 8
1.7. Justificación 9
1.8. Hipótesis 10
2. Marco teórico 11
2.1. Base de datos 11
2.2. Sistema Manejador de Base de Datos (DBMS) 11
2.3. Administrador de Base de Datos (DBA) 11
2.4. Bodega de datos (DW) 11
2.5. Modelos de bases de datos multidimensionales 13
2.6. Hipercubo 14
2.7. Hecho 14
2.8. Dimensiones 14
2.9. Mercados de datos (DATA MARTS) 15
2.10. Minería de datos 15
2.11. Tipos de minería de datos 20
2.12. Funciones de la minería de datos 21
2.13. Herramientas de Minería de Datos 22
3. Metodología de la investigación 24
3.1. Metodologías de minería de datos 24

3.2. Metodología KDD 24
3.3. Metodología CRISP-DM 26
3.4. Selección de la metodología 29
4. Metodología 30
4.1. Comprensión del negocio 30
4.1.1. Determinar los objetivos del negocio 30
4.1.1.1. Contexto 30
4.1.1.2. Objetivos del negocio 31
4.1.1.3. Criterios de éxito 31
4.1.2. Evaluar la situación 32
4.1.2.1. Inventario de recursos 32
4.1.2.2. Terminología 33
4.1.2.3. Factibilidad 33
4.1.3. Determinar los objetivos de minería de datos 34
4.1.3.1. Objetivo de minería de datos 34
4.1.3.2. Criterios de éxito de minería de datos 34
4.1.4. Producir el plan de proyecto 35
4.1.4.1. Evaluación inicial de funciones y algoritmos 35
4.2. Comprensión de los datos 36
4.2.1. Recolectar datos iniciales 36
4.2.1.1. Reporte de recolección de datos iniciales 36
4.2.2. Descripción de los datos 40
4.2.2.1. Reporte de descripción de los datos 40
4.2.3. Exploración de los datos 44
4.2.3.1. Reporte del modelo descriptivo 44

4.2.4. Verificar la calidad de los datos 47
4.3. Preparación de los datos 48
4.3.1. Seleccionar y verificar los datos 48
4.3.1.1. Razones para la inclusión o exclusión de los datos 48
4.3.2. Limpiar los datos 51
4.3.3. Integración de los datos 51
4.3.4. Dar formato a los datos 53
4.4. Modelado 60
4.4.1. Selección de la técnica de modelado 60
4.4.1.1. Técnica de modelado 61
4.4.2. Generar el diseño de pruebas 62
4.4.3. Construcción del modelo 62
4.4.3.1. Algoritmo de clasificación 63
Prueba #1 63
Resultados Prueba #1 65
Prueba #2 66
Resultados Prueba #2 68
Prueba #3 69
Resultados Prueba #3 71
4.4.3.2. Algoritmo de clúster 73
Prueba #1 73
Resultados Prueba #1 75
Prueba #2 76
Resultados Prueba #2 78
Prueba #3 79

Resultados Prueba #3 81
4.5. Interpretación de los Resultados 83
4.5.1. Comprensión e interpretación de los resultados obtenidos. 83
4.5.1.1. Variables o características de mayor relevancia en los votantes. 83
4.5.1.2. Patrón de comportamiento en los votantes 85
4.5.2. Evaluar el impacto de los resultados 89
5. Conclusiones y Recomendaciones 91
5.1. Conclusiones 91
5.2. Recomendaciones 93
6. Referencias. 95

1. Introducción

1.1. Contexto

Actualmente vivimos en un mundo donde las nuevas tecnologías crecen a pasos agigantados y en intervalos de tiempo muy cortos. Tal es el caso especial del manejo y almacenamiento de la información en dispositivos electrónicos mediante tecnologías computacionales, basándose directamente en aumentar la capacidad de almacenamiento de los mismos, es decir, mayor volumen de información.

Gracias a este avance surge la necesidad de crear los métodos y técnicas necesarias para el manejo, procesamiento, visualización y análisis de grandes cantidades de datos.

El uso de bodegas de datos permite recopilar y almacenar grandes volúmenes de información de diferentes orígenes de datos, para posteriormente analizar la información por medio de patrones y parámetros establecidos para la búsqueda de conocimiento. Para crear los patrones de búsqueda y sus parámetros, se establece el uso de Minería de Datos.

Hoy en día en el Estado de Baja California Sur se aplica un sistema democrático para el proceso de elección a gobernador constitucional, mediante el cual, los ciudadanos de la entidad pueden ejercer el voto en favor de un candidato.

Para llevar a cabo el proceso de elección a gobernador en la entidad se organizan las elecciones electorales, las cuales se llevan a cabo con la participación de ciudadanos y del Instituto Estatal Electoral de Baja California Sur (IEEBCS).

Cada vez que se realizan elecciones electorales se genera una gran cantidad de información valiosa, donde las más importantes son los resultados electorales y la participación ciudadana. Los resultados electorales son el conteo final de todos y cada uno de los votos ejercidos de los ciudadanos por su candidato o partido de preferencia. La participación ciudadana es el concentrado de las personas que

votaron y las que no votaron en la elección. Esta información es el fundamento y/o respaldo de la constancia de mayoría que se le otorga al candidato ganador de la elección y es la única utilidad que se aplica a los resultados electorales.

Mediante la construcción de una bodega de datos alimentada por la información de la participación ciudadana y otras fuentes con información demográfica de los ciudadanos de la entidad y la creación de un modelo multidimensional, podemos obtener más información que normalmente no se encuentra a la vista. Por ejemplo, datos estadísticos de los votantes como: la edad promedio, el número de mujeres y de hombres que votaron, etc. Además, mediante la aplicación de técnicas de Minería de Datos, podemos obtener patrones de comportamiento en la participación ciudadana y dar respuesta a problemas con una reducción considerable en tiempo y costo. Por ejemplo la planeación estratégica de la promoción del voto.

1.2. Antecedentes

Hasta el día de hoy el IEEBCS ha organizado 3 elecciones para el cargo de Gobernador del Estado (en los años 1999, 2005, 2011). En la Tabla 1 se muestran los datos obtenidos en las elecciones realizadas por el IEEBCS:

Proceso Electoral	Personas que podían votar	Personas que votaron	Participación Ciudadana
1999	231,091	159,924	69%
2005	308,984	164,473	53%
2011	419,944	252,253	60%

Tabla 1.- Participación Ciudadana en Procesos Electorales organizados por el IEEBCS.

En cada uno de estos procesos se han generado bases de datos con información de los resultados electorales y de la participación ciudadana. Los resultados electorales contienen los votos totales recibidos por los candidatos y que fueron depositados en las urnas el día de la elección. Su principal y única función es obtener un ganador en la elección realizada, habiendo obtenido el mayor número

de votos. La participación ciudadana es el porcentaje de personas que ejercieron su derecho a votar entre el número de personas que podían votar.

Una de las atribuciones más importantes del IEEBCS es la Promoción del Voto, cuyo principal objetivo es el de incentivar a los ciudadanos a que acudan a las urnas el día de la jornada electoral para ejercer su derecho a voto libre y secreto, para la elección de los cargos de elección popular en la entidad.

1.3. Descripción del problema

Actualmente la promoción del voto se realiza mediante la creación de programas de fomento cívico, talleres, cursos, difusión en medios de comunicación y promoción de la imagen institucional del IEEBCS, pero todo de forma muy general. Es decir, un único plan de acción para todos los municipios del estado, sin tomar en cuenta diferentes factores que pueden influir en la promoción del voto. Como por ejemplo la edad de las personas a las que se aplicará, el nivel socioeconómico, grado de estudios promedio, la ubicación geográfica e incluso el clima, por mencionar algunos.

Para llevar a cabo una planeación estratégica por zonas dentro de la entidad, se debe realizar un análisis de la participación ciudadana presentada en procesos electorales de años anteriores. A la fecha el análisis de esta información no es posible ya que no se cuenta con los datos debidamente integrados y estructurados, además que no existe un modelo, método o sistema informático que permita llevar a cabo el análisis de la información.

Es por ello que se plantea el análisis de los datos históricos de los procesos electorales pasados para poder descubrir e identificar los patrones de comportamiento en la participación ciudadana y de este modo planificar una promoción del voto más versátil, estratégica y eficaz en la ciudadanía.

1.4. Objetivo general

Diseñar un modelo de minería de datos para el análisis de la participación ciudadana en procesos electorales del Estado de Baja California Sur y determinar las variables que intervienen en sus patrones de comportamiento.

1.5. Objetivos específicos

- Analizar el problema y comprender los datos correspondientes a los resultados electorales y a la participación ciudadana.
- Plantear el problema de minería de datos, en base al análisis de la información y selección de las bases de datos.
- Generar una bodega de datos mediante técnicas de ETL.
- Generar un modelo multidimensional con base a la información contenida en la bodega de datos para la generación de vistas para poder llevar a cabo un análisis de la información más adecuada.
- Analizar y seleccionar los algoritmos de minería de datos existentes y la compatibilidad o necesidad de utilizarlos en nuestro modelo.
- Identificar las variables más representativas en la aplicación de los algoritmos de minería de datos en el modelo.
- Construir un modelo de minería de datos, aplicando los algoritmos y variables seleccionadas.
- Evaluar y presentar los resultados obtenidos en la aplicación del modelo de minería de datos.

1.6. Alcances y limitaciones

Alcances.

- Estructurar y ordenar la información de procesos anteriores.
- Generar nueva información a partir del análisis de los datos.
- Crear un banco de información estadística de los procesos electorales realizados, clasificado por casilla, distrito y municipio.
- Obtener los patrones de comportamientos en la participación ciudadana de los procesos electorales.

- Identificar los factores de mayor influencia en el comportamiento de la participación ciudadana.

Limitaciones.

- Solo se aplica a los resultados electorales de las elecciones a gobernador.
- Solo se aplica en las elecciones realizadas en el estado de Baja California Sur.
- Se requiere información adicional de otros orígenes de datos.
- No todos los datos son de uso público.
- Los datos no tienen la misma estructura lógica.
- Información incompleta.

1.7. Justificación

Actualmente no existe un sistema, implementación informática o proceso automatizado que permita analizar los datos generados en los procesos electorales dentro del IEEBCS. En el mejor de los casos, en otros institutos estatales de otras entidades del país, solo se cuenta con información estadística a partir de los resultados. Solamente el Instituto Nacional Electoral (INE) cuenta con una plataforma web que permite consultar información estadística en los procesos nacionales pasados.

Por tal motivo se plantea el uso de una Bodega de Datos para almacenar la información y, mediante la aplicación de técnicas y algoritmos de Minería de Datos, obtener patrones de comportamiento capaces de describir la participación ciudadana en los procesos electorales.

Todo esto nos permite manejar un gran volumen de información con una inversión reducida, en un menor tiempo (casi inmediato una vez que se cargan los datos correctamente) y sin la necesidad de un gran despliegue de capital humano para realizar el análisis.

También permitirá al IEEBCS el análisis de la participación ciudadana generada en los procesos electorales, y así poder brindar apoyo en la planeación y toma de decisiones para la implementación de estrategias en la promoción del voto para futuros procesos electorales.

Se podrá dar respuesta a preguntas como:

- ¿Por qué la participación ciudadana en algunas regiones es mayor y en otras es menor?
- ¿Qué variables influyen?, es decir, el motivo por el cual existe un mayor interés por participar o ejercer su voto en ciertas regiones y cuáles fueron las variables de mayor influencia.

1.8. Hipótesis

Aplicando algoritmos de minería de datos es posible determinar los patrones de comportamiento de la participación ciudadana en procesos electorales que permita una mejor planeación de la promoción del voto en futuros procesos.

2. Marco teórico

2.1. Base de datos

Es una colección de archivos interrelacionados, creados con un sistema manejador de bases de datos. El contenido de una base de datos engloba la información que se almacena en archivos de una organización, de tal manera que los datos estén disponibles para los usuarios. Una finalidad de las bases de datos es eliminar la redundancia o al menos minimizarla. Los tres componentes principales de un sistema en base de datos son el hardware, el software DBMS y los datos a manejar, así como el personal encargado del manejo del sistema. [1]

2.2. Sistema Manejador de Base de Datos (DBMS)

Un DBMS es una colección de numerosas rutinas de software interrelacionadas, cada una de las cuales es responsable de una tarea específica.

El objetivo primordial de un sistema manejador de base de datos es proporcionar un entorno que sea a la vez conveniente y eficiente para ser utilizado al extraer, almacenar y manipular información de la base de datos. Todas las peticiones de acceso a la base de datos se manejan centralizadamente por medio del DBMS, por lo que este paquete funciona como interfaz entre los usuarios y la base de datos. [1]

2.3. Administrador de Base de Datos (DBA)

Es la persona o equipo de personas profesionales responsables de control y manejo del sistema de base de datos. Generalmente tiene(n) experiencia en DBMS, diseño de bases de datos, sistemas operativos, comunicación de datos, hardware y programación. [1]

2.4. Bodega de datos (Data Warehouse-DW)

Es un conjunto de datos integrados u orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales apoyan el proceso de toma de

decisiones de la administración y está orientada al manejo de grandes volúmenes de datos provenientes de diversas fuentes o diversos tipos. [2]

Estos datos cubren largos períodos de tiempo, lo que trae consigo que se tengan diferentes esquemas de los datos fuente. La concentración de esta información está orientada a su análisis para apoyar la toma de decisiones oportunas y fundamentadas. Previo a su utilización se deben aplicar procesos de análisis, selección y transferencia de datos seleccionados desde las fuentes.

El ciclo del desarrollo de la bodega de datos no difiere en mucho de las fases de perfeccionamiento de todos los desarrollos de software. Las fases y las secuencias son las mismas, pero existen variantes únicas asociadas a la bodega de datos y son las siguientes:

Planeación

En esta fase se determinan:

- El enfoque que se optará para la implementación: Top-Down (De arriba hacia abajo), Bottom-up (De abajo hacia arriba) o una combinación de estas dos.
- La metodología de desarrollo: Las más usuales son el método de análisis y diseño estructurado y el método del desarrollo en espiral.

Requerimientos

Especificación clara y precisa de las funciones que se esperan obtener de la bodega de datos. Estas funciones deben definirse desde varias perspectivas: propietario, arquitecto o desarrollador de la bodega de datos y desde la visión del usuario. Se definen las áreas tema que apoyará la bodega de datos, las dimensiones de categorización (tiempo, geografía, industria, grupo de clientes, línea de producto, etc.).

Análisis

Consiste en convertir todos los requerimientos conseguidos en la fase anterior en especificaciones concretas que sirvan de base para el diseño. Se definen los modelos lógicos de los datos para la bodega de datos, los mercados de datos, se establecen los procedimientos de conexión con las fuentes de datos y la bodega de datos y las herramientas de acceso del usuario final.

Diseño

Los modelos lógicos definidos en la fase anterior se convierten en modelos físicos. Se generan los diseños para programas y procesos que se requieren según la arquitectura, tanto a nivel de los datos como de aplicación.

Construcción

Se conoce también como diseño físico y consiste en plasmar en la práctica los diseños lógicos de la fase anterior. Incluye la construcción de programas que creen y modifiquen las bases de datos, que extraigan datos de las fuentes, programas para transformación de datos tales como integración, resumen y adición, programas para la actualización de los datos, programas para búsquedas en bases de datos muy grandes.

Montaje

Se refiere a la instalación, puesta en marcha y uso de la bodega de datos. Un elemento importante consiste en concientizar a los usuarios sobre la disponibilidad, beneficios y presentación de la bodega de datos, esto se conoce como comercialización de la información.

2.5. Modelos de bases de datos multidimensionales

En un modelo de datos multidimensional los datos se organizan alrededor de los temas de la organización, formando así la llamada tabla de hechos. La estructura de datos manejada en este modelo son matrices multidimensionales o hipercubos

que pueden ser estructurados en diferentes arquitecturas (dependiendo del uso que se le vaya a dar a los datos y del tipo de los mismos). [3]

2.6. Hipercubo

Un hipercubo consiste en un conjunto de celdas. Cada celda se identifica por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones. Un hipercubo, por tanto, deberá ser reestructurado cada vez que se le agreguen datos o se modifiquen los ya existentes, ya que la información no está en tablas sino organizada de manera dimensional.

2.7. Hecho

Es el objeto a analizar. Posee atributos de tipo cuantitativo llamados de hechos o de síntesis. Sus valores (medidas) se obtienen generalmente por la aplicación de una función estadística que resume un conjunto de valores en un único valor. Por ejemplo: ventas en dólares, cantidad de unidades en inventario, cantidad de unidades de producto vendidas, horas trabajadas, promedio de piezas producidas, consumo de combustible de un vehículo, etcétera.

2.8. Dimensiones

Representan cada uno de los ejes en un espacio multidimensional. Suministran el contexto en el que se obtienen las medidas de un hecho. Algunos ejemplos son: tiempo, producto, cliente, departamento, entre otras. Las dimensiones se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado. Los componentes de una dimensión se denominan niveles y se organizan en jerarquías. Por ejemplo la dimensión tiempo puede tener niveles día, mes y año.

Los hechos se guardan en tablas de hechos y las dimensiones en tablas de dimensiones. Sin embargo hay diferentes diseños que podemos usar, dependiendo de cómo se desee acceder a la información y del tipo de aplicación que vayamos a desarrollar.

2.9. Mercados de datos (DATA MARTS)

Si bien existen diversas estructuras de datos, a través de los cuales se pueden representar los datos del DW, solamente se entrará en detalle acerca de los cubos multidimensionales, por considerarse que esta estructura de datos es una de las más utilizadas y cuyo funcionamiento es el más complejo de entender.

Un cubo multidimensional o cubo, representa o convierte los datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones.

Los objetos más importantes que se pueden incluir en un cubo multidimensional son los siguientes:

- **Indicadores:** Sumas que se efectúan sobre algún hecho o expresiones basadas en sumas, pertenecientes a una tabla de hechos.
- **Atributos:** Campos o criterios de análisis, pertenecientes a tablas de dimensiones.
- **Jerarquías:** Representa una relación lógica entre dos o más atributos. [4]

2.10. Minería de datos

La minería de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y gestión de datos, procesamiento de datos, el modelo y las consideraciones de inferencia, métricas de Intereses, consideraciones de la Teoría de la Complejidad Computacional, post-procesamiento de las estructuras descubiertas, la visualización y actualización en línea. [4]

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias. Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada y puede ser utilizado en el análisis adicional o, por ejemplo, en la máquina de aprendizaje y análisis predictivo.

Por ejemplo, el paso de minería de datos podría identificar varios grupos en los datos que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones. Ni la recolección de datos, preparación de datos, ni la interpretación de los resultados y la información son parte de la etapa de minería de datos, pero pertenecen a todo el proceso KDD (Knowledge Discovery in Databases) como pasos adicionales.

Los términos relacionados con el dragado de datos, la pesca de datos y espionaje de los datos se refieren a la utilización de métodos de minería de datos a las partes de la muestra que son (o pueden ser) demasiado pequeños para las inferencias estadísticas fiables que se hicieron acerca de la validez de cualquiera de los patrones descubiertos. Estos métodos pueden, sin embargo, ser utilizados en la creación de nuevas hipótesis que se prueban contra las poblaciones de datos más grandes.

Un proceso típico de minería de datos consta de los siguientes pasos generales:

- **Selección del conjunto de datos.** Tanto en lo que se refiere a las variables objetivo (aquellas que se quieren predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.

- **Análisis de las propiedades de los datos.** En especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

- **Transformación del conjunto de datos de entrada**. Se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema. A este paso también se le conoce como procesamiento de los datos.

- **Seleccionar y aplicar la técnica de minería de datos**. Se construye el modelo predictivo, de clasificación o segmentación.

- **Extracción de conocimiento**. Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre procesamiento diferente de los datos.

- **Interpretación y evaluación de datos**. Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Si el modelo final no superara esta evaluación el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los

pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.

