Sistema para la toma de decisiones para la valoración de los programas de bienestar en el estado de Puebla

Dra. Somodevilla García María Josefa[1],
Rodríguez Serrano Alejandro[1], Reyes Peralta Alberto Esteban[1],
Padilla Luis Edgar Abidán[1]

[1*]Facultad De Ciencias de la Computación, Benemérita universidad Autónoma de Puebla, Puebla, 72570, Puebla, México.

Autores:
maria.somodevilla@correo.buap.mx;
alejandro.rodriguezs@alumno.buap.com.mx;
alberto.reyes@alumno.buap.com.mx;
edgar.padillal@alumno.buap.com.mx;

Resumen

La Secretaría de Bienestar en México busca mejorar el bienestar social a través de programas que fomenten la autosuficiencia alimentaria, la inclusión productiva y la reconstrucción del tejido social. En este artículo se propone una investigación sobre la secreteria de bienestar del estado de puebla utilizando el software de licencia libre Weka y sus técnicas de minería de datos como lo son los algoritmo apriori, simple k-means y el J48. Después de realizar diversos experimentos, se muestra que la dispersión de la base de datos juegan un papel muy importante en los resultados finales y que no es posible determinar características novedosas debido a esta característica de la base de datos.

Palabras Clave:: Bienestar, Minería de datos, aprendizaje automático, toma de decisiones.

1 Introducción

La secretaría de bienestar en Puebla es una dependencia del gobierno estatal que tiene como objetivo mejorar las condiciones de vida de las personas en situación de vulnerabilidad, mediante programas sociales que promueven la alimentación, la vivienda, la salud y la participación comunitaria. Algunos de los programas que ofrece son: Coinversión Vivienda, Programa Integral Alimentario, Centros Preventivos de Bienestar y Promueve Bienestar soberanía alimentaria1. Además, la secretaría de bienestar coordina con la delegación federal la entrega de las tarjetas de pensión para adultos mayores, que se pueden recoger en diferentes módulos ubicados en el estado. La secretaría de bienestar tiene su sede en la calle 20 Oriente 2036 colonia Humboldt y su teléfono es 222 7 77 97 00 [1].

La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados [2]. Se puede utilizar para analizar la información que la secretaría de bienestar genera, en relación a los apoyos que distribuye y las zonas donde los entrega, para mejorar la toma de decisiones y optimizar los recursos. Por ejemplo, se podría aplicar la minería de datos para identificar las necesidades de las personas beneficiarias, evaluar el impacto de los programas sociales, detectar posibles fraudes o irregularidades, segmentar a la población según sus características o preferencias, etc.

Weka es un software libre de minería de datos y aprendizaje automático escrito en Java [3]. Es desarrollado en la Universidad de Waikato y es ampliamente utilizado en la industria y la investigación académica. Weka proporciona una interfaz gráfica de usuario para la construcción de modelos de aprendizaje automático a partir de datos y cuenta con una amplia gama de algoritmos de clasificación, regresión, agrupamiento y asociación. También incluye herramientas para el preprocesamiento y visualización de datos, y es compatible con varios formatos de archivos de datos.

En este artículo se presenta una investigación que tiene como objetivo evaluar el impacto de los programas de bienestar en Puebla. Para ello la estructura del documento es la siguiente, en el sección 2 se da un repaso sobre el estado del arte. Por otra parte el planteamiento del problema, la metodología ocupada, el análisis de los datos y la construcción del data werehouse se muestran en el apartado 3. La aplicación de los algoritmos apriori, simple k-means y j48 junto a sus resultados se muestran en la sección 5. Finalmente, se dan las conclusiones en la parte 6 del documento.

2 Estado del arte

En esta sección se realiza una recopilación de artículos relacionados a minería de datos, su relación con la herramienta de software libre Weka y su uso con información gubernamental. En el artículo [4] hace una revisión de la literatura sobre la minería de datos y sus aplicaciones en la toma de decisiones. Habla sobre la aplicación de técnicas de minería de datos como una herramienta estratégica en la toma de decisiones empresariales. El objetivo principal es demostrar cómo la minería de datos puede ser una solución eficiente para la gestión de grandes cantidades de información y cómo puede proporcionar información valiosa para la toma de decisiones empresariales.