Una vez validado el modelo, éste ya está listo para su explotación. Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje PMML (Predictive Model Markup Language), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar.

Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes o bodegas de datos. De hecho, muchas grandes empresas e instituciones han creado y alimentan bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos no estructurados como información contenida en archivos de texto, en Internet, etc.

Como ya se ha comentado, las técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística. Dichas técnicas no son más que algoritmos más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

- **Redes neuronales**.- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red

que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

  - El Perceptrón.
  - El Perceptrón Multicapa.
  - Los Mapas Auto organizados, también conocidos como redes de Kohonen.

- **Regresión lineal**.- Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

- **Árboles de decisión**.- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

- **Modelos estadísticos**.- Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

- **Agrupamiento o Clustering**.- Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Por ejemplo:

  - Algoritmo K-medias.

- **Reglas de asociación**.- Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos

La Minería de Datos ha sufrido transformaciones en los últimos años de acuerdo con cambios tecnológicos, de estrategias de marketing, la extensión de los modelos de compra en línea, etc. Las transformaciones más importantes son:

- La importancia que han cobrado los datos no estructurados (texto, páginas de Internet, etc.).
- La necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.
- La exigencia de que los procesos funcionen prácticamente en línea (por ejemplo, en casos de fraude con una tarjeta de crédito).
- Los tiempos de respuesta. El gran volumen de datos que hay que procesar en muchos casos para obtener un modelo válido es un inconveniente pues esto implica grandes cantidades de tiempo de proceso y hay problemas que requieren una respuesta en tiempo real.

2.11. Tipos de minería de datos

Predicción

Muchas formas de minería de datos son predictivos. Por ejemplo, un modelo podría predecir el ingreso basado en la educación y otros factores demográficos. Las predicciones tienen una probabilidad asociada (es decir, ¿qué tan probable es que esta predicción se cumpla?). Las probabilidades de predicción son también conocidas como confianza (¿qué tanta confianza tengo de que esta predicción se cumpla?).

Algunas formas de minería de datos predictiva generan reglas, las cuales son condiciones que implican una salida dada. Por ejemplo, una regla podría especificar que una persona que tiene un grado universitario y vive en cierta

colonia probablemente tiene un ingreso mayor que el promedio en la región. Las reglas tienen un soporte asociado (¿Qué porcentaje de la población satisface esa regla?). [5]

Agrupación

La agrupación es otra forma en la que la minería de datos identifica grupos naturales en los datos. Por ejemplo, un modelo podría identificar el segmento de la población que tiene un ingreso dentro de un rango específico, que tiene un buen registro de manejo, y que arrienda un carro nuevo con base anual. [5]

2.12. Funciones de la minería de datos

Las funciones de minería de datos se dividen en dos categorías, supervisadas y no supervisadas.

Minería de datos supervisada.

El aprendizaje supervisado es también conocido como aprendizaje dirigido. El proceso de aprendizaje es dirigido por un atributo u objetivo dependiente previamente conocido.

El aprendizaje supervisado generalmente resulta en modelos predictivos. Este es el contraste para el aprendizaje no supervisado, donde la meta es la detección de patrones.

La construcción de un modelo supervisado involucra el entrenamiento, un proceso mediante el cual el software analiza muchos casos donde el valor objetivo ya es conocido.

En el proceso de entrenamiento, el modelo “aprende” la lógica de hacer la predicción. Por ejemplo, un modelo que busca identificar los clientes que probablemente respondan a una promoción, debe ser entrenado para que analice

las características de muchos clientes que ya se sabe que respondieron o no respondieron a una promoción en el pasado. [5]

Minería de datos no supervisada.

El aprendizaje no supervisado es no dirigido. No hay distinción entre atributos dependientes e independientes. Es decir, no hay un resultado previamente conocido que guie al algoritmo en la construcción del modelo. Por lo tanto, la minería de datos no supervisada puede ser usada para propósitos descriptivos. Aunque también puede ser usada para hacer predicciones. [5]

2.13. Herramientas de Minería de Datos

Microsoft SQL Server 2012

Microsoft SQL Server es un Sistema de manejo de bases de datos relacionales desarrollado por Microsoft. Como un software de bases de datos, su función principal es la de almacenar y recuperar información a través de consultas realizadas por otras aplicaciones de software, ya sea en la misma computadora o en otras conectadas por red. Hay muchas implementaciones de SQL server específicas para cargas de trabajo de distintos tamaños y para distintos tipos de aplicaciones, incluyendo distintos números de usuarios concurrentes. Su lenguaje primario de consultas es T-SQL y ANSI SQL. [6]

Microsoft SQL Server Business Intelligence Development Studio

Businnes Intelligence Development Studio (Estudio de desarrollo de inteligencia de negocios) es un ambiente de desarrollo integrado de Windows y se usa para desarrollar análisis de datos e inteligencia de negocios utilizando los servicios de análisis de Microsoft SQL Server, servicios de reportes y servicios integrados.

Está basado en el ambiente de desarrollo de Microsoft Visual Studio, pero se le agregaron servicios específicos de SQL server, así como tipos de proyectos, incluyendo herramientas, controles y proyectos para reportes, flujos de datos ETL, cubos OLAP, y estructuras de minería de datos. [7]

Microsoft SQL Server Analysis Services

Microsoft SQL Server Analysis Services (Servicios de Análisis de SQL Server) es una herramienta OLAP, de minería de datos y de reportes de Microsoft SQL Server. Es usada como una herramienta para analizar y dar sentido a información que podría estar repartida en múltiples bases de datos o en distintas tablas. [8]

Este recurso viene integrado en SQL Server como una herramienta de inteligencia de negocios y de bodegas de datos.

Microsoft Visual Studio 2012

MS Visual Studio es un ambiente de desarrollo integrado de Microsoft. Se usa para desarrollar programas de computadora para la familia de sistemas operativos Microsoft Windows, así como sitios web, aplicaciones web y servicios web. Visual Studio usa plataformas de desarrollos de software de Windows tales como las Windows API, Formas de Windows, Microsoft Silverlight etc.

3. Metodología de la investigación

3.1. Metodologías de minería de datos

Permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Estas metodologías nos ayudan a entender el proceso de descubrimiento de conocimiento para proveer una guía de planificación y ejecución en los proyectos.

Algunos modelos conocidos como metodologías son en realidad un modelo de proceso: un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo.

La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer y la metodología especifica cómo hacerlo. Dentro de las metodologías más adecuadas para la planificación de este proyecto se encuentran KDD y CRISP-DM.

3.2. Metodología KDD

La Extracción de conocimiento está principalmente relacionado con el proceso de descubrimiento conocido como *Knowledge Discovery in Databases* (KDD), que descubre conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. [5]

No es un proceso automático, es un proceso repetitivo que explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos. La Figura 1 ilustra las etapas del proceso KDD:

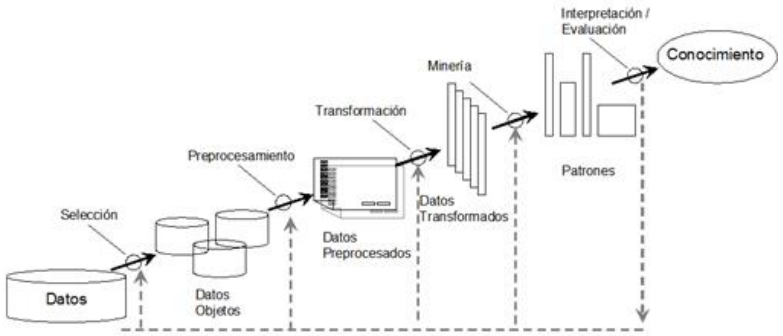

Figura 1.- Etapas del Proceso KDD.

Las etapas del proceso KDD son:

1. **Selección de datos.** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

2. **Pre-procesamiento.** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

3. **Transformación.** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

4. **Minería de Datos.** Es la fase del modelamiento en donde los algoritmos de minería de datos son aplicados, con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u “ocultos” en los datos.

5. **Interpretación y evaluación.** Se identifican los patrones obtenidos y se seleccionan los que realmente son útiles, basándose en las medidas planteadas y se realiza una evaluación de los resultados obtenidos.

3.3. Metodología CRISP-DM

CRISP-DM se encuentra dentro de las metodologías más utilizadas para la elaboración de proyectos de minería de datos. Está basada en actividades ordenadas en seis fases que recorren todo el proceso de minería de datos, desde la definición de los objetivos del negocio que se pretende obtener, hasta la vigilancia y el mantenimiento del modelo que se proponga e implemente.

Cada una de esas fases se ha subdividido a su vez en tareas ordenadas en un esquema jerárquico, desde un mayor a un menor nivel de detalle. En la Figura 2 se muestran las fases de la metodología CRISP-DM.

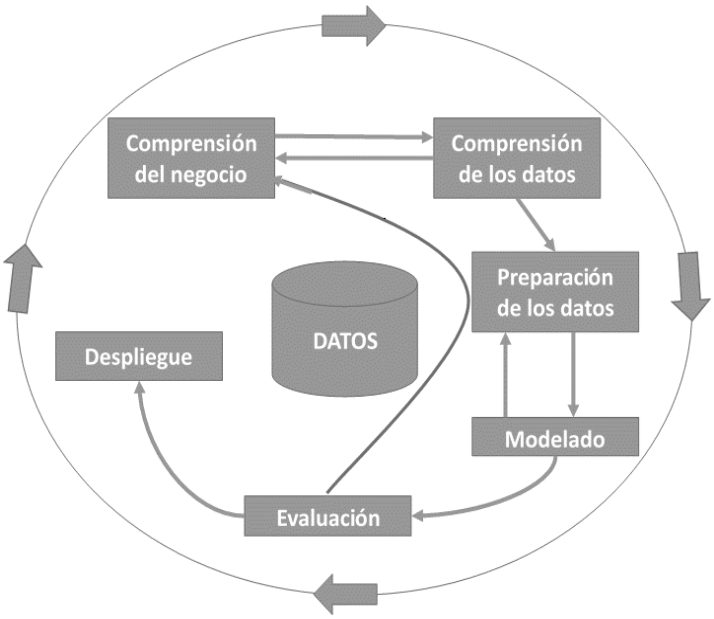

Figura 2.- Fases de la Metodología CRISP-DM.

A continuación se describen en mayor detalle las fases de la metodología CRISP-DM:

1. Fase de comprensión del negocio: Se centra en la comprensión de los objetivos del proyecto de minería de datos desde un punto de vista de negocios. Esta fase es equivalente, por lo tanto, a una fase de análisis de requerimientos de un proyecto de desarrollo de software y es importante porque el cliente puede no tener claro qué es lo que quiere. Las tareas a realizar en esta fase incluyen determinar los objetivos de negocios, evaluar la situación del proyecto en términos de recursos, restricciones y suposiciones, determinar objetivos de minería de datos que traduzcan a criterios técnicos los objetivos de negocios y, finalmente, producir el plan del proyecto.

2. Fase de comprensión de los datos: Comprende la recolección inicial de datos, identificando la calidad de estos y estableciendo las relaciones más evidentes entre ellos. Incluye la tarea de recolección de datos iniciales, en la cual deben inscribirse los datos en términos de número de registros, número de campos por registro y significado de cada campo. Incluye también la tarea de descripción de los datos en términos de tipo, distribución, tablas de frecuencia y estadígrafos. Tareas adicionales de esta fase son la exploración de los datos mediante gráficos y tablas, y la verificación de la calidad de los mismos. La verificación sobre la calidad de los datos debe efectuarse para asegurar la consistencia de la información proveniente de bases de datos diferentes, proporcionadas por diferentes entidades y con fechas de proceso diferentes.

3. Fase de preparación de los datos: En esta fase debe construirse una base de datos que debe contener todas las características consideradas candidatas para estimar el valor de una variable que se espera predecir. Esta fase incluye la tarea de selección de los datos a los que se va aplicar la técnica del modelo, la tarea de limpieza de los mismos para alcanzar el nivel de calidad requerido por las técnicas de minería de datos que sean seleccionadas, la tarea de construir datos adicionales, la tarea de integrar diferentes bases de datos, y la tarea de formatear los datos. La fase de preparación de los datos debe entregar datos que estén en un formato adecuado para la técnica del modelo que se empleará en la siguiente

fase. Por esto, la fase de modelado puede requerir regresar una o más veces a la fase de preparación de los datos.

4. Fase de modelado: Esta es la fase medular de un proyecto de minería de datos y consiste en descubrir una relación entre un conjunto de variables y una variable que se espera predecir. Contempla la selección de una técnica de modelado, entre las cuales pueden mencionarse las redes de KOHONEN o modelos K-MEAN para agrupamiento, árboles de decisión C5 o C&R para segmentación, redes neuronales o regresión logística para predicción, inducción de reglas generalizadas para descubrimiento de patrones y análisis de factores para reducir la complejidad de los datos, entre otras. Entre sus tareas se encuentran la selección de la técnica del modelo, la generación del diseño de las pruebas del modelo, la construcción del mismo y, finalmente, la evaluación técnica a la que debe someterse el modelo a través de criterios estadísticos.

5. Fase de evaluación de negocios: En la fase de modelado se evalúa el modelo de forma técnica en relación a factores tales como su precisión y generalidad. En esta fase, en cambio, debe evaluarse el nivel de satisfacción de los objetivos de negocios perseguidos por el proyecto de minería de datos. Incluye la tarea de evaluar los resultados, la tarea de revisar el proceso de minería de datos y, finalmente, la tarea de determinar los próximos pasos a seguir (momento en el que debe decidirse si debe darse por terminado el proyecto de minería de datos y entrar en la fase de despliegue, si deben iniciarse iteraciones adicionales, o si debe iniciarse un nuevo proyecto de minería de datos).

6. Fase de despliegue del modelo: En esta fase deberá definirse una estrategia para implementar los resultados de la minería de datos. Incluye las tareas de planificar el despliegue del modelo, de planificar el monitoreo y el mantenimiento de los modelos, de generar el reporte final del proyecto, y de revisar el proyecto en relación a evaluar lo que ocurrió correctamente y lo que necesita ser mejorado.

3.4. Selección de la metodología

La selección de la metodología se hizo con base a la comparación de las metodologías existentes y las más utilizadas. La metodología más conveniente para llevar a cabo nuestro proyecto es la metodología CRISP-DM.

La metodología CRISP-DM es una de las más completas y constituye un mapa de ruta que permite determinar qué actividades desarrollar en qué etapa, para alcanzar los objetivos finales del proyecto. Es una de las principales metodologías utilizadas, junto con la metodología SEMMA.

Contiene las fases de un proyecto, sus tareas respectivas y las relaciones entre estas tareas.

Algunas de las fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores.

Una de las ventajas de la metodología CRISP-DM es que este sistema está diseñado como una metodología independiente de la herramienta tecnológica a utilizar en la explotación de datos. Esto último la hace más flexible. Otro punto importante es que es de acceso libre y cumple con la característica de ser orientada al negocio.

4. Metodología

4.1. Comprensión del negocio

4.1.1. Determinar los objetivos del negocio

4.1.1.1. Contexto

El Instituto Estatal Electoral de Baja California Sur tiene su residencia en la Ciudad de la Paz, capital del Estado de Baja California Sur. Actualmente está compuesto por las Direcciones Ejecutivas:

- Dirección Ejecutiva de Organización Electoral (DEOE).
- Dirección Ejecutiva de Educación Cívica y Capacitación Electoral (DEECCE).
- Dirección Ejecutiva de Prerrogativas y Partidos Políticos (DEPPP).
- Dirección de Quejas y Denuncias y de Procedimiento Contencioso Electoral (DQDPCE).
- Dirección de Transparencia y Acceso a la Información y del Servicio Profesional Electoral (DTAISPE).
- Dirección Ejecutiva de Administración y Finanzas (DEAF).
- Unidad de Cómputo y Servicios Informáticos (UCSI).

Las direcciones que llevan a cabo actividades relacionadas con la participación ciudadana y que son de importancia para el presente proyecto de investigación son la DEOE y la DEECCE.

La DEOE es la encargada de llevar a cabo el proceso de ubicación de casillas, el cual consiste en la planeación e instalación de las casillas a utilizar en la Jornada Electoral. Con base a este proceso se genera una base de datos con los datos técnicos y de ubicación de cada una de las casillas a instalar durante el proceso electoral.

Dentro de las funciones y actividades más importantes de la DEECCE se encuentra la Promoción del Voto, esto se lleva a cabo mediante la creación de

programas de promoción y difusión de la imagen institucional y el incentivo de la participación ciudadana. Además es la encargada de medir la participación ciudadana que se dio en un proceso electoral.

En general, una de las actividades principales realizadas por el Instituto y que se lleva previamente a un proceso electoral es la Promoción del Voto, que consiste en la creación y planeación de estrategias y programas para promover la participación ciudadana, es decir, motivar a la población para que acuda a ejercer su voto el día de la elección.

4.1.1.2. Objetivos del negocio

Los objetivos principales del Instituto son:

- Coordinar, preparar, desarrollar y vigilar los procesos electorales.
- Promover y difundir los principios y valores de la cultura democrática.
- Asegurar a los ciudadanos el ejercicio de los derechos político-electorales y vigilar el cumplimiento de sus obligaciones.
- Velar por la autenticidad y efectividad del sufragio.
- Llevar a cabo la promoción del voto y la difusión de la educación cívica y la cultura democrática.

4.1.1.3. Criterios de éxito

Obtener un alto porcentaje en la participación ciudadana de procesos electorales futuros realizados por el IEEBCS, para lo que se requiere llevar a cabo una promoción eficaz del voto. También se desea poder realizar una promoción del voto más específica y enfocada con base a los diferentes sectores sociales de la ciudadanía, ya que actualmente se realiza una promoción del voto muy general.

Por lo tanto si mediante el análisis de los datos se pueden identificar los patrones de comportamiento en la participación ciudadana y con ello poder llevar a cabo

una promoción del voto más específica por zonas, tendríamos un alto porcentaje de participación ciudadana en procesos futuros.

4.1.2. Evaluar la situación

4.1.2.1. Inventario de recursos

Actualmente el IEEBCS cuenta con las herramientas tecnológicas y con el capital humano necesario para llevar a cabo el proyecto. Además, se cuenta con los datos y el software necesarios para su análisis.

En la parte del personal, el IEEBCS cuenta con una Unidad de Cómputo y Servicios Informáticos, integrada por un Director de la Unidad, un Coordinador de la Unidad, un Desarrollador de Sistemas y un Ingeniero de Soporte.

En la parte de la infraestructura, el Instituto cuenta con servidores físicos:

- **Servidor de datos.**
  - Marca: HP.
  - Modelo: ProLiant ML350p GenB.
  - Procesador: Intel Xeon E5-2600.
  - Memoria Ram: 16Gb DDR3.
  - Almacenamiento: 500Gb.
  - Sistema Operativo: Windows Server 2012.
- **Servidor de dominio.**
  - Marca: HP.
  - Modelo: ProLiant ML110.
  - Procesador: Intel Xeon.
  - Memoria Ram: 4Gb DDR3.
  - Almacenamiento: 160Gb.
  - Sistema Operativo: Windows Server 2008.

- **Servidor proxy/firewall**.
    - Marca: HP.
    - Modelo: DX2300.
    - Procesador: Intel Pentium 4.
    - Memoria Ram: 2Gb DDR2.
    - Almacenamiento: 80Gb.
    - Sistema Operativo: IPFire.

En la parte de software, se cuenta con Microsoft Windows Server 2012, Microsoft SQL Server 2012, Microsoft Visual Studio 2012 Ultimate y Profesional.

4.1.2.2. Terminología

Participación ciudadana.

Es el conjunto de acciones o iniciativas que pretenden impulsar el desarrollo local y la democracia participativa a través de la integración de la comunidad al ejercicio de la política. Está basada en varios mecanismos para que la población tenga acceso a las decisiones del gobierno de manera independiente sin necesidad de formar parte de la administración pública o de un partido político.

Minería de datos.

Es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

4.1.2.3. Factibilidad

Para el desarrollo del proyecto ya se cuenta con los datos necesarios para el análisis, los equipos de cómputo y tecnológicos necesarios, con el personal

adecuado para el soporte y con el software requerido. Por lo tanto no generaría un costo adicional.

Para la implementación también se cuenta con todo lo necesario en cuanto a datos, equipos de cómputo y tecnológicos, personal y software. En conclusión tampoco genera costo adicional.

Por lo tanto la implementación de este proyecto es completamente factible considerando que los beneficios serían mayores y la inversión necesaria para implementarlo es mucho menor, casi nulo.

4.1.3. Determinar los objetivos de minería de datos

4.1.3.1. Objetivo de minería de datos

El objetivo principal por el cual se desea implementar un modelo de minería de datos para el análisis de la participación ciudadana es poder identificar los factores o características similares que existen entre las personas que acudieron a votar y también las que no ejercieron tal derecho.

Esto con la finalidad de identificar si existe un patrón similar entre los votantes y con base a ello realizar una promoción del voto basado en esos factores identificados en el patrón obtenido.

4.1.3.2. Criterios de éxito de minería de datos

Para poder decir que el análisis de la información mediante minería de datos fue exitoso en nuestro proyecto, debemos obtener los patrones de comportamiento en la participación ciudadana y poder identificar los factores o variables que influyen en el comportamiento, para llevar a cabo una planeación de la promoción del voto mucho más estratégica y especializada.

Con esto se obtendría un mayor porcentaje de participación ciudadana en los procesos electorales futuros a realizar por el IEEBCS.

4.1.4. Producir el plan de proyecto

4.1.4.1. Evaluación inicial de funciones y algoritmos

Para llevar a cabo el proyecto se utilizará la herramienta de Microsoft SQL Server 2012, la cual cuenta con un módulo dedicado a la Inteligencia de Negocios, que permite la construcción de bodega de datos, el uso de minería de datos y sus diferentes algoritmos así como la generación de cubos y vistas para modelos multidimensionales.

Los algoritmos con los que viene incluido dicha herramienta son:

Algoritmo de Asociación

Es un algoritmo de asociación suministrado por Analysis Services, útil para los motores de recomendación. Un motor de recomendación recomienda productos a los clientes basándose en los elementos que ya han adquirido o en los que tienen interés. El algoritmo de asociación de Microsoft recorre un conjunto de datos para hallar elementos que aparezcan juntos en un caso.

Algoritmo de Clústeres

Es un algoritmo de segmentación suministrado por Analysis Services. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares.

Algoritmo de Arboles de Decisión

Es un algoritmo de clasificación y regresión. Usado para el modelo de predicción de atributos discretos y continuos.

Algoritmo de Regresión Lineal

Es una variación del algoritmo de árboles de decisión que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación, utilizar esa relación para la predicción.

Algoritmo de Regresión Logística

Este algoritmo se ha implementado utilizando una variación del algoritmo de Red neuronal. Este algoritmo comparte muchas de las cualidades de las redes neurales pero es más fácil de entrenar.

Algoritmo Bayes-Naive

Es un algoritmo de clasificación basado en los teoremas de Bayes y que Microsoft SQL Server Analysis Services proporciona para el modelado de predicción.

Desde el punto de vista computacional, el algoritmo es menos complejo que otros algoritmos y, por tanto, resulta útil para generar rápidamente modelos de minería de datos que detectan las relaciones entre las columnas de entrada y las columnas de predicción.

4.2. Comprensión de los datos

4.2.1. Recolectar datos iniciales

La información a utilizar se encuentra distribuida en diferentes bases de datos, de las cuales en su mayoría son conformadas por el IEEBCS, por ejemplo la información de la participación ciudadana en las elecciones llevadas a cabo en los años 2005 y 2011. Además se cuenta con una base de datos que contiene la ubicación de las casillas electorales en los mismos procesos, 2005 y 2011.

También se requiere de información de orígenes externos como por ejemplo datos estadísticos y demográficos (proporcionados por el Instituto Nacional de Estadística y Geografía) en específico de la base de datos denominada Datos Geo Electorales.

4.2.1.1. Reporte de recolección de datos iniciales

Una vez llevada a cabo la recolección de los datos iniciales, correspondientes a los diferentes orígenes de datos que se van a utilizar, obtuvimos la información que se muestra en las Tablas 2, 3, 4, 5, 6 y 7:

Participación Ciudadana Proceso Electoral 2005

Campo	Datos de Ejemplo	
PATERNO	PEREZ	HERNANDEZ
MATERNO	LOPEZ	LEON
NOMBRE	JUAN	MARIA
EDAD	35	23
SEXO	H	M
CALLE	NAVEGANTES #45	PASEO DEL CORTEZ
COLONIA	CENTRO	PEDREGAL
ENTIDAD	3	3
MUNICIPIO	1	1
DISTRITO	1	1
SECCIÓN	121	122
CASILLA	B	C
CLAVE_ELEC	PRLPJU75012303H500	HRLNMA87020303M200
VOTÓ	SI	NO

Tabla 2.- Ejemplo Participación Ciudadana Proceso Electoral 2005

Ubica Tu Casilla Proceso Electoral 2005

Campo	Datos de Ejemplo	
ENTIDAD	3	3
MUNICIPIO	1	1
DISTRITO	1	1
SECCIÓN	123	124
CASILLA	B	C
TIPO_DOMICILIO	PARTICULAR	ESCUELA
DOMICILIO	IGNACIO ALLENDE #456	ESC. PRIMARIA 18 DE MARZO
MANZANA	23	56

Tabla 3.- Ejemplo Ubica Tu Casilla Proceso Electoral 2005

Datos Geo Electorales 2005

Campo	Datos de Ejemplo	
ENTIDAD	3	3
MUNICIPIO	1	1
DISTRITO_FEDERAL	1	2
SECCIÓN	315	127
GRA_PROM_ESC	5	10
18YMAS	1410	613
18YMAS_F	505	307
18YMAS_M	905	306
18YMAS_EDUC_PB	185	340
18YMAS_EDUC_PB_F	91	165
18YMAS_EDUC_PB_M	94	175
CASADOS	360	379
VIVHAB_MAT	181	248
VIVHAB_SINMAT	29	2
VIVHAB_PC	67	117
VIVHAB_DREN	178	248
VIVHAB_TV	205	246
VIVHAB_AUT	162	195
TOTAL_VIVIENDAS	210	250

Tabla 4.- Ejemplo Datos Geo Electorales 2005

Participación Ciudadana Proceso Electoral 2011

Campo	Datos de Ejemplo	
PATERNO	PEREZ	HERNANDEZ
MATERNO	LOPEZ	LEON
NOMBRE	JUAN	MARIA
EDAD	35	23
SEXO	H	M
CALLE	NAVEGANTES #45	PASEO DEL CORTEZ
COLONIA	CENTRO	PEDREGAL
ENTIDAD	3	3

MUNICIPIO	1	1
DISTRITO	1	1
SECCIÓN	121	122
CASILLA	B	C
CLAVE_ELEC	PRLPJU75012303H500	HRLNMA87020303M200
VOTÓ	SI	NO

Tabla 5.- Ejemplo Participación Ciudadana Proceso Electoral 2011

Ubica Tu Casilla Proceso Electoral 2011

Campo	Datos de Ejemplo	
ENTIDAD	3	3
MUNICIPIO	1	1
DISTRITO	1	1
SECCIÓN	135	136
CASILLA	B	C
TIPO_DOMICILIO	PARTICULAR	ESCUELA
DOMICILIO	GUILLERMO PRIETO #987	ESC. PRIMARIA 3 DE MAYO
MANZANA	78	23

Tabla 6.- Ejemplo Ubica Tu Casilla Proceso Electoral 2011

Datos Geo Electorales 2010

Campo	Datos de Ejemplo	
ENTIDAD	3	3
MUNICIPIO	1	1
DISTRITO_FEDERAL	1	2
SECCIÓN	127	315
GRA_PROM_ESC	10	5
18YMAS	613	1410
18YMAS_F	307	505
18YMAS_M	306	905
18YMAS_EDUC_PB	340	185
18YMAS_EDUC_PB_F	165	91
18YMAS_EDUC_PB_M	175	94

CASADOS	379	360
VIVHAB_MAT	248	181
VIVHAB_SINMAT	2	29
VIVHAB_PC	117	67
VIVHAB_DREN	248	178
VIVHAB_TV	246	205
VIVHAB_AUT	195	162
TOTAL_VIVIENDAS	250	210

Tabla 7.- Ejemplo Datos Geo Electorales 2010

4.2.2. Descripción de los datos

Una vez que se llevó a cabo la recolección y análisis de los datos se procede a la descripción de los mismos.

4.2.2.1. Reporte de descripción de los datos

A continuación se presenta el reporte obtenido de la descripción de los datos en las Tablas 8, 9, 10, 11, 12 y 13.

Tabla: Participación Ciudadana Proceso Electoral 2005

Número de Registros: 308,984

Campo	Tipo de Datos	Descripción
PATERNO	Texto	Apellido Paterno del Ciudadano
MATERNO	Texto	Apellido Materno del Ciudadano
NOMBRE	Texto	Nombre(s) del Ciudadano
EDAD	Número	Edad del Ciudadano al momento de votar
SEXO	Carácter	Sexo del Ciudadano
CALLE	Texto	Calle del domicilio registrado por el Ciudadano
COLONIA	Texto	Colonia del domicilio registrado por el Ciudadano
ENTIDAD	Número	Entidad a la que pertenece el Ciudadano
MUNICIPIO	Número	Municipio al que pertenece el Ciudadano
DISTRITO	Número	Distrito al que pertenece el Ciudadano
SECCIÓN	Número	Sección Electoral a la que pertenece el

		Ciudadano
CASILLA	Carácter	Casilla a la que pertenece el Ciudadano
CLAVE_ELEC	Texto	Clave de Elector del Ciudadano
VOTÓ	Lógico	Especifica si votó o no el ciudadano

Tabla 8.- Descripción Participación Ciudadana Proceso Electoral 2005.

Tabla: Ubica Tu Casilla Proceso Electoral 2005

Número de Registros: 643

Campo	Tipo de Datos	Descripción
ENTIDAD	Número	Entidad donde está ubicada la casilla
MUNICIPIO	Número	Municipio donde está ubicada la casilla
DISTRITO	Número	Distrito donde está ubicada la casilla
SECCIÓN	Número	Sección Electoral donde está ubicada la casilla
CASILLA	Carácter	Tipo de casilla
TIPO_DOMICILIO	Texto	Tipo de domicilio donde está ubicada la casilla
DOMICILIO	Texto	Domicilio donde está ubicada la casilla
MANZANA	Número	Número de manzana de la ubicación geográfica de la casilla

Tabla 9.- Descripción Ubica Tu Casilla Proceso Electoral 2005.

Tabla: Datos Geo Electorales 2005

Número de Registros: 434

Campo	Tipo de Datos	Descripción
ENTIDAD	Número	Entidad a la que pertenece la sección.
MUNICIPIO	Número	Municipio al que pertenece la sección.
DISTRITO_FEDERAL	Número	Distrito federal al que pertenece la sección.
SECCIÓN	Número	Sección electoral.
GRA_PROM_ESC	Carácter	Grado Promedio Escolar de la sección.
18YMAS	Número	Personas de 18 años o más.
18YMAS_F	Número	Mujeres de 18 años o más.
18YMAS_M	Número	Hombres de 18 años o más.
18YMAS_EDUC_PB	Número	Personas de 18 años o más con educación pos básica.

18YMAS_EDUC_PB_F	Número	Mujeres de 18 años o más con educación pos básica.
18YMAS_EDUC_PB_M	Número	Hombres de 18 años o más con educación pos básica.
CASADOS	Número	Personas casadas.
VIVHAB_MAT	Número	Viviendas con piso de material.
VIVHAB_SINMAT	Número	Viviendo sin piso de material.
VIVHAB_PC	Número	Viviendas con computadora.
VIVHAB_DREN	Número	Viviendas con drenaje.
VIVHAB_TV	Número	Viviendas con televisión.
VIVHAB_AUT	Número	Viviendas con automóvil.
TOTAL_VIVIENDAS	Número	Total de viviendas en la sección.

Tabla 10.- Descripción Datos Geo Electorales 2005.

Tabla: Participación Ciudadana Proceso Electoral 2011

Número de Registros: 419,944

Campo	Tipo de Datos	
PATERNO	Texto	Apellido Paterno del Ciudadano
MATERNO	Texto	Apellido Materno del Ciudadano
NOMBRE	Texto	Nombre(s) del Ciudadano
EDAD	Número	Edad del Ciudadano al momento de votar
SEXO	Carácter	Sexo del Ciudadano
CALLE	Texto	Calle del domicilio registrado por el Ciudadano
COLONIA	Texto	Colonia del domicilio registrado por el Ciudadano
ENTIDAD	Número	Entidad a la que pertenece el Ciudadano
MUNICIPIO	Número	Municipio al que pertenece el Ciudadano
DISTRITO	Número	Distrito al que pertenece el Ciudadano
SECCIÓN	Número	Sección Electoral a la que pertenece el Ciudadano
CASILLA	Carácter	Casilla a la que pertenece el Ciudadano
CLAVE_ELEC	Texto	Clave de Elector del Ciudadano
VOTÓ	Lógico	Especifica si votó o no el ciudadano

Tabla 11.- Descripción Participación Ciudadana Proceso Electoral 2011.

Tabla: Ubica Tu Casilla Proceso Electoral 2011

Número de Registros: 825

Campo	Tipo de Datos	
ENTIDAD	Número	Entidad donde está ubicada la casilla
MUNICIPIO	Número	Municipio donde está ubicada la casilla
DISTRITO	Número	Distrito donde está ubicada la casilla
SECCIÓN	Número	Sección Electoral donde está ubicada la casilla
CASILLA	Carácter	Tipo de casilla
TIPO_DOMICILIO	Texto	Tipo de domicilio donde está ubicada la casilla
DOMICILIO	Texto	Domicilio donde está ubicada la casilla
MANZANA	Número	Número de manzana de la ubicación geográfica de la casilla

Tabla 12.- Descripción Ubica Tu Casilla Proceso Electoral 2011.

Tabla: Datos Geo Electorales 2010

Número de Registros: 434

Campo	Tipo de Datos	Descripción
ENTIDAD	Número	Entidad a la que pertenece la sección.
MUNICIPIO	Número	Municipio al que pertenece la sección.
DISTRITO_FEDERAL	Número	Distrito federal al que pertenece la sección.
SECCIÓN	Número	Sección electoral.
GRA_PROM_ESC	Carácter	Grado Promedio Escolar de la sección.
18YMAS	Número	Personas de 18 años o más.
18YMAS_F	Número	Mujeres de 18 años o más.
18YMAS_M	Número	Hombres de 18 años o más.
18YMAS_EDUC_PB	Número	Personas de 18 años o más con educación pos básica.
18YMAS_EDUC_PB_F	Número	Mujeres de 18 años o más con educación pos básica.
18YMAS_EDUC_PB_M	Número	Hombres de 18 años o más con educación pos básica.
CASADOS	Número	Personas casadas.
VIVHAB_MAT	Número	Viviendas con piso de material.

VIVHAB_SINMAT	Número	Viviendo sin piso de material.
VIVHAB_PC	Número	Viviendas con computadora.
VIVHAB_DREN	Número	Viviendas con drenaje.
VIVHAB_TV	Número	Viviendas con televisión.
VIVHAB_AUT	Número	Viviendas con automóvil.
TOTAL_VIVIENDAS	Número	Total de viviendas en la sección.

Tabla 13.- Descripción Datos Geo Electorales 2010.

4.2.3. Exploración de los datos

Después de realizar el análisis y descripción de los datos se procede a examinar la información y seleccionar los datos necesarios para nuestro modelo, desde cada uno de los orígenes de datos recolectados. Así como una descripción breve de por qué fueron seleccionados los datos.

4.2.3.1. Reporte del modelo descriptivo

A continuación se presenta el reporte de las tablas y los campos seleccionados para el modelo descriptivo, así como una breve descripción de la importancia del campo seleccionado. Las Tablas 14, 15, 16, 17, 18 y 19 son las que conforman el reporte.

Tabla: Participación Ciudadana Proceso Electoral 2005

Campo	Importancia
EDAD	Este campo nos permite clasificar y/o agrupar a los ciudadanos por medio de su edad, incluso crear conjuntos por rangos de edad.
SEXO	Nos permite saber y clasificar el sexo de la persona.
COLONIA	Nos permite agrupar a los ciudadanos por la colonia de su domicilio.
MUNICIPIO	Nos permite filtrar a nivel de detalle por medio del Municipio y agruparlos por este mismo.
DISTRITO	Nos permite filtrar a nivel de detalle por medio del Distrito y agruparlos por este mismo.
SECCIÓN	Nos permite filtrar a nivel de detalle por medio de la Sección y agruparlos por esta mismo.
CASILLA	Nos permite saber en qué casilla votó, el tipo de casilla y su ubicación.

VOTÓ	Clasificar y separar los ciudadanos que votaron y los que no votaron.

Tabla 14.- Exploración Participación Ciudadana Proceso Electoral 2005.

Tabla: Ubica Tu Casilla Proceso Electoral 2005

Campo	Tipo de Datos
SECCIÓN	Nos permite hacer la relación y cruce, en conjunto con la casilla, para los datos del ciudadano.
CASILLA	Nos permite hacer la relación y cruce, en conjunto con la sección electoral, para los datos del ciudadano. Ya que en una misma sección puede contener diversas casillas.
TIPO_DOMICILIO	Nos permite identificar el tipo de domicilio de la casilla y si tuvo influencia en el número de votos que se registraron en la casilla.
MANZANA	Nos permite hacer la relación y cruce con la información demográfica de un origen de datos externo.

Tabla 15.- Exploración Ubica Tu Casilla Proceso Electoral 2005.

Tabla: Datos Geo Electorales 2005

Campo	Tipo de Datos
SECCIÓN	Nos permite hacer la relación y cruce para los datos del ciudadano.
GRA_PROM_ESC	Determinar si el grado promedio de escolaridad está estrictamente relacionado con el porcentaje de la participación ciudadana en una sección electoral.
18YMAS_F	Determinar el porcentaje de mujeres mayores de edad en una sección y definir si es parte de los parámetros influyentes en la participación ciudadana.
18YMAS_M	Determinar el porcentaje de hombres mayores de edad en una sección y definir si es parte de los parámetros influyentes en la participación ciudadana.
18YMAS_EDUC_PB	Determinar si las personas con educación pos básica son los que menos votan en una sección o su relación con otros datos.
CASADOS	Determinar que tanto influye el estado civil en una sección para la participación ciudadana.
VIVHAB_MAT	Determinar si el factor de contar con un piso de material en la vivienda en relevante para determinar si las personas acuden o no a votar.

VIVHAB_SINMAT	Determinar si el factor de NO contar con un piso de material en la vivienda en relevante para determinar si las personas acuden o no a votar.
VIVHAB_PC	Determinar si tiene influencia contar con una computadora en la vivienda para la participación ciudadana.
VIVHAB_TV	Determinar si tiene influencia contar con una televisión en la vivienda para la participación ciudadana.
VIVHAB_AUT	Determinar si tiene influencia contar con un automóvil en la vivienda para la participación ciudadana.

Tabla 16.- Exploración Datos Geo Electorales 2005.