El artículo titulado "A review of decision-making methods based on multi-criteria analysis in sustainable energy planning" [5] realiza una revisión exhaustiva de los diferentes métodos de toma de decisiones basados en análisis multicriterio utilizados en la planificación de la energía sostenible señalando que se necesita una mayor investigación para mejorar la eficacia y eficiencia de estos métodos de análisis.

Otro este otro artículo [6] se ofrece una visión general de la aplicación de la minería de datos en la toma de decisiones empresariales. Se hace también una revisión del uso de la minería de datos en diferentes áreas empresariales, incluyendo marketing, finanzas, recursos humanos, logística, entre otras.

Uno de los artículos más relevantes encontrados se enfoca en el análisis del sector energético en México mediante la aplicación de técnicas de minería de datos. En este estudio, los autores utilizan datos gubernamentales para extraer información relevante sobre los patrones de consumo de energía en diferentes regiones del país [7].

En relación con la seguridad, se encontró otro artículo que se centra en el análisis de tendencias y patrones de homicidios en México mediante el uso de técnicas de minería de datos. Los autores procesaron los datos gubernamentales y analizaron los patrones y tendencias de los homicidios en distintas regiones del país [8].

En cuanto a la calidad del aire en la Ciudad de México, se identificó un estudio que aplicó técnicas de minería de datos para el análisis de los datos proporcionados por el gobierno. El objetivo fue explorar patrones y tendencias en la contaminación del aire en diferentes regiones de la ciudad [9].

Por otro lado, se encontró un artículo que aborda la identificación de patrones de fraude fiscal en México mediante el uso de técnicas de minería de datos. Los autores aplicaron estas técnicas a los datos gubernamentales para analizar los patrones y tendencias de los fraudes fiscales en distintas regiones del país [10].

Finalmente, se encontró otro estudio que se enfoca en el análisis de los delitos en México mediante el uso de técnicas de minería de datos. Los autores analizaron los datos proporcionados por el gobierno para identificar patrones y tendencias en distintas regiones del país, explorando así las características comunes de los delitos cometidos [11].

Berón y Méjía [12] describen el proceso de desarrollo de una metodología para la identificación de las principales causas de ausentismo en una empresa de la región central de Colombia. Para ello, se utilizó el algoritmo de clasificación J48 desde la plataforma libre de aprendizaje automático y minería de datos Weka.

Castrillón, Castillo y Castaño utilizan la herramienta de minería de datos Weka para llevar a cabo el análisis de clasificación [13]. Primero, se preprocesaron los datos, realizando la eliminación de registros con datos faltantes y la transformación de los datos categóricos a numéricos. Luego, se dividió el conjunto de datos en dos grupos: uno de entrenamiento (70% de los datos) y uno de prueba (30% de los datos). Finalmente, se aplicaron diferentes técnicas de clasificación, tales como árboles de decisión (J48), redes neuronales, Naïve Bayes y Máquinas de Soporte Vectorial (SVM), para predecir la variable dependiente de interés, en este caso, el género de la persona.

En Spositto *et al* recopila información relevante sobre incendios forestales [14] , sus causas y consecuencias, y revisa las diferentes técnicas de detección y prevención existentes. Luego, describe la metodología utilizada para el desarrollo del modelo,

que incluye la selección y preprocesamiento de datos, la aplicación de técnicas de clasificación y la evaluación del modelo. Se utiliza la herramienta Weka para aplicar los algoritmos de clasificación, como Naive Bayes, J48 y Random Forest, a los datos recopilados.

Quezada *et al* describen una metodología para predecir y analizar los factores socio-económicos que influyen en el embarazo adolescente en Colombia [15]. Se utiliza el algoritmo de clasificación J48, ejecutado en la plataforma Weka, para predecir el comportamiento de la variable dependiente "Tiene Hijos (TH)" y para identificar las principales causas que generan este comportamiento.