Tabla: Participación Ciudadana Proceso Electoral 2011

Campo	Importancia
EDAD	Este campo nos permite clasificar y/o agrupar a los ciudadanos por medio de su edad, incluso crear conjuntos por rangos de edad.
SEXO	Nos permite saber y clasificar el sexo de la persona.
COLONIA	Nos permite agrupar a los ciudadanos por la colonia de su domicilio.
MUNICIPIO	Nos permite filtrar a nivel de detalle por medio del Municipio y agruparlos por este mismo.
DISTRITO	Nos permite filtrar a nivel de detalle por medio del Distrito y agruparlos por este mismo.
SECCIÓN	Nos permite filtrar a nivel de detalle por medio de la Sección y agruparlos por esta mismo.
CASILLA	Nos permite saber en qué casilla votó, el tipo de casilla y su ubicación.
VOTÓ	Clasificar y separar los ciudadanos que votaron y los que no votaron.

Tabla 17.- Exploración Participación Ciudadana Proceso Electoral 2011.

Tabla: Ubica Tu Casilla Proceso Electoral 2011

Campo	Tipo de Datos
SECCIÓN	Nos permite hacer la relación y cruce, en conjunto con la casilla, para los datos del ciudadano.
CASILLA	Nos permite hacer la relación y cruce, en conjunto con la sección electoral, para los datos del ciudadano. Ya que en una misma sección puede contener diversas casillas.
TIPO_DOMICILIO	Nos permite identificar el tipo de domicilio de la casilla y si tuvo

	influencia en el número de votos que se registraron en la casilla.
MANZANA	Nos permite hacer la relación y cruce con la información demográfica de un origen de datos externo.

Tabla 18.- Exploración Ubica Tu Casilla Proceso Electoral 2011.

Tabla: Datos Geo Electorales 2005

Campo	Tipo de Datos
SECCIÓN	Nos permite hacer la relación y cruce para los datos del ciudadano.
GRA_PROM_ESC	Determinar si el grado promedio de escolaridad está estrictamente relacionado con el porcentaje de la participación ciudadana en una sección electoral.
18YMAS_F	Determinar el porcentaje de mujeres mayores de edad en una sección y definir si es parte de los parámetros influyentes en la participación ciudadana.
18YMAS_M	Determinar el porcentaje de hombres mayores de edad en una sección y definir si es parte de los parámetros influyentes en la participación ciudadana.
18YMAS_EDUC_PB	Determinar si las personas con educación pos básica son los que menos votan en una sección o su relación con otros datos.
CASADOS	Determinar que tanto influye el estado civil en una sección para la participación ciudadana.
VIVHAB_MAT	Determinar si el factor de contar con un piso de material en la vivienda en relevante para determinar si las personas acuden o no a votar.
VIVHAB_SINMAT	Determinar si el factor de NO contar con un piso de material en la vivienda en relevante para determinar si las personas acuden o no a votar.
VIVHAB_PC	Determinar si tiene influencia contar con una computadora en la vivienda para la participación ciudadana.
VIVHAB_TV	Determinar si tiene influencia contar con una televisión en la vivienda para la participación ciudadana.
VIVHAB_AUT	Determinar si tiene influencia contar con un automóvil en la vivienda para la participación ciudadana.

Tabla 19.- Exploración Datos Geo Electorales 2010.

4.2.4. Verificar la calidad de los datos

La calidad de los datos es muy buena, ya que muchos de los datos son limpiados con anterioridad para su uso. Además casi todos los campos cuentan con valores, es decir no pueden contener valores nulos.

Los únicos campos que pueden contener valores nulos son el apellido paterno y apellido materno, en el caso de los ciudadanos. Pero estos datos no son de mucha importancia en nuestro modelo, ya que no afecta a la información que se va a utilizar en la bodega de datos, porque no forma parte ella y además no son de gran relevancia.

4.3. Preparación de los datos

Una vez que hemos realizado la recolección y comprensión de la información contenida en diferentes orígenes de datos se procede a la selección, preparación, limpieza, formato e integración de los datos, para poder ser utilizados en el modelo de minería de datos.

4.3.1. Seleccionar y verificar los datos

Con base a la recolección de los datos realizada con anterioridad y su análisis, procedemos a seleccionar los atributos útiles y de mayor relevancia de nuestros datos, así como la exclusión de aquellos que no sean necesarios para cumplir nuestros objetivos de minería de datos.

4.3.1.1. Razones para la inclusión o exclusión de los datos

Participación Ciudadana

En la información contenida en la tabla tenemos campos que podemos excluir, ya que no son necesarios para nuestro modelo, tales como los datos personales del ciudadano como el NOMBRE, PATERNO, MATERNO y CLAVE_ELEC. Además se excluyen el campo CALLE y ENTIDAD.

Los demás campos se van a conservar ya que representan información muy importante y necesaria para nuestro modelo. Por ejemplo, la edad, sexo, colonia, municipio, distrito, sección, casilla, voto. La Tabla 20 muestra la inclusión y exclusión mencionada.

Campo	Estatus
PATERNO	EXCLUIDO
MATERNO	EXCLUIDO
NOMBRE	EXCLUIDO
EDAD	INCLUIDO
SEXO	INCLUIDO
CALLE	EXCLUIDO
COLONIA	INCLUIDO
ENTIDAD	EXCLUIDO
MUNICIPIO	INCLUIDO
DISTRITO	INCLUIDO
SECCIÓN	INCLUIDO
CASILLA	INCLUIDO
CLAVE_ELEC	EXCLUIDO
VOTÓ	INCLUIDO

Tabla 20.- Inclusión/Exclusión Participación Ciudadana.

Ubica Tu Casilla

En esta tabla también tenemos campos a excluir, tales como ENTIDAD, MUNICIPIO, DISTRITO y DOMICILIO.

Los campos que se van a conservar SECCIÓN, CASILLA, TIPO_DOMICILIO y MANZANA. La Tabla 21 muestra dicha inclusión y exclusión.

Campo	Estatus
ENTIDAD	EXCLUIDO
MUNICIPIO	EXCLUIDO
DISTRITO	EXCLUIDO
SECCIÓN	INCLUIDO
CASILLA	INCLUIDO

TIPO_DOMICILIO	INCLUIDO
DOMICILIO	EXCLUIDO
MANZANA	INCLUIDO

Tabla 21.- Inclusión/Exclusión Ubica Tu Casilla.

Datos Geo Electorales

En la información contenida en esta tabla tenemos campos que deben ser excluidos, tales como: ENTIDAD, MUNICIPIO, DISTRITO_FEDERAL, 18YMAS, 18YMAS_EDUC_PB_F, 18YMAS_EDUC_PB_M, VIVHAB_DREN y TOTAL_VIVIENDAS.

Los campos que se van a conservar, ya que representan información muy importante y necesaria para nuestro modelo son: la SECCIÓN, GRA_PROM_ESC, 18YMAS_F, 18YMAS_M, 18YMAS_EDUC_PB, CASADOS, VIVHAB_MAT, VIVHAB_SINMAT, VIVHAB_PC, VIVHAB_TV y VIVHAB_AUT. La Tabla 22 muestra la inclusión y exclusión mencionada.

Campo	Estatus
ENTIDAD	EXCLUIDO
MUNICIPIO	EXCLUIDO
DISTRITO_FEDERAL	EXCLUIDO
SECCIÓN	INCLUIDO
GRA_PROM_ESC	INCLUIDO
18YMAS	EXCLUIDO
18YMAS_F	INCLUIDO
18YMAS_M	INCLUIDO
18YMAS_EDUC_PB	INCLUIDO
18YMAS_EDUC_PB_F	EXCLUIDO
18YMAS_EDUC_PB_M	EXCLUIDO
CASADOS	INCLUIDO
VIVHAB_MAT	INCLUIDO
VIVHAB_SINMAT	INCLUIDO

VIVHAB_PC	INCLUIDO
VIVHAB_DREN	EXCLUIDO
VIVHAB_TV	INCLUIDO
VIVHAB_AUT	INCLUIDO
TOTAL_VIVIENDAS	EXCLUIDO

Tabla 22.- Inclusión/Exclusión Datos Geo Electorales.

4.3.2. Limpiar los datos

Debido a que la calidad de los datos es muy buena y todos están debidamente estructurados, no se requiere llevar a cabo una limpieza de los datos. Todos los campos que contiene cada una de las tablas cuentan con la información necesaria, no cuentan con valores nulos o incompletos, no presentan incongruencia o valores incorrectos.

4.3.3. Integración de los datos

Como nuestra información está contenida en diferentes orígenes de datos y/o tablas, es necesario realizar una integración de los datos.

En primer lugar vamos a realizar la integración de los datos entre la tabla de Participación Ciudadana y la tabla Ubica Tu Casilla. La integración será posible mediante los campos en común: “sección” y “casilla”. De este modo podemos integrar los datos del tipo de casilla y la manzana a la que pertenece, como se muestra en las tablas de la Figura 3 y el resultado de la integración se muestra en la Tabla 23.

Participación Ciudadana

Campo	Datos de Ejemplo	
EDAD	35	23
SEXO	H	M
COLONIA	CENTRO	PEDREGAL
MUNICIPIO	1	1
DISTRITO	1	1
SECCIÓN	121	122
CASILLA	B	C
VOTÓ	SI	NO

Ubica Tu Casilla

Campo	Datos de Ejemplo	
SECCIÓN	123	124
CASILLA	B	C
TIPO_DOMICILIO	PARTICULAR	ESCUELA
MANZANA	23	56

Figura 3.- Campos en común en las tablas de Participación Ciudadana y Ubica Tu Casilla.

Campo	Datos de Ejemplo	
SECCIÓN	121	122
CASILLA	B	C
TIPO_DOMICILIO	PARTICULAR	ESCUELA
MANZANA	23	56
EDAD	35	23
SEXO	H	M
COLONIA	CENTRO	PEDREGAL
MUNICIPIO	1	1
DISTRITO	1	1
VOTÓ	SI	NO

Tabla 23.- Integración de la tabla Participación Ciudadana y Ubica Tu Casilla.

Posteriormente se lleva a cabo la integración de la tabla generada en la primer integración con el tercer origen de datos seleccionado para nuestro modelo, el cual se refiere a los datos geo electorales. En este caso en particular la integración se realiza mediante el campo en común “sección”. El cual está contenido en ambos orígenes, como se muestra en la Figura 4. El resultado de la integración se muestra en la Tabla 24.

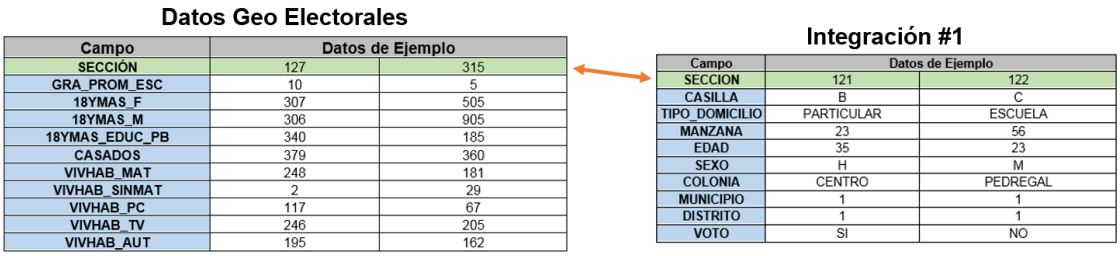

Figura 4.- Campos en común en las tablas de Integración #1 y Datos Geo Electorales.

Campo	Datos de Ejemplo	
SECCIÓN	121	122
CASILLA	B	C
TIPO_DOMICILIO	PARTICULAR	ESCUELA
MANZANA	23	56
EDAD	35	23
SEXO	H	M
COLONIA	CENTRO	PEDREGAL
MUNICIPIO	1	1

DISTRITO	1	1
VOTÓ	SI	NO
GRA_PROM_ESC	10	5
18YMAS_F	307	505
18YMAS_M	306	905
18YMAS_EDUC_PB	340	185
CASADOS	379	360
VIVHAB_MAT	248	181
VIVHAB_SINMAT	2	29
VIVHAB_PC	117	67
VIVHAB_TV	246	205
VIVHAB_AUT	195	162

Tabla 24.- Tabla final generada mediante el proceso de integración de los datos.

4.3.4. Dar formato a los datos

Una vez que se realiza la integración de los datos, se debe analizar el origen final, creado a partir de las tablas de participación ciudadana, ubica tu casilla y datos geo electorales. Realizar los ajustes correspondientes, ya que la etapa siguiente es el modelado, donde los datos integrados serán la entrada de nuestro modelo de minería de datos.

Con base al análisis realizado a los datos integrados a la bodega de datos denominada “ParticipaciónCiudadana” se observa que algunos datos deben ser transformados mediante el proceso ETL (Extracción, Transformación y Carga), para poder realizar un análisis más preciso en nuestro modelo y, con base a la transformación, generar la tabla de hechos y el modelo multidimensional. La Tabla 25 muestra los datos de la participación ciudadana.

id	municipio	distrito	seccion	casilla	edad	sexo	colonia	voto	escolaridad	mujeres	hombres	posbasica	casados	piso	sinpiso	vivpc	vivtv	vivau
1	3	1	135	C1	56	H	BARR LAS LADRILLERAS 2	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
2	3	1	135	C1	42	M	COL LAS LADRILLERAS 23(	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
3	3	1	135	C1	43	M	COL LA LADRILLERA 2302(	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
4	3	1	135	C1	53	M	COL LADRILLERA 23020	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
5	3	1	135	C1	39	M	COL LADRILLERA 23020	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
6	3	1	135	C1	39	M	COL LADRILLERA 23020	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
7	3	1	135	C1	49	H	COL LA LADRILLERA 2302(	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
8	3	1	135	C1	37	M	BARR LAS LADRILLERAS 2	NO	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
9	3	1	135	C1	19	M	COL LAS LADRILLERAS 23(	SI	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
10	3	1	135	C1	20	M	BARR LAS LADRILLERAS 2	SI	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
11	3	1	135	C1	69	M	COL COLINA DE LA CRUZ (	SI	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
12	3	1	135	C1	61	H	COL COLINA DE LA CRUZ (	SI	8.85	46.92	53.08	38.15	54.21	96.88	3.13	36.88	98.44	65.94
13	3	1	136	B	87	M	COL COLINA DE LA CRUZ (	SI	9.53	49.52	50.48	43.71	57.12	97.93	2.07	38.62	97.24	68.28
14	3	1	136	B	35	M	BARR COLINA DE LA CRU(	NO	9.53	49.52	50.48	43.71	57.12	97.93	2.07	38.62	97.24	68.28
15	3	1	136	B	63	M	COL COLINA DE LA CRUZ (	SI	9.53	49.52	50.48	43.71	57.12	97.93	2.07	38.62	97.24	68.28

Tabla 25.- Tabla “ParticipaciónCiudadana” antes del proceso ETL.

Proceso de Extracción, Transformación y Carga.

Los campos que fueron sometidos al proceso de transformación son: escolaridad, posbasica, casados, vivpc, vivau y se agregó un nuevo campo llamado edadr.

- **ESCOLARIDAD.-** En el caso del campo ESCOLARIDAD, que muestra los grados cursados, se hizo una transformación de los valores para clasificarlos por Primaria (P1), Secundaria (S), Preparatoria (P2) y Universidad o posteriores (U).

  Para su clasificación se utilizaron las siguientes condiciones:

```
IF escolaridad<7
       escolaridad='P1'
ELSE IF escolaridad<10
       escolaridad='S'
ELSE IF escolaridad<13
       esclaridad='P2'
ELSE
       escolaridad='U'
```

- **POSBASICA.-** En el caso de este campo, que muestra las personas con educación posbasica, es decir con secundaria o mayor, se hizo una transformación de los valores para clasificarlos por Casi ninguno (CN), Menos de la mitad (ME), Mas de la mitad (MA) y Casi todos (CT).

  Para su clasificación se utilizaron las siguientes condiciones:

```
IF posbasica<25
       posbasica ='CN'
ELSE IF posbasica <50
       posbasica ='ME'
ELSE IF posbasica <75
       posbasica ='MA'
ELSE
       posbasica ='CT'
```

- **CASADOS.-** En el caso de este campo, que muestra las personas con estado civil casados, se hizo una transformación de los valores para clasificarlos por Menos de la mitad (ME) y Más de la mitad (MA).

  Para su clasificación se utilizaron las siguientes condiciones:

```
IF casados<50
       casados ='ME'
ELSE
       casados ='MA'
```

- **VIVPC.-** En el caso de este campo, que muestra las viviendas con computadora, se hizo una transformación de los valores para clasificarlos por Casi ninguno (CN), Menos de la mitad (ME), Mas de la mitad (MA) y Casi todos (CT).

  Para su clasificación se utilizaron las siguientes condiciones:

```
IF vivpc<25
       vivpc ='CN'
ELSE IF vivpc <50
       vivpc ='ME'
ELSE IF vivpc <75
       vivpc ='MA'
ELSE
       vivpc ='CT'
```

- **VIVAU.-** En el caso de este campo, que muestra las viviendas con automóvil, se hizo una transformación de los valores para clasificarlos por Casi ninguno (CN), Menos de la mitad (ME), Mas de la mitad (MA) y Casi todos (CT).

Para su clasificación se utilizaron las siguientes condiciones:

```
IF vivau<25
       vivau ='CN'
ELSE IF vivau <50
       vivau ='ME'
ELSE IF vivau <75
       vivau ='MA'
ELSE
       vivau ='CT'
```

- **EDADR.-** Este campo se va a generar en base a la edad de las personas, el cual nos va a permitir clasificar a las mismas por rango de edad. Los rangos creados son: Joven (JO), Adulto (AD) y Tercera Edad (3A).

  Para su clasificación se utilizaron las siguientes condiciones:

```
IF edad<30
       edadr ='JO'
ELSE IF edad <60
       edadr ='AD'
ELSE
       edadr ='3E'
```

Tabla de hechos

Después de haber implementado el proceso ya contamos con los datos requeridos para nuestro análisis, ya con un formato sin valores nulos o erróneos, y se creó una tabla de hechos. Esta tabla sirve como origen de datos y contiene los siguientes campos: ID, MUNICIPIO, DISTRITO, SECCIÓN, CASILLA, EDAD, SEXO, COLONIA, VOTÓ, ESCOLARIDAD, MUJERES, HOMBRES, POSBÁSICA, CASADOS, PISO, SINPISO, VIVPC, VIVTV, VIVAU y EDADR. La Tabla 26 muestra una parte de la tabla de hechos generada:

id	municipi	distrit	seccio	casill	edad	sexo	colonia	voto	esc	muje	hombr	posbas	casado	piso	sinpiso	vivpc	vivtv	vivau	edadr
4924	3	1	138	B	31	M	COL VICENTE GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4925	3	1	138	B	51	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4926	3	1	138	B	49	H	COL GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4927	3	1	138	B	20	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4928	3	1	138	B	24	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4929	3	1	138	B	62	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	3E
4930	3	1	138	B	64	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	3E
4931	3	1	138	B	31	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4932	3	1	138	B	28	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4933	3	1	138	B	49	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4934	3	1	138	B	39	H	COL GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4935	3	1	138	B	35	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4936	3	1	138	B	48	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4937	3	1	138	B	44	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4938	3	1	138	B	34	M	COL VICENTE GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4939	3	1	138	B	37	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4940	3	1	138	B	27	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4941	3	1	138	B	30	H	COL VICENTE GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4942	3	1	138	B	63	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	3E
4943	3	1	138	B	36	H	COL GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4944	3	1	138	B	37	H	COL GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4945	3	1	138	B	31	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4946	3	1	138	B	24	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4947	3	1	138	B	19	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4948	3	1	138	B	49	H	COL VICENTE GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4949	3	1	138	B	52	H	COL GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4950	3	1	138	B	55	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4951	3	1	138	B	26	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4952	3	1	138	B	21	M	COL COLINA DE LA CRUZ 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4953	3	1	138	B	29	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4954	3	1	138	B	28	H	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	JO
4955	3	1	138	B	58	M	COL VICENTE GUERRERO 23020	SI	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	AD
4956	3	1	138	B	69	M	COL VICENTE GUERRERO 23020	NO	S	50	50	ME	MA	99.54954954	0.450450450	ME	96.396396396396398	MA	3E

Tabla 26.- Tabla de hechos denominada “Hechos_ParticipaciónCludadana”.

Generación del modelo multidimensional

Una ventaja que nos brinda el generar un modelo multidimensional, para el caso de nuestro estudio, es que permite concentrar los datos de diferentes tablas u orígenes de datos en una sola tabla, que normalmente es la tabla de hechos. Además permite llevar a cabo un mejor análisis de la información, ya que los algoritmos de minería de datos que vamos a utilizar generan los resultados en un menor tiempo y con un mayor grado de efectividad.

También nos permite la definición de dimensiones, con las cuales se pueden generar vistas o perspectivas de los datos contenidos en la tabla de hecho de diferentes medidas o hechos.

El modelo que se creó para nuestro objeto de estudio es un modelo multidimensional de tipo Copo de Nieve. El cual cuenta con 3 dimensiones principales, las cuales son: Personas, Sección y Población.

La dimensión Personas cuenta con los datos personales de los ciudadanos que acudieron a votar y los que no. Por ejemplo, si votó, la edad, sexo, rango de edad

y escolaridad. A su vez esta dimensión cuenta con las dimensiones Edad, que muestra la edad de las personas y su rango de edad al que pertenecen los ciudadanos. De la dimensión Edad se desprende la dimensión de RangoEdad, que muestra el rango de edad a la que pertenecen las personas, Jóvenes (JO), Adultos (AD) y 3ra Edad (3E). Además cuenta con una dimensión Escolaridad, la cual se refiere al grado de estudios de las personas, Primaria (P1), Secundaria (S), Preparatoria (P2) y Universidad o posteriores (U). Estas dimensiones son necesarias ya que con base a este modelo de dimensiones, podemos visualizar información de diferentes aspectos, por ejemplo las edades de las personas que votaron. También podemos consultar el número de personas que votaron por su rango de edad, y con ello identificar el rango de edad que más voto en una elección. Incluso si queremos un nivel más de detalle, por ejemplo, saber la edad de las personas que más votaron.

La dimensión Sección nos permite agrupar a los ciudadanos por distintas secciones electorales, distrito al que pertenecen y el municipio al que pertenece el distrito y las secciones que conforman un distrito. También cuenta con las dimensiones Distrito, la cual nos permite mostrar a los ciudadanos de un distrito y poder obtener información como por ejemplo el distrito que más votó y a su vez identificar cual sección fue la que más voto de ese mismo distrito. De igual forma cuenta con la dimensión Municipio, con la que podemos mostrar los ciudadanos que votaron por municipio, identificar qué municipio fue el que más votó y de ese municipio, cual fue el distrito que más votó.

La dimensión Población agrupa la información de datos estadísticos poblacionales de los ciudadanos que acudieron a votar. Cuenta con la dimensión Auto, la cual nos permite conocer si las personas que votan cuentan con al menos un automóvil. La dimensión PC nos permite conocer si las personas cuentan con un equipo de cómputo. La dimensión TV nos muestra información relativa a si las personas cuentan con una televisión en su vivienda.

El modelo multidimensional fue creado dentro de la herramienta “SQL Server Data Tools 2012” de Microsoft y lo denominamos “ParticipaciónCiudadana”.

En la Figura 5 se muestra un diagrama del modelo multidimensional, en la Figura 6 y 7 ejemplos de la herramienta SQL Server Data Tools.

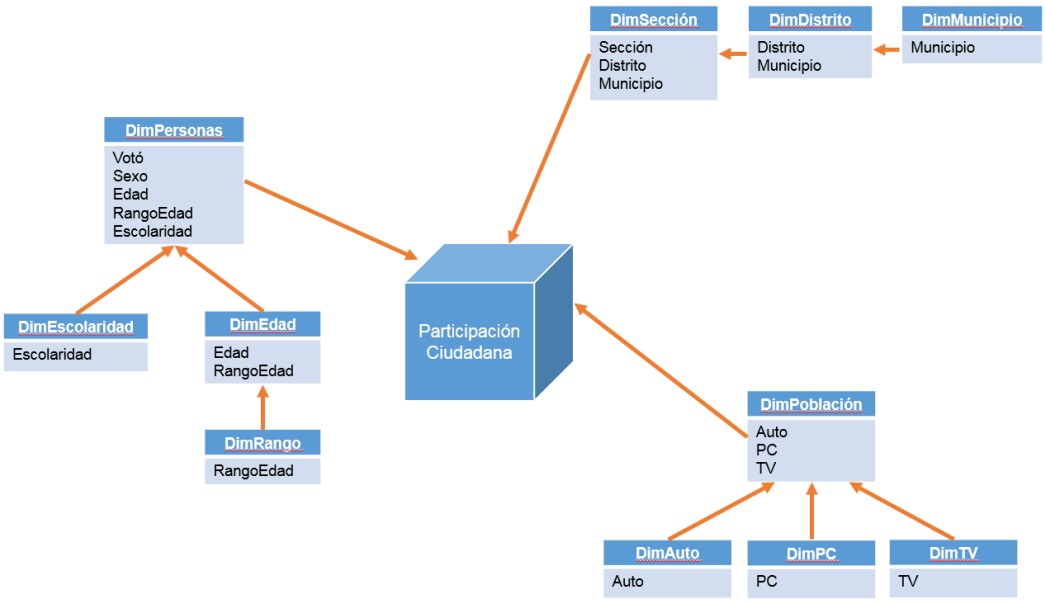

Figura 5.- Modelo multidimensional denominado “Participación Ciudadana”.

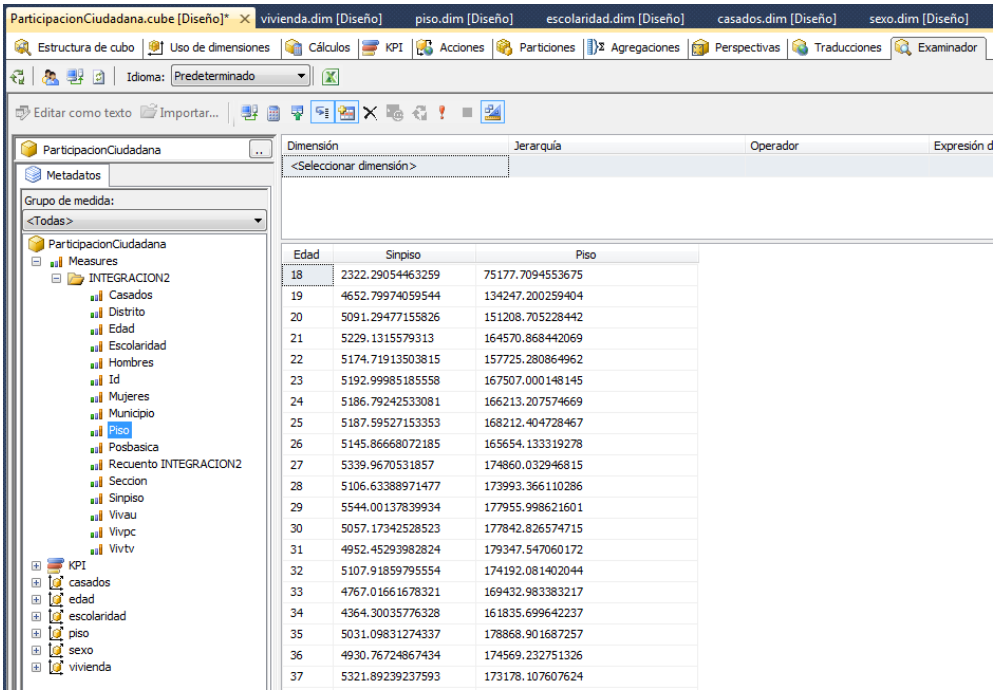

Figura 6.- Examinador del cubo usando las medidas y dimensiones en SQL Server Data Tools 2012.

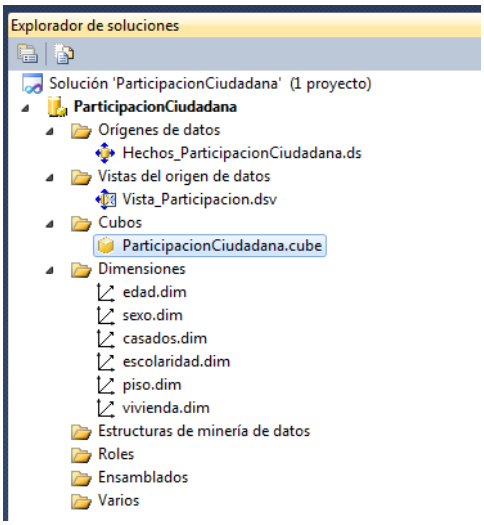

Figura 7.- Ejemplo de los elementos del modelo multidimensional en SQL Server Data Tools 2012.

4.4. Modelado

4.4.1. Selección de la técnica de modelado

Con base al análisis y al modelo multidimensional, creado en la etapa de Preparación de los Datos (y tomando en cuenta los objetivos que se desean cumplir con el modelo de minería de datos), se procede a seleccionar las técnicas de minería de datos que vamos a utilizar en nuestro modelo, las cuales son la aplicación de algoritmos para la minería de los datos. En concreto se aplicaron el algoritmo de Clasificación y Algoritmo de Clúster.

Se decidió por la aplicación de estos algoritmos debido a que son las técnicas que nos permiten analizar información histórica (en este caso de los procesos electorales pasados) y con base a ese análisis poder extraer información y/o conocimiento que no se puede apreciar fácilmente, tales como los factores de influencia en la participación ciudadana y patrones de comportamiento de las personas que acuden a votar.

4.4.1.1. Técnica de modelado

1.- Algoritmo de Clasificación

Es un algoritmo que permite clasificar los datos con base a ciertos criterios y atributos. Nos permite identificar los atributos de mayor relevancia o influencia sobre algún campo en particular que previamente seleccionamos para su análisis.

Para nuestro modelo en particular se utilizó el método de clasificación Arboles de Decisión.

Árbol de Decisión

Este método hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, identifica las columnas de entrada que se correlacionan con la columna de predicción.

Por lo tanto vamos a aplicar este método para identificar los campos que se relacionan con el campo seleccionado para predicción: VOTÓ, es decir, identificar los valores o parámetros que influyen en que la persona vote o no. En pocas palabras los factores tanto personales como sociales que influyen en la participación ciudadana.

2.- Algoritmo de Clúster

Es un algoritmo que nos permite la segmentación de los datos con base a ciertos criterios y poder obtener datos similares o con características parecidas y con base a ello generar clústeres o grupos de datos. Utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares.

Para nuestro modelo en particular se utilizó el método de agrupamiento K-Means.

K-Means

Es un método de agrupamiento que tiene como objetivo la partición de un conjunto de N observaciones en K grupos, en el que cada observación pertenece al grupo más cercano a la media. Tiene por objeto encontrar grupos de observaciones (clústeres) con características semejantes. Las observaciones de un grupo deben ser parecidas a las otras observaciones del grupo, pero deben ser lo más diferentes de las observaciones de otros grupos.

Este modelo permite identificar o agrupar a las personas que votaron y que no votaron y que cuentan con características similares. Estas características a identificar serán las columnas de entrada que seleccionaremos. Y, con base a este agrupamiento, poder definir un patrón o patrones de comportamiento en la participación ciudadana.

4.4.2. Generar el diseño de pruebas

Para poder llevar a cabo el análisis de la información y obtener resultados en menor tiempo posible, toda vez que el universo de los datos es muy extenso (hablamos de una bodega de datos con 960,019 registros), se tomó como objeto de estudio solo los datos correspondientes al municipio de La Paz (73,901 registros). Además que particularmente cada municipio es muy diferente uno de otro en cuanto a su nivel de participación ciudadana, así como sus datos estadísticos poblaciones.

4.4.3. Construcción del modelo

Una vez que seleccionamos nuestra técnica de modelado y diseñamos las pruebas para nuestro modelo, procedemos a realizar la construcción del modelo.

Previamente se hizo una selección y preparación de la información contenida en los diferentes orígenes de datos, ya que esta información nos va a servir como datos de entrada para el análisis de la información. Posteriormente se hizo la selección de los algoritmos de minería de datos, en especial algoritmos de

Clasificación y Agrupamiento. Para el algoritmo de clasificación se seleccionó el método de Árbol de Decisión y por el algoritmo de agrupamiento se seleccionó el método K-Means. Después se procede al análisis de la información, esto mediante la implementación y ejecución de los algoritmos de minerías de datos seleccionados. La implementación, ejecución y los resultados obtenidos se describen en la siguiente etapa:

4.4.3.1. Algoritmo de clasificación

Primeramente analizamos nuestra información por medio del algoritmo de clasificación, Árbol de Decisión. Esto con la finalidad de detectar los valores más representativos o de mayor relevancia en las personas que acuden a votar. Es decir, los factores que influyen en las personas para ejercer su voto el día de la elección.

Prueba #1

Como primera prueba seleccionamos la columna a analizar, que en este caso es: VOTO. Después procedimos a seleccionar las columnas de entrada: EDAD, ESCOLARIDAD, POSBASICA, CASADOS, PISO, SINPISO, VIVPC, VIVTV y VIVAU.

Estas columnas de entrada fueron seleccionadas con base al resultado de las etapas de comprensión de los datos y selección de los datos, etapas aplicadas con anterioridad en nuestra investigación y que forman parte de la metodología aplicada, ya que muestran información relevante en cuanto a los ciudadanos que votaron.

Una vez que se realizó el primer análisis con el algoritmo seleccionado y la configuración de sus variables de entrada y la variable de análisis, se generó el árbol de decisión resultante, que se muestra en la Figura 8:

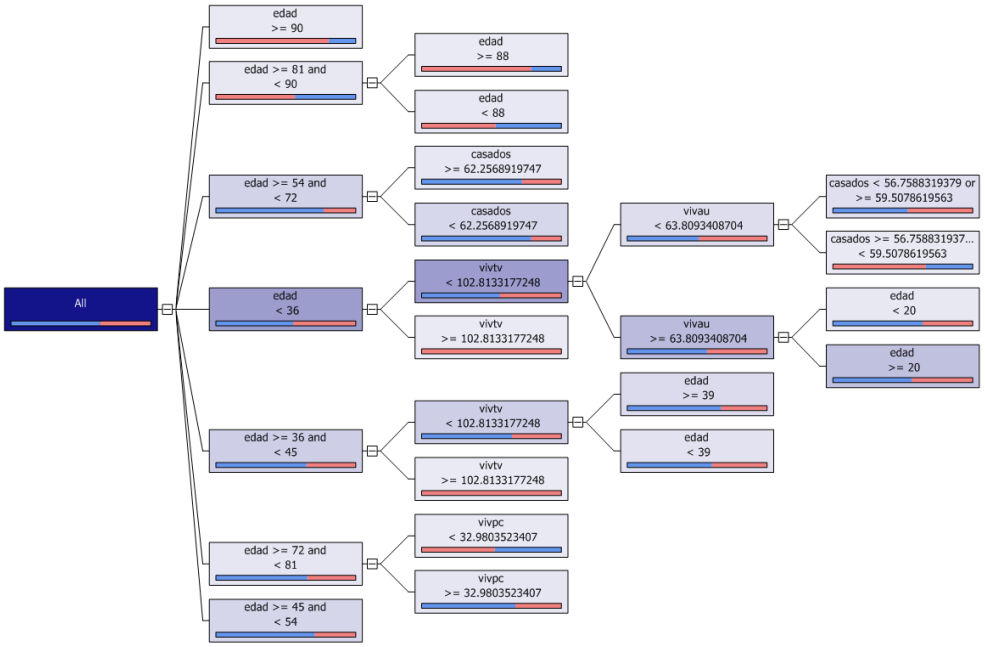

Figura 8.- Árbol resultante de la aplicación del algoritmo de clasificación en la prueba #1.

Como se puede observar en la Figura 8, el árbol resultante que nos arroja el algoritmo una vez que analizó los datos, no proporciona una información relevante y muy clara de las variables más representativas, sino que muestra un árbol con múltiples opciones. Este resultado no es el adecuado para nuestros objetivos, por lo que se debe realizar una nueva ejecución del algoritmo pero considerando los datos de entrada, es decir omitiendo algunos datos de entrada de la primera prueba.

Otros resultados que nos proporciona la aplicación del algoritmo de clasificación es una red de dependencias con base en las variables de entradas. Esta red muestra el nodo central, variable a evaluar (VOTO), y los demás nodos, variables de entradas, como los nodos que predicen el nodo central. Como se muestra en la Figura 9:

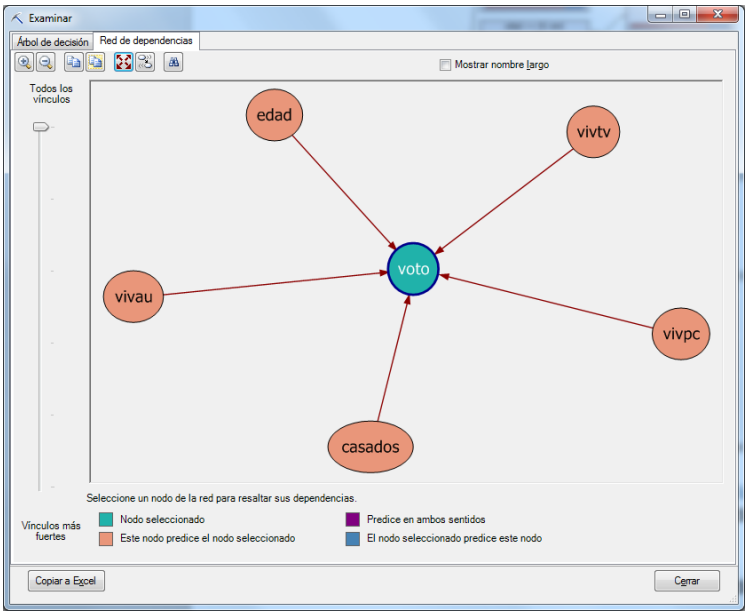

Figura 9.- Red de dependencias en la aplicación del algoritmo de clasificación en la prueba #1.

En la Figura 9 se observa que la red de dependencias da como resultado que las variables más representativas para el caso de la variable de análisis, VOTO, son las variables EDAD, CASADOS, VIVAU, VIVPC y VIVTV. En este caso el número de variables no arroja un resultado muy claro y concreto para nuestros objetivos, sin embargo nos brinda al menos una idea de cuales variables tomar en cuenta para la siguiente prueba.

Resultados Prueba #1

En la Tabla 27 se detallan los parámetros establecidos en la prueba 1, resultados obtenidos y las conclusiones.

Método de clasificación	Árbol de decisión
Columna de análisis	Voto
Columnas de entrada	Edad, escolaridad, posbasica, casados, piso, sinpiso, vivpc, vivtv y vivau

Tabla 27.- Parámetros del algoritmo establecidos en la prueba #1.

Resultados: La red de dependencias nos da como resultado que las variables más representativas para el caso de la variable de análisis, VOTO, son las variables EDAD, CASADOS, VIVAU, VIVPC y VIVTV.

Conclusiones: Los resultados obtenidos de esta primera prueba nos permitieron realizar una nueva prueba tomando en cuenta las variables más representativas como columnas de entrada para la siguiente prueba.

Ejemplos: La Figura 16 muestra un ejemplo de los resultados de la prueba #1.

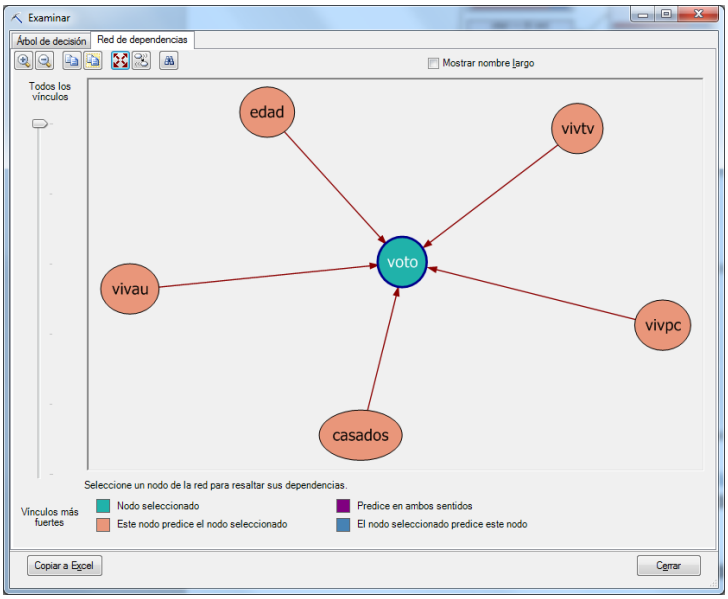

Figura 16.- Variables más representativas en la prueba #1 del algoritmo de clasificación.

Prueba #2

Para la segunda prueba seleccionamos la misma columna a analizar que en la prueba anterior: VOTO. Después procedimos a seleccionar las nuevas columnas de entrada, omitiendo algunas de la primera prueba y tomando en cuenta también la red de dependencias de la primera evaluación. Las columnas de entrada seleccionadas para esta prueba fueron: EDADR, ESCOLARIDAD, POSBASICA, VIVPC, VIVTV y VIVAU.

Estas columnas de entrada fueron seleccionadas con base al resultado de la prueba realizada con anterioridad, ya que muestran información relevante en cuanto a los ciudadanos que votaron.

Una vez que se realizó el segundo análisis con el algoritmo seleccionado, con la modificación de las variables de entrada y conservando la misma variable de análisis, se generó un nuevo árbol de decisión resultante como se muestra en la Figura 10:

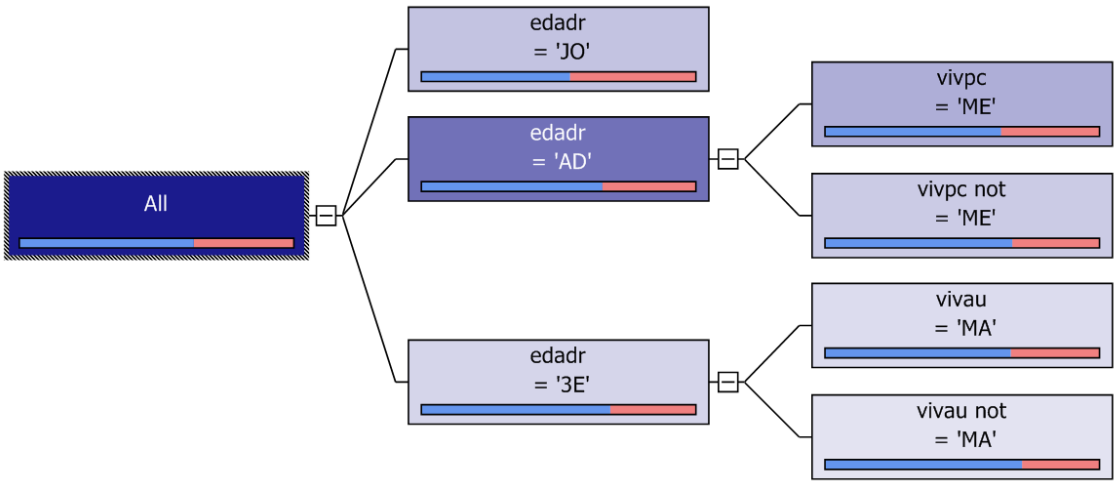

Figura 10.- Árbol resultante de la aplicación del algoritmo de clasificación en la prueba #2.

Como se puede observar en la Figura 10, el árbol resultante que nos arroja el algoritmo una vez que analizó los datos, proporciona información mucho más clara y precisa de las variables más representativas. Incluso se puede observar la variable de mayor influencia y sus variables subsecuentes. Este resultado es favorable para nuestro estudio y nos arroja suficiente información para poder llevar a cabo una tercera prueba.

Como parte de los resultados también obtuvimos la red de dependencias con base a las variables de entradas ya modificadas. Como se muestra en la Figura 11:

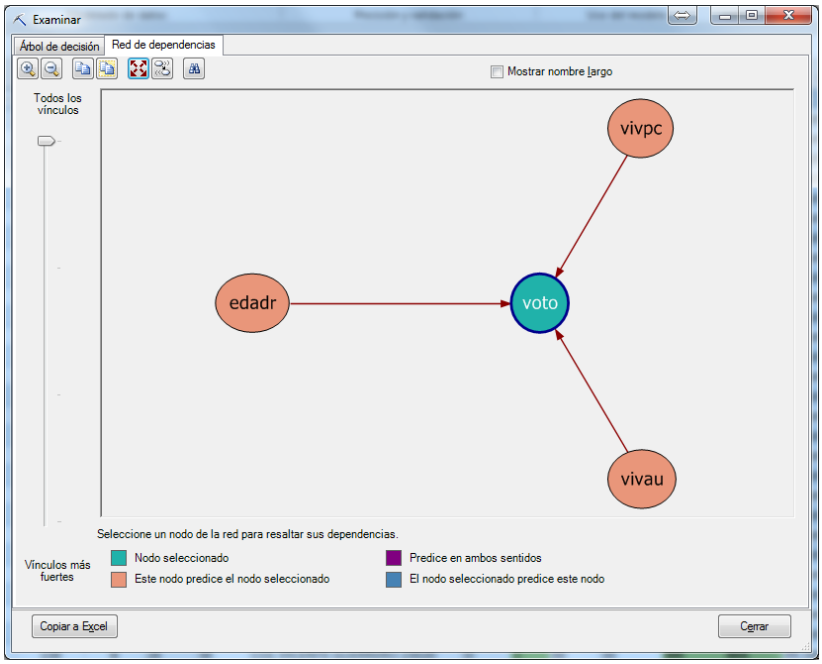

Figura 11.- Red de dependencias en la aplicación del algoritmo de clasificación en la prueba #2.

En la Figura 11, se observa que la red de dependencias nos da como resultado que las variables más representativas para el caso de la variable de análisis, VOTO, son las variables EDADR, VIVAU y VIVPC. En este caso las variables resultantes si ayudan para el cumplimiento de los objetivos de estudio. Además sirve de soporte para llevar a cabo una tercera prueba y obtener datos aún más precisos.

Resultados Prueba #2

En la Tabla 28 se detallan los parámetros establecidos en la prueba 2, resultados obtenidos y las conclusiones.

Método de clasificación	Árbol de decisión
Columna de análisis	Voto
Columnas de entrada	Edadr, escolaridad, posbasica, vivpc, vivtv y vivau

Tabla 28.- Parámetros del algoritmo establecidos en la prueba #2.

Resultados: La red de dependencias nos da como resultado que las variables más representativas para el caso de la variable de análisis, VOTO, son las variables EDADR, VIVAU y VIVPC.

Conclusiones: Los resultados obtenidos en esta segunda prueba ya nos permiten realizar una conclusión sobre la información analizada, sin embargo se realizó una nueva prueba con las variables más representativas como columnas de entrada para obtener resultados aún más precisos.

Ejemplos: La Figura 17 muestra un ejemplo de los resultados de la prueba #2.

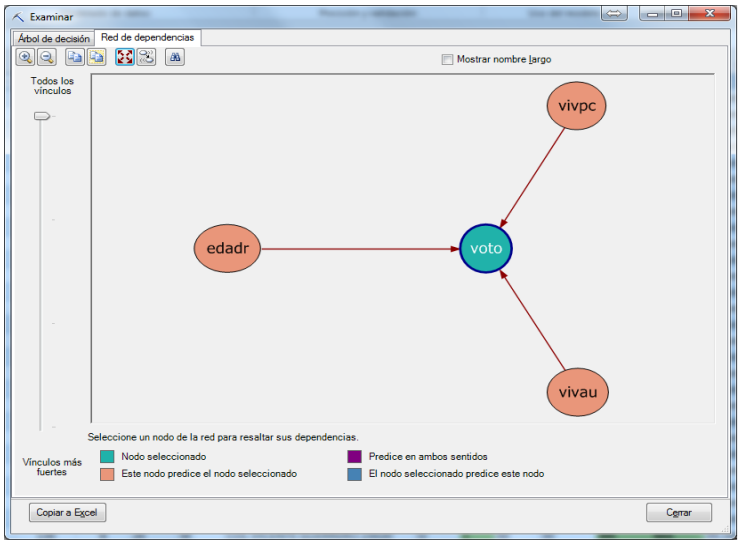

Figura 17.- Variables más representativas en la prueba #2 del algoritmo de clasificación.

Prueba #3

Para la tercera prueba seleccionamos la misma columna a analizar que en las pruebas anteriores: VOTO. Después procedimos a seleccionar las nuevas columnas de entrada, omitiendo algunas de la segunda prueba y tomando en cuenta también la red de dependencias de la segunda evaluación. Las columnas de entrada seleccionadas para esta prueba fueron: EDADR, VIVPC y VIVAU.

Estas columnas de entrada fueron seleccionadas con base al resultado de la segunda prueba realizada con anterioridad, ya que muestran información relevante en cuanto a los ciudadanos que votaron.

Una vez que se realizó el tercer análisis con el algoritmo seleccionado, con la modificación de las variables de entrada y conservando la misma variable de análisis, se generó un nuevo árbol de decisión resultante como se muestra en la Figura 12:

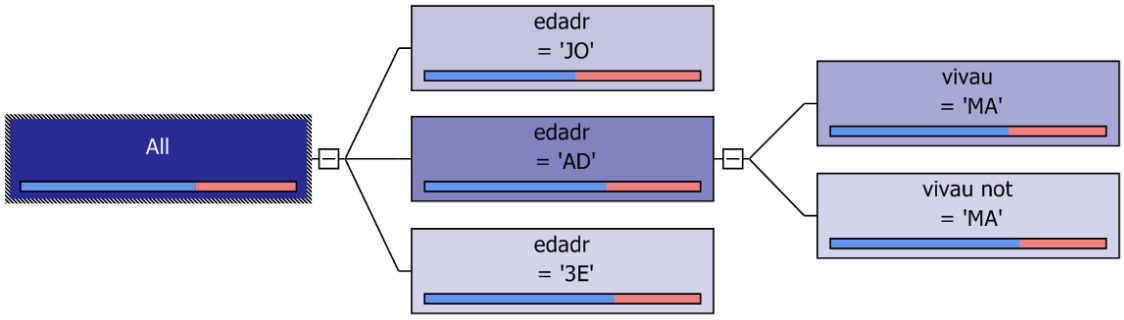

Figura 12.- Árbol resultante de la aplicación del algoritmo de clasificación en la prueba #3.

Como se puede observar en la Figura 12, el árbol resultante que nos arroja el algoritmo proporciona la información adecuada de las variables más representativas e incluso se cumplen los resultados deseados en la aplicación del algoritmo de clasificación. Con estos resultados ya podemos tomar una conclusión e interpretación clara y precisa del análisis de la información.

Como parte de los resultados también obtuvimos la red de dependencias, la cual nos dio como resultado que las variables más representativas son EDADR y VIVAU. A su vez nos indica que la variable de mayor influencia en cuanto al campo de análisis, VOTO, es principalmente EDADR y posteriormente VIVAU. En la Figura 13 se muestra la red de dependencias resultante:

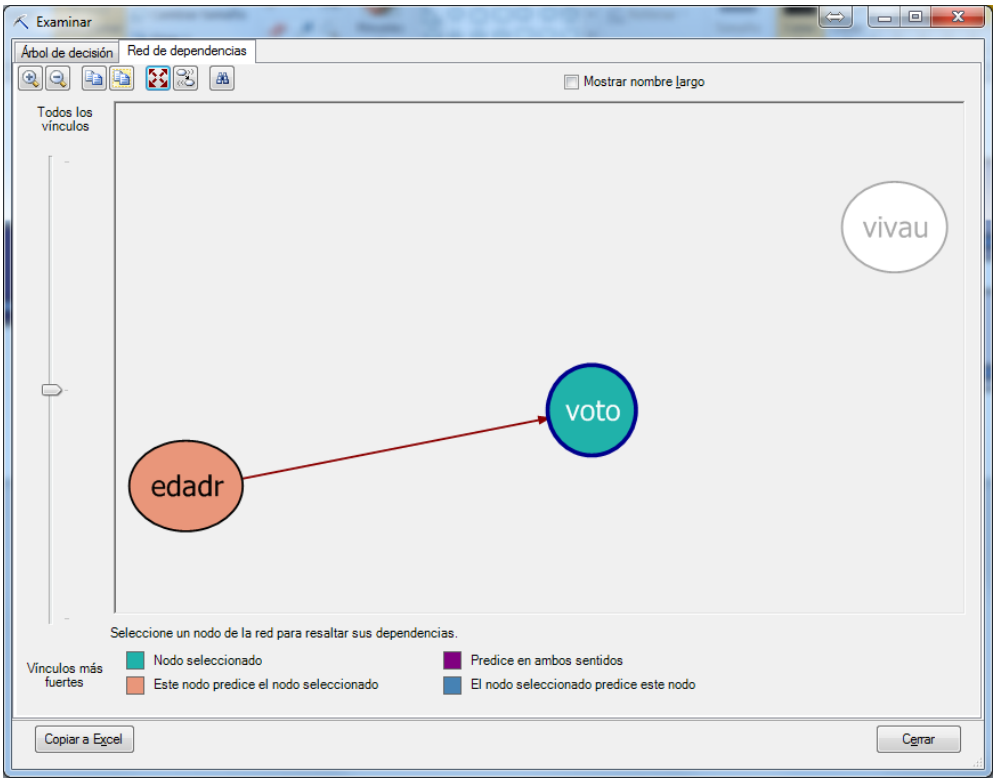

Figura 13.- Red de dependencias en la aplicación del algoritmo de clasificación en la prueba #3.

Después de haber realizado las pruebas necesarias con el algoritmo de clasificación y toda vez que obtuvimos la información deseada, continuamos con la ejecución y análisis del siguiente algoritmo, en este caso el de agrupamiento, que seleccionamos como parte del modelado.

Resultados Prueba #3

En la Tabla 29 se detallan los parámetros establecidos en la prueba 3, resultados obtenidos y las conclusiones.

Método de clasificación	Árbol de decisión
Columna de análisis	Voto
Columnas de entrada	Edadr, vivpc y vivau

Tabla 29.- Parámetros del algoritmo establecidos en la prueba #3.

Resultados: La red de dependencias nos da como resultado que las variables más representativas para el caso de la variable de análisis, VOTO, son las variables EDADR y VIVAU. Aunque en este caso nos especifica que la variable de

mayor influencia en los votantes es la EDADR, y como segundo parámetro es VIVAU.

Conclusiones: Los resultados obtenidos en esta tercera prueba nos permiten establecer las variables de mayor influencia en la participación ciudadana así como su orden jerárquico. Parte de los resultados nos dice que la variable de mayor influencia en los votantes es la EDADR, es decir su rango de edad. Como segunda variable VIVAU, es decir si la persona cuenta con automóvil en su vivienda. También como tercera variable influyente VIVPC, la cual nos indica si la persona cuenta con computadora en su vivienda.

Ejemplos: La Figura 18 muestra un ejemplo de los resultados de la prueba #3.

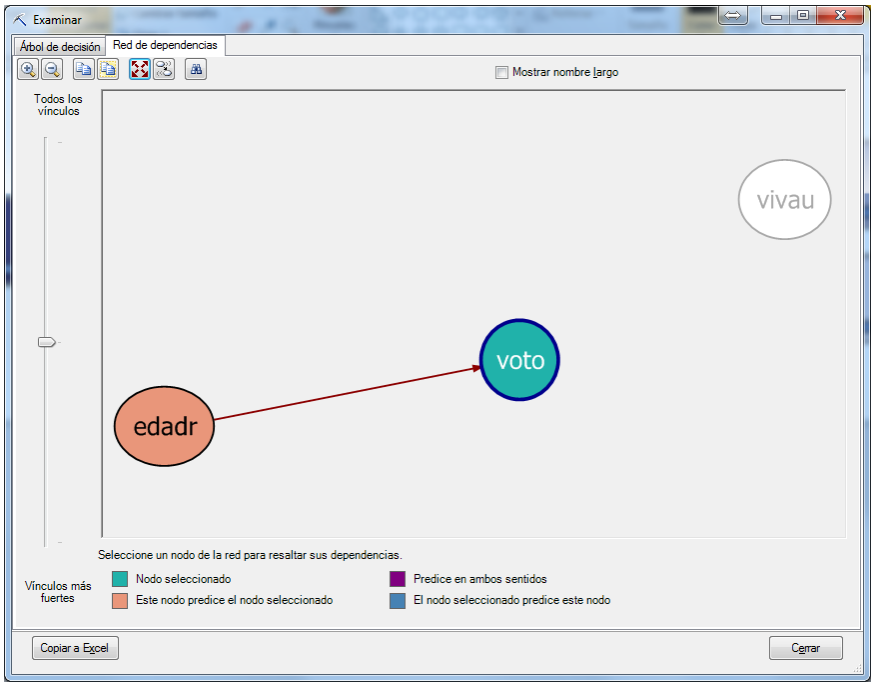

Figura 18.- Variables más representativas en la prueba #3 del algoritmo de clasificación.

Cabe señalar que hasta este momento solo contamos con las variables de mayor influencia en la participación ciudadana y su orden jerárquico. Sin embargo aún no hemos descubierto cuales son los valores específicos que toman esas variables y así poder identificar los patrones de comportamiento.

4.4.3.2. Algoritmo de clúster

En esta etapa analizamos nuestra información por medio del algoritmo de clúster o agrupamiento, utilizando el método de agrupación K-Means. Esto con la finalidad de agrupar los votantes con características similares. Con esto podremos identificar y/o establecer los patrones de comportamiento.

Prueba #1

Como primera prueba seleccionamos las columnas de entrada, las cuales fueron: EDAD, ESCOLARIDAD, VOTO, CASADOS, PISO, SINPISO, VIVPC, VIVTV y VIVAU.

Estas columnas de entrada fueron seleccionadas con base al resultado de las pruebas realizadas con el algoritmo de clasificación, pruebas realizadas con anterioridad en nuestra investigación y que forman parte de la metodología aplicada. Estas columnas fueron seleccionadas porque representan información relevante en cuanto a la participación ciudadana, es decir, tienen influencia en las personas que votan.

Adicionalmente se establecieron algunos parámetros necesarios para la ejecución del análisis con el algoritmo. Para ello seleccionamos el método de agrupamiento (CLUSTERING_METHOD) el cual fue: K-Means Escalable. Así mismo seleccionamos el número de clúster (CLUSTER_COUNT) el cual fue: 7.

Una vez que se realizó el primer análisis con el algoritmo seleccionado y la configuración de sus variables de entrada y parámetros correspondientes, se generó un diagrama de clúster (grupos) como se muestra en la Figura 14:

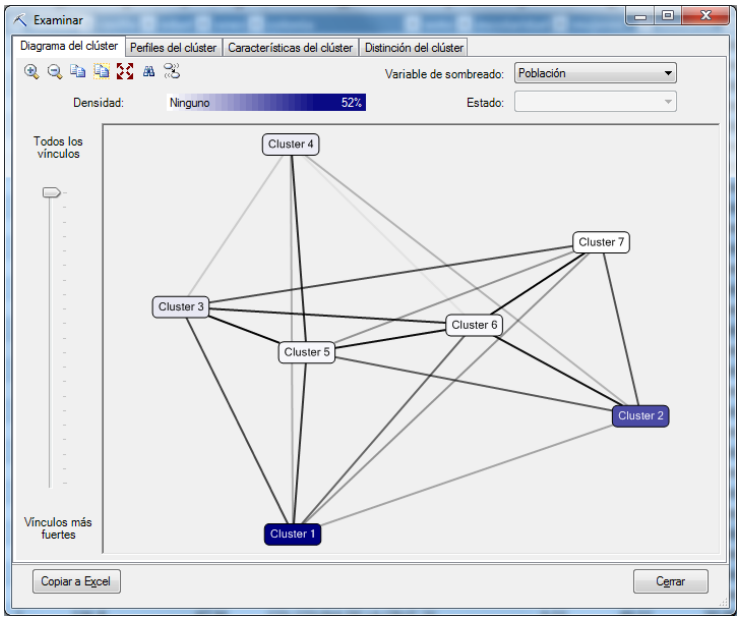

Figura 14.- Diagrama de clúster obtenido al aplicar el algoritmo de agrupamiento K-Means en la prueba #1.

Como se puede apreciar en la Figura 14, el diagrama resultante no muestra una información muy clara de los clústeres generados, ya que sus interconexiones entre clústeres son muy similares en casi todos los casos y no se puede hacer una inferencia precisa del resultado del análisis.

En la Figura 15 se muestra un resumen de los perfiles de clústeres generados en el análisis de la información.

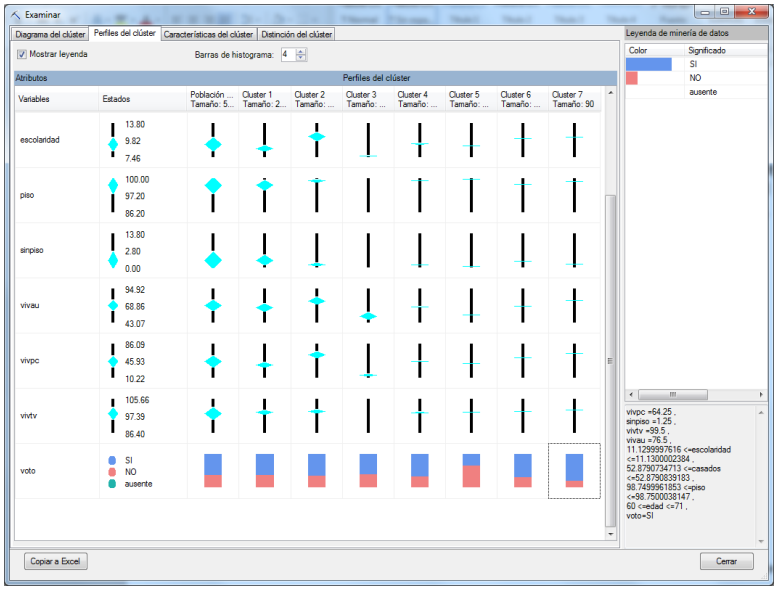

Figura 15.- Perfiles de clúster obtenidos al aplicar el algoritmo de agrupamiento K-Means en la prueba #1.

Como se puede apreciar en la Figura 15, los perfiles de clústeres resultantes no muestran una información muy clara de los clústeres generados, ya que tienen similitudes entre los clústeres y no muestran una tendencia muy marcada en la participación ciudadana para establecer patrones de comportamiento.

Por lo tanto, se procedió a realizar una nueva prueba con datos de entrada y parámetros distintos.

Resultados Prueba #1

En la Tabla 30 se detallan los parámetros establecidos en la prueba 1, resultados obtenidos y las conclusiones.

Método de agrupamiento	K-Means
Número de clúster	7
Columnas de entrada	Edadr, escolaridad, voto, casados, piso, sinpiso, vivpc, vivtv y vivau

Tabla 30.- Parámetros del algoritmo establecidos en la prueba #1.

Resultados: Los clúster generados en esta prueba contaban con similitudes entre algunos clúster, es decir con valores muy similares entre cada uno de los clúster generados. Por lo tanto no se puede establecer una tendencia muy marcada entre los votantes.

Conclusiones: Los resultados obtenidos en esta primera prueba no permiten establecer una tendencia o patrones de comportamiento, sin embargo nos sirven de apoyo para establecer nuevos parámetros en la siguiente prueba.

Ejemplos: La Figura 19 muestra un ejemplo de los resultados de la prueba #1.

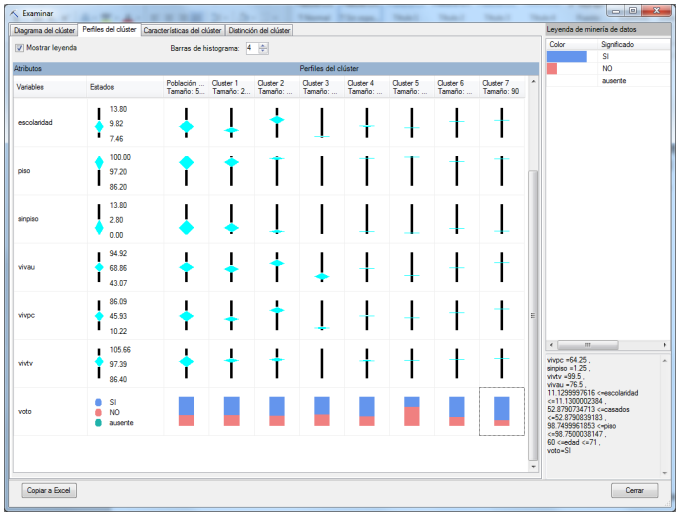

Figura 19.- Clúster obtenidos en la prueba #1 del algoritmo de agrupamiento K-Means.

Prueba #2

Como segunda prueba, se modificaron las columnas de entrada. En especial se omitieron algunas de ellas. Las columnas seleccionadas fueron: ESCOLARIDAD, VOTO, EDADR, VIVPC, VIVTV y VIVAU.

Estas columnas de entrada fueron seleccionadas con base al resultado de la primera prueba realizada con el algoritmo. Las columnas que se omitieron son las que no representaban gran diferencia entre los clústeres de la primera prueba.

Además se volvieron a establecer los parámetros necesarios para la ejecución del análisis. Para ello seleccionamos el mismo método de agrupamiento (CLUSTERING_METHOD) el cual fue: K-Means Escalable, pero en el caso del número de clúster (CLUSTER_COUNT) lo modificamos por un valor menor: 5.

Una vez que se llevó a cabo el análisis con el algoritmo seleccionado, después de haber configurado de nuevo sus variables de entrada y parámetros correspondientes, se generó un nuevo diagrama de clúster (grupos) como se muestra en la Figura 16:

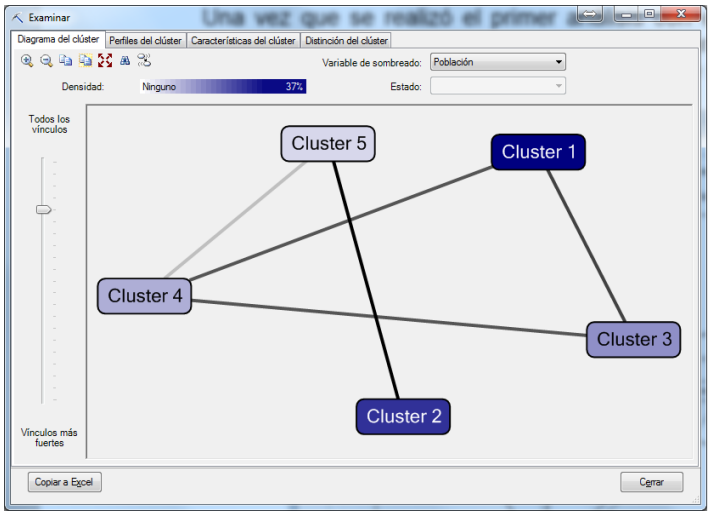

Figura 16.- Diagrama de clúster obtenido al aplicar el algoritmo de agrupamiento K-Means en la prueba #2.

Como se puede apreciar en la Figura 16, el diagrama resultante muestra información un poco más clara de los clústeres generados, pero observamos que los clústeres 1 y 2 tienen densidad muy similar, al igual que los clústeres 3 y 4. Por lo tanto todavía no podemos llegar a un conclusión concreta.

En la Figura 17 tenemos el resumen de los perfiles de los clústeres generados en el análisis de la información.

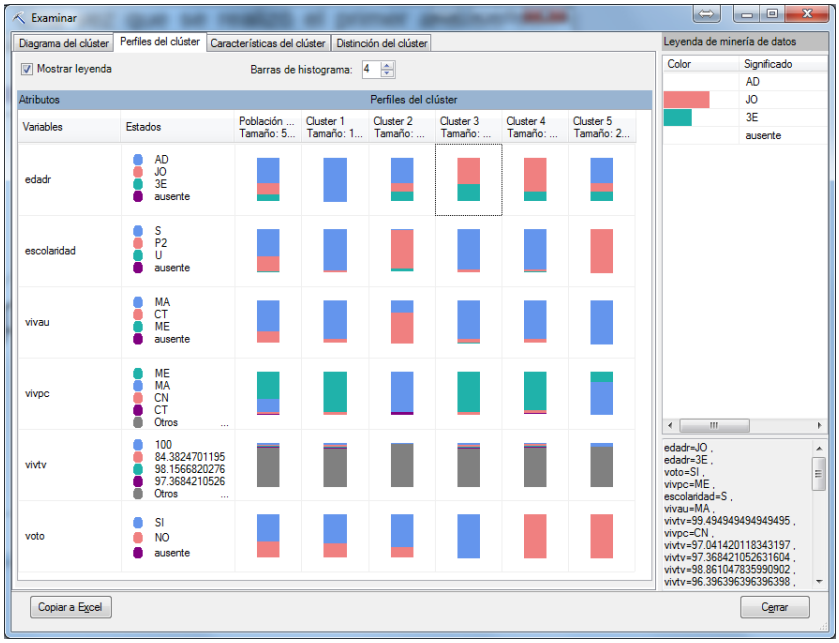

Figura 17.- Perfiles de clúster obtenidos al aplicar el algoritmo de agrupamiento K-Means en la prueba #2.

Como se puede apreciar en la Figura 17, los perfiles de clústeres resultantes todavía no muestran información muy clara de los clústeres generados, ya que sigue existiendo similitudes entre los clústeres, por ejemplo entre el 1 y 2 y además el 4 y 5. Por lo tanto no se puede establecer una tendencia muy marcada en la participación ciudadana y poder obtener patrones de comportamiento.

Es por esto que se procedió a realizar una nueva prueba con datos de entrada y parámetros distintos.

Resultados Prueba #2

En la Tabla 31 se detallan los parámetros establecidos en la prueba 2, resultados obtenidos y las conclusiones.

Método de agrupamiento	K-Means
Número de clúster	5
Columnas de entrada	Edadr, escolaridad, voto, vivpc, vivtv y vivau

Tabla 31.- Parámetros del algoritmo establecidos en la prueba #2.

Resultados: Los clúster generados en esta prueba aún sigue mostrando similitudes entre clúster, en particular los clúster 1 y 2 son similares entre sí, así como los clúster 4 y 5 respectivamente. Por lo tanto sigue siendo imposible establecer una tendencia entre los votantes.

Conclusiones: Los resultados obtenidos en esta segunda prueba no permiten establecer la tendencia o los patrones de comportamiento, sin embargo nos sirven de apoyo para establecer nuevos parámetros en la siguiente prueba.

Ejemplos: La Figura 20 muestra un ejemplo de los resultados de la prueba #2.

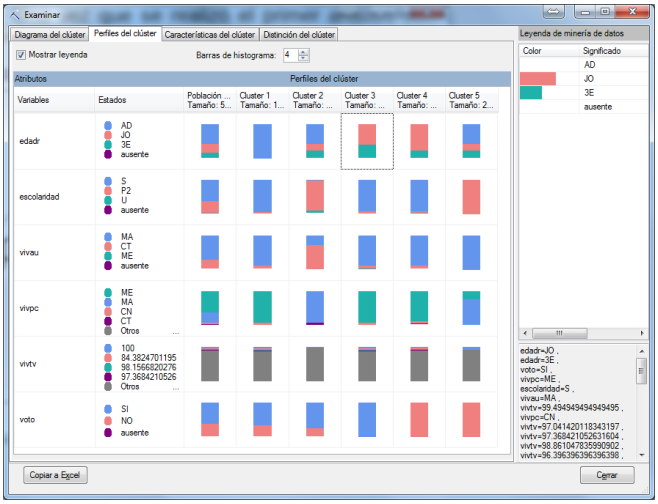

Figura 20.- Clúster obtenidos en la prueba #2 del algoritmo de agrupamiento K-Means.

Prueba #3

Como tercera prueba volvemos a realizar una nueva selección de las columnas de entrada, tomando en cuenta el resultado de la segunda prueba, quedando las columnas de entrada siguientes: ESCOLARIDAD, VOTO, EDADR y VIVAU.

Estas columnas de entrada fueron seleccionadas con base al resultado de la segunda prueba realizada con el algoritmo. Las columnas que se omitieron son las que no representaban gran diferencia entre los clústeres de la segunda prueba.

Además, se volvieron a establecer los parámetros necesarios para la ejecución del análisis. Se siguió conservando el mismo método de agrupamiento (CLUSTERING_METHOD) el cual fue: K-Means Escalable. En el caso del número de clúster (CLUSTER_COUNT) lo modificamos por un valor menor: 3.

Una vez que se llevó a cabo el análisis con el algoritmo seleccionado, después de haber configurado de nuevo sus variables de entrada y parámetros correspondientes, se generó un nuevo diagrama de clústeres, como se muestra en la Figura 18.

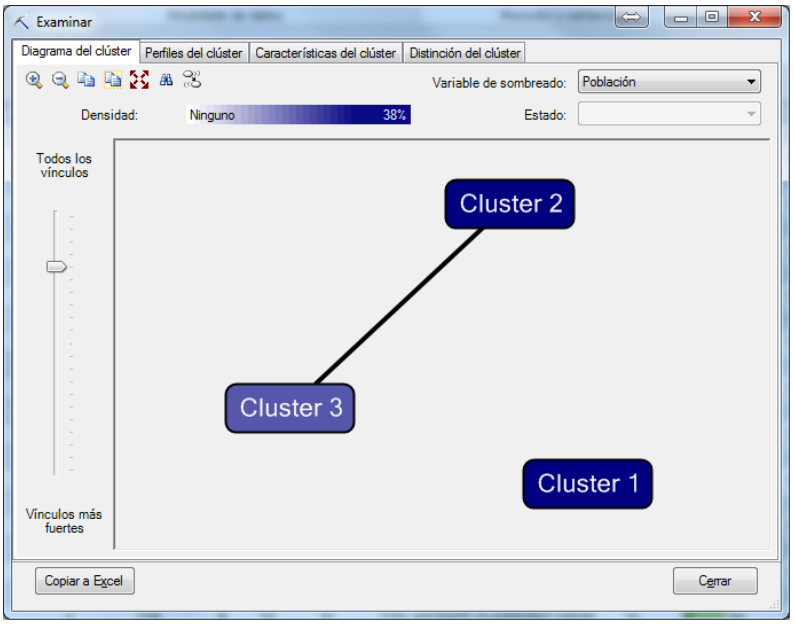

Figura 18.- Diagrama de clúster obtenido al aplicar el algoritmo de agrupamiento K-Means en la prueba #3.

Como se puede apreciar en la Figura 18, el diagrama resultante nos muestra información clara y precisa de los clústeres generados, ya que sus conexiones son muy distintas. Por lo tanto ya podemos hacer una interpretación del resultado del análisis y llegar a una conclusión.

Además analizando el resumen de los perfiles de cada uno de los clústeres generados en el análisis de la información, podemos observar que los 3 clústeres generados son muy distintos unos de otros y, por lo tanto, ya se pueden establecer las tendencias o patrones de comportamiento entre los clústeres generados, es decir entre la participación ciudadana. En la Figura 19 se muestra un ejemplo del resumen de los perfiles.

Después de haber realizado las pruebas necesarias con el algoritmo de agrupamiento y toda vez que obtuvimos la información deseada, ya podemos hacer una conclusión de los resultados obtenidos.

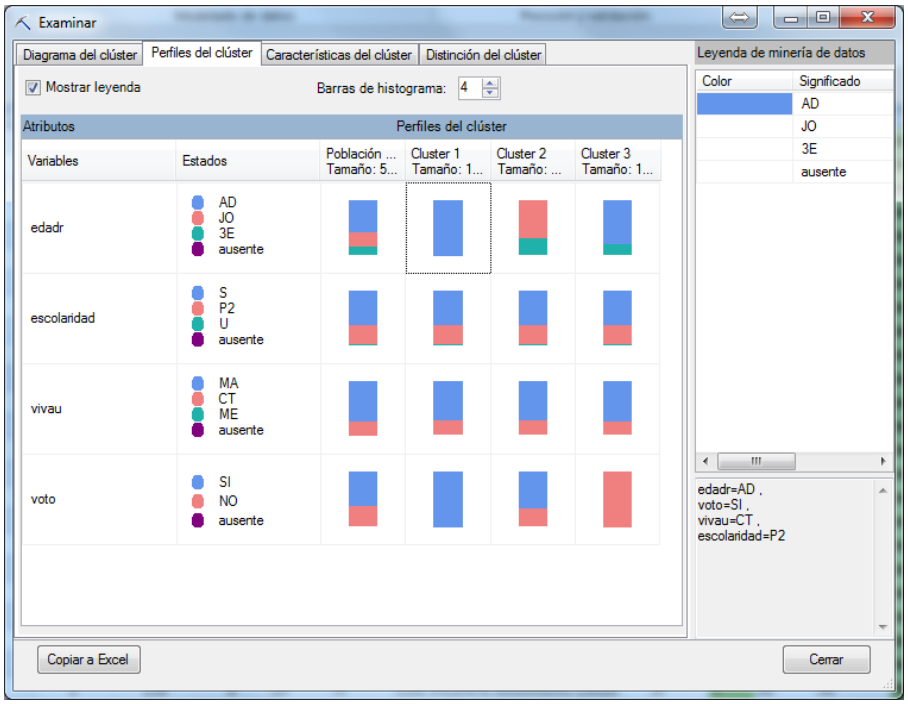

Figura 19.- Perfiles de clúster obtenidos al aplicar el algoritmo de agrupamiento K-Means en la prueba #3.

Resultados Prueba #3

En la Tabla 32 se detallan los parámetros establecidos en la prueba 3, resultados obtenidos y las conclusiones.

Método de agrupamiento	K-Means
Número de clúster	3
Columnas de entrada	Edadr, escolaridad, voto y vivau

Tabla 32.- Parámetros del algoritmo establecidos en la prueba #3.

Resultados: Los clústeres generados en esta prueba ya muestran diferencias muy marcadas entre cada uno de los clústeres. Por ejemplo, en el clúster 1 tenemos a las personas que sí votaron y sus variables similares entre cada uno de ellos. En el clúster 3 tenemos agrupadas a las personas que no votaron y sus características similares. Por lo tanto ya se pueden establecer las tendencias o patrones de comportamiento entre los clústeres, es decir entre la participación ciudadana.

Conclusiones: Los resultados obtenidos en esta tercera prueba ya permiten establecer una tendencia en la participación ciudadana y los patrones de comportamiento de los ciudadanos. Por lo tanto ya no fue necesario realizar pruebas posteriores.

Ejemplos: La Figura 21 muestra un ejemplo de los resultados de la prueba #3.

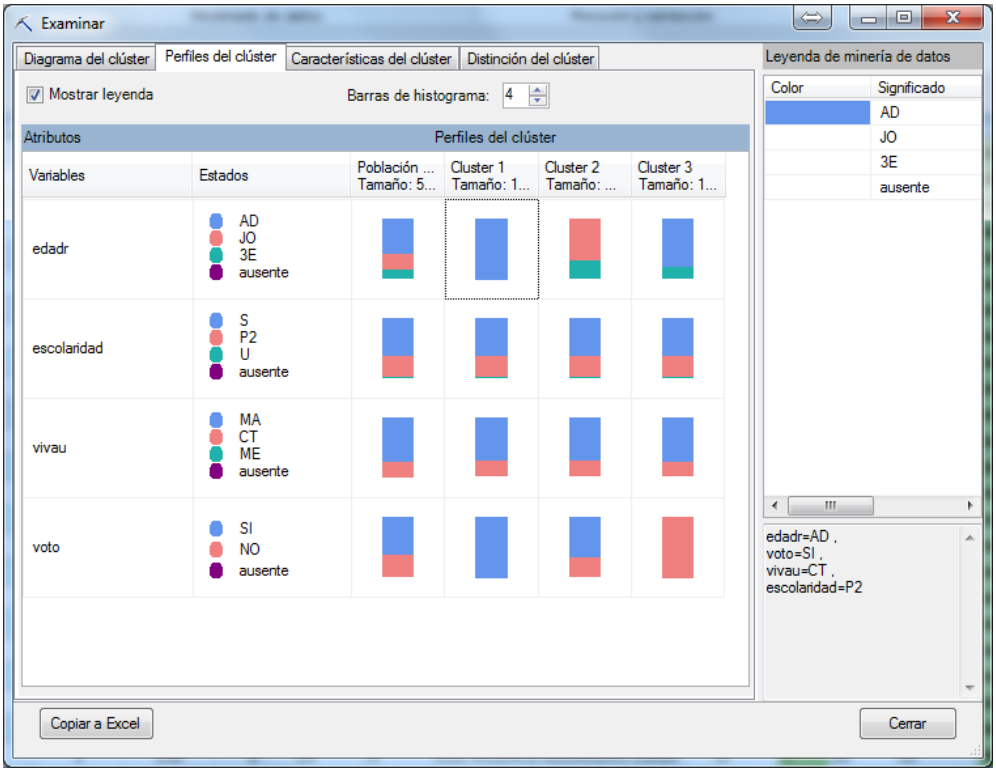

Figura 21.- Clúster obtenidos en la prueba #3 del algoritmo de agrupamiento K-Means.

Después de llevar a cabo todas las pruebas necesarias con el algoritmo de agrupamiento y de haber obtenidos los resultados necesarios para cumplir nuestros objetivos, procedemos a la siguiente etapa que consiste en la interpretación de los resultados.

4.5. Interpretación de los Resultados

4.5.1. Comprensión e interpretación de los resultados obtenidos.

Después de haber descrito los resultados obtenidos en cada una de las pruebas realizadas a los algoritmos de minería de datos seleccionados, se debe realizar una interpretación y comprensión de esos resultados. Esto consiste en describir o especificar la información obtenida en las pruebas para una interpretación más entendible y ordenada, de acuerdo al tema de investigación.

A continuación se describen cuáles son las variables de mayor influencia en los votantes y posteriormente se describe el patrón de comportamiento. Estas dos descripciones forman parte de los objetivos principales de la investigación.

4.5.1.1. Variables o características de mayor relevancia en los votantes.

Como parte de los objetivos específicos, los cuales se definieron con anterioridad en el capítulo 1.5 de este documento, tenemos “identificar las variables más representativas en la participación ciudadana”. Para esto se llevó a cabo la selección y ejecución del algoritmo de clasificación, el cual nos permitió encontrar esas variables más representativas.

Después de varias iteraciones del algoritmo y con la afinación de sus variables de entrada, descartando en cada iteración las variables menos representativas, pudimos obtener resultados más precisos, hasta llegar a la identificación de las variables más representativas y nivel de influencia de cada una de ellas en la participación ciudadana. En la Figura 22 se muestra los ejemplos de los resultados obtenidos.

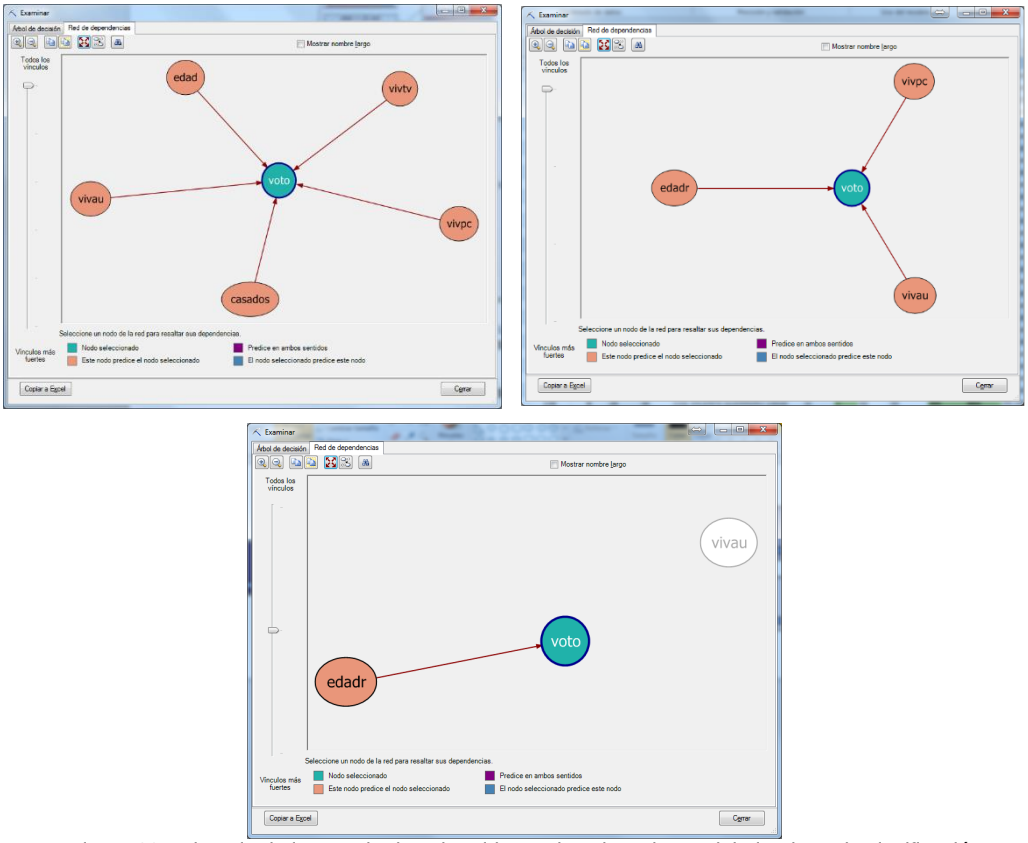

Figura 22.- Ejemplo de los resultados obtenidos en las ejecuciones del algoritmo de clasificación.

Las variables más representativas en la participación ciudadana, ordenadas de mayor a menor nivel de influencia se muestran en la Tabla 33:

EDADR	Rango de edad de los votantes.
VIVAU	Vivienda con automóvil.
VIVPC	Vivienda con computadora.
ESCOLARIDAD	Grados de escolaridad.
VIVTV	Vivienda con televisión.

Tabla 33.-Variables más representativas ordenadas de mayor a menor nivel de influencia.

Como conclusión se establece que la variable de mayor influencia en la participación ciudadana es la edad del votante. Posteriormente, influye mucho también si cuenta con automóvil en su vivienda. Con menor grado de influencia

obtuvimos si la persona cuenta con una computadora en su vivienda, su grado de escolaridad y si cuenta con televisión.

4.5.1.2. Patrón de comportamiento en los votantes

El objetivo general de la investigación es “identificar los patrones de comportamiento en la participación ciudadana”. Para cumplir el objetivo antes descrito se llevó a cabo la selección e implementación del algoritmo de agrupamiento o clúster, el cual permite crear grupos (clústeres) con características similares (columnas de entrada) entre los datos a evaluar. De este modo con la detección de grupos de votantes con características muy parecidas o iguales, se pueden detectar patrones de comportamiento entre las personas que votaron o no, según sea el caso.

Después de varias ejecuciones del algoritmo de agrupamiento, tomando en cuenta las variables más representativas (resultado del algoritmo de clasificación) y donde en cada una de estas ejecuciones se modificaban las columnas de entrada y el número de clúster a generar, se obtuvieron los resultados necesarios y con ello la detección de los patrones de comportamiento. En este caso particular, el objetivo era encontrar el patrón de comportamiento de las personas que acudieron a votar el día de la jornada electoral. En la Figura 23 y 24 se muestran los ejemplos de los resultados obtenidos

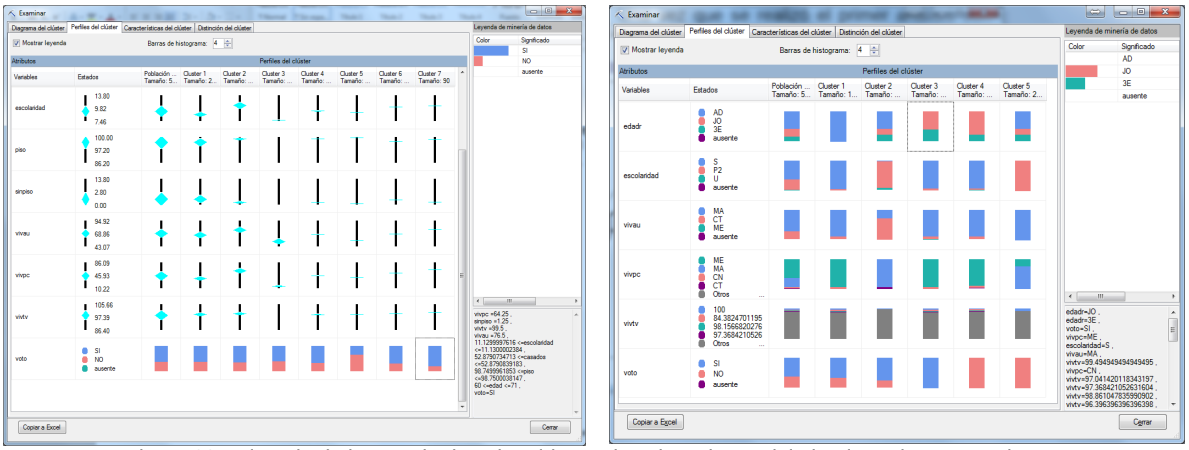

Figura 23.- Ejemplo de los resultados obtenidos en las ejecuciones del algoritmo de agrupamiento.

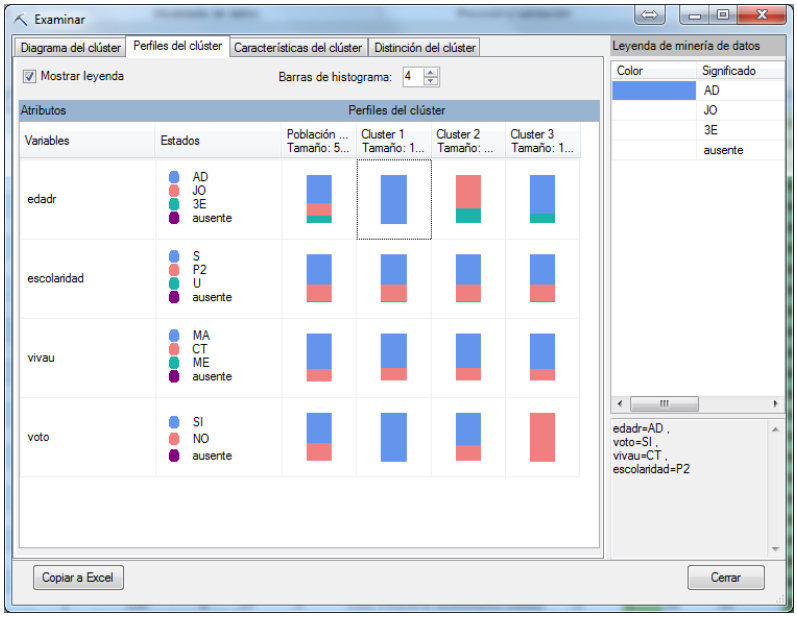

Figura 24.- Ejemplo de los resultados finales obtenidos en las ejecuciones del algoritmo de agrupamiento.

En la ejecución final del algoritmo de agrupamiento, como se muestra en la Figura 24, se generaron 3 clústeres. El clústeres #1 agrupa a las personas que SI votaron, como se puede observar en la Figura 25:

Figura 25.- Clúster #1 y la columna de entrada VOTO agrupado por el valor SI.

Posteriormente observamos que todas las personas que votaron, pertenecientes al clúster #1, también pertenecen (todas) al rango de edad “AD”, el cual corresponde a las personas adultas, es decir, personas entre 30 y 60 años. El resultado se muestra en la Figura 26:

Figura 26.- Clúster #1 y la columna de entrada EDADR.

Continuando con el análisis del clúster #1, observando ahora la columna de entrada de ESCOLARIDAD, se establece que la mayoría cuentan con el valor “S”, el cual corresponde al grado escolar SECUNDARIA. Por lo tanto, podemos concluir que la mayoría de las personas que votaron cuentan con estudios de secundaria. El resultado se muestra en la Figura 27:

Figura 27.- Clúster #1 y la columna de entrada ESCOLARIDAD.

Por último, observando la columna de entrada de VIVAU, se establece que la mayoría cuentan con el valor “MA” (más de la mitad), el cual corresponde o describe si la persona cuenta con automóvil en su vivienda. Por lo tanto podemos concluir que más de la mitad de las personas que votaron cuentan con un automóvil en su vivienda. El resultado se muestra en la Figura 28:

Figura 28.- Clúster #1 y la columna de entrada VIVAU.

Ya que hemos analizado todas y cada una de las columnas de entrada y sus respectivos valores que conforman el clúster #1 (que corresponde a las personas que SI votaron) procedemos a describir el patrón de comportamiento encontrado en el análisis del clúster #1, mediante la Figura 29:

Figura 29.- Patrón de comportamiento de la participación ciudadana del municipio de La Paz.

4.5.2. Evaluar el impacto de los resultados

Previo a la jornada electoral se lleva a cabo una actividad llamada “Promoción del Voto”, la cual consiste en motivar a la ciudadanía para que asistan a votar el día de las elecciones. Esta promoción se lleva a cabo de manera muy general, es decir, aplicando el mismo programa o método de promoción a las ciudades y los poblados del estado de Baja California Sur.

Sin embargo, en las ciudades y los poblados muchas veces se cuentan con características diferentes, por lo que la mayoría de las veces la promoción del voto no tiene el mismo impacto o éxito en los diferentes lugares donde se aplica.

Por lo tanto gracias al análisis de la participación ciudadana de procesos anteriores y la detección de patrones de comportamiento en las personas que acudieron a votar en el municipio de La Paz, se puede llevar a cabo una

planeación estratégica de la promoción del voto para futuros procesos electorales, mucho más enfocada hacia las personas con características particulares de ese municipio.

5. Conclusiones y Recomendaciones

5.1. Conclusiones

En el presente trabajo se obtuvieron los patrones de comportamiento en la participación ciudadana y la identificación de las variables más representativas en un ciudadano a la hora de acudir a votar.

Con base a las variables más representativas encontradas las conclusiones son:

- La participación ciudadana en el municipio de La Paz está mayormente influenciada por la edad del ciudadano, es decir, dependiendo la edad de la persona es la posibilidad de que acuda o no a votar.

- También influye mucho si la persona cuenta con automóvil en su vivienda. Esto quiere decir que si la persona cuenta con transporte para trasladarse a las casillas acude a votar.

- Otro rasgo importante de influencia en la participación ciudadana es la escolaridad de la persona, es decir, los grados de estudios o nivel académico con los que cuenta el posible votante.

- Factores como el piso de la vivienda (si es de material o no) es un factor que no influye en la participación ciudadana.

- Otro factor que tampoco influye es el sexo de las personas (no tiene relevancia si es hombre o mujer) ya que ambos cuentan con una participación similar de 50 y 50.

- Si la vivienda cuenta con drenaje o con servicios públicos, son factores que tampoco influyen a la hora acudir a votar.

Tomando en cuenta el patrón de comportamiento encontrado en la participación ciudadana, las conclusiones son:

- Los ciudadanos del municipio de La Paz que acudieron a votar el día de la jornada electoral fueron personas adultas de entre 30 y 60 años.

- La mayoría de las personas que votaron tienen escolaridad máxima de secundaria, es decir, algunos cuentan con educación básica, secundaria y en algunos casos no cuentan con estudios.

- Más de la mitad, o casi todos, cuentan con un automóvil en su vivienda.

- Otro dato importante es que los jóvenes (en comparación con los adultos) casi no votaron, es decir, personas entre los 18 y 29 años no mostraron interés en ejercer su derecho a votar. Sin embargo, votaron un poco más que las personas de la tercera edad.

- Las personas de la tercera edad (mayores a 60 años) tampoco acuden a votar comúnmente y fueron de las que menos porcentaje de participación registraron.

- Las personas que cuentan con estudios de preparatoria aun representan un bajo porcentaje de participación. Sin embargo, una parte de ellos si acude a votar.

- Las personas con nivel licenciatura o posterior no son muy concurrentes a la hora de acudir a votar el día de la elección, ya que presentaron un porcentaje muy bajo o casi nulo de participación ciudadana.

- En general la población presentó un porcentaje mayor a la mitad de participación el día de las elecciones, generando un 60% de participación ciudadana.

Con estas conclusiones se puede evaluar la promoción del voto realizada en los procesos electorales y, con base a ello, apoyar en la toma de decisiones para una planeación estratégica y focalizada de la promoción del voto en el municipio de La Paz y obtener una mayor participación ciudadana en futuros procesos electorales.

5.2. Recomendaciones

- Planeación de una promoción del voto especializada para el municipio de La Paz, tomando en cuenta los resultados obtenidos en este trabajo.

- Aplicar el modelo de minería de datos para los demás municipios del Estado de Baja California Sur, y de este modo llevar a cabo una promoción del voto especializada con base al comportamiento de la participación ciudadana en cada uno de los municipios.

- Incluir más variables o métricas en el modelo establecido como datos de entrada. Por ejemplo, datos demográficos, culturales, sociales, censos de población.

- Aumentar la generación de bases de datos con mayor volumen de información, que puedan enriquecer el análisis de la información en la búsqueda de nuevos patrones en los próximos procesos electorales.

- Generación de nuevos modelos de minería de datos con objetivos distintos y el descubrimiento de nueva información y/o conocimiento en los procesos electorales.

- Llevar a cabo convenios de colaboración con instituciones y dependencias para el acceso e intercambio de información relevante para la participación ciudadana, tales como: datos censales, demográficos, sociales, culturales, para lograr un mayor rango de análisis en los datos.

- Creación de una bodega de datos para almacenar toda la información generada durante un proceso electoral.

- Creación de un modelo multidimensional para el procesamiento y visualización de la información contenida en una bodega de datos correspondiente a los procesos electorales.

- Aplicar el mismo modelo de este trabajo en los próximos procesos electorales y comparar sus resultados con los obtenidos en esta investigación.

6. Referencias.

1. Petrovic, D. (2010). SQL Server 2008 Manual de referencia. Alemania: Mc Graw Hill.

2. Torres-Jiménez, J. 2011, Breve Introducción a las Bodegas de Datos. http://www.tamps.cinvestav.mx/~jtj/courses/dbs/slides/Bodegas%20de%20datos.pdf

3. Gupta, A. y I. Mumick. 1999, Materialized views: techniques, implementations, and applications. The MIT Press.

4. Dario, B. R. (2010). Metodología para la construcción de un data warehouse. Córdoba, Argentina: Hefesto.

5. Hernández Orallo, J. (2004). Introducción a la Minería de Datos: Pearson.

6. Stanek, W. R. 2012, Microsoft SQL Server 2012 Pocket Consultant.

7. Microsoft.com 2013, Database System, Performance and Scalability, SQL Server 2012 Business Intelligence Editions.

8. Jacobson R. 2000, Microsoft SQL Server 2000 Analysis Services Step by Step. Microsoft Press.