En [16], Franco y Martínez experimentan con varias técnicas de weka respecto a minería de datos educativas para generar modelos predictivos de rendimiento académico con un buen nivel de confiabilidad y exactitud, lo que permite identificar casos de alumnos en situación de riesgo académico. Se analizaron diversas fuentes de datos y se obtuvo un archivo con 415 instancias de alumnos cada uno con 65 atributos.

3 Desarrollo

3.1 Planteamiento del problema

Los programas de la Secretaría de Bienestar del estado de Puebla son una estrategia del gobierno para mejorar las condiciones de vida de las personas en situación de vulnerabilidad, mediante acciones que promueven la autosuficiencia alimentaria, la inclusión productiva, la atención a la salud y la cohesión social.

En este trabajo se plantea utilizar técnicas de minería de datos para analizar y poder generar modelos que muestren resultados sobre la eficacia de los programas de bienestar, para ello se emplea la herramienta Weka, un software de código abierto que implementa diversos algoritmos y técnicas para realizar minería de datos.

Los modelos generados pueden servir de apoyo en la toma de decisiones de futuros programas gubernamentales.

3.2 Metodología

Para el desarrollo del trabajo se sigue la siguiente metodología, la cual se muestra en la Figura 1. Primero se realiza un análisis a la base de datos en donde se estudian las tablas y atributos en busca de datos relevantes para el estudio, una vez elegidos los datos a utilizar se verifica la integridad de las tablas, se eliminan o rellanan valores nulos utilizando el criterio de media o mediana, posteriormente se procede a crear el diseño del Data Ware House, finalmente con Weka se obtienen los modelos y resultados.

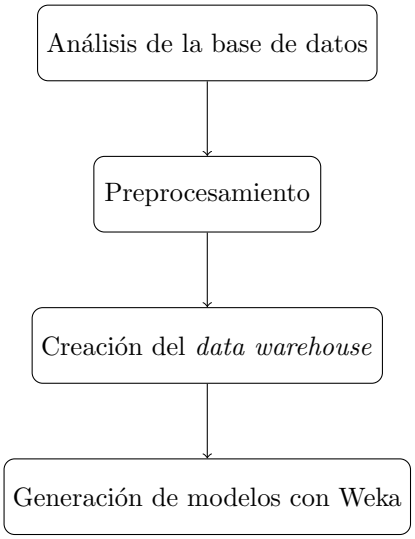


Figura 1: Metodología.

3.2.1 Análisis de los datos

La base de datos de bienestar del estado de Puebla consta de 42 tablas, los datos que se presentan en la base de datos de bienestar corresponden a los años 2021 y 2022.

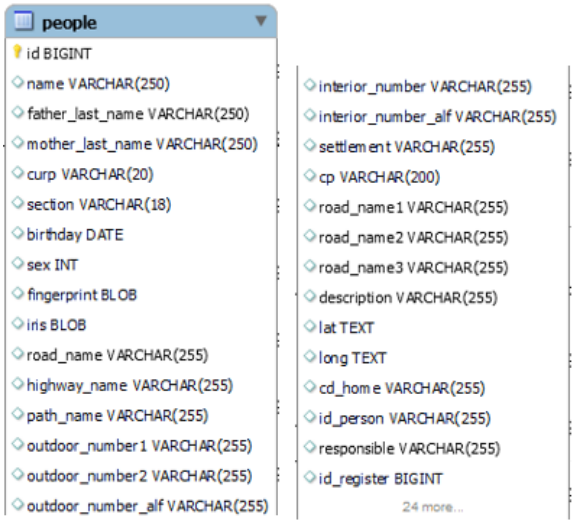


Figura 2: Tabla *people* de la base de datos de bienestar

La tabla principal, *people*, tiene 197295 registros que contienen la información principal de las personas registrados en algún programa de bienestar, esta tabla se muestra en la figura 2, debido al tipo de datos tanto numéricos como de caracteres esta tabla es de suma importancia en los análisis posteriores.

Después de un análisis de la base de datos, se recuperó información de otras 2 tablas *program_details* y *municipalities_catalogs*, estas tablas se aprecian en las figuras 3 y 4 respectivamente.

La tabla de programas consta con 26 registros, en la siguiente lista se muestran los nombres de estos programas y sus objetivos [17].

- Calentadores solares: Suministrar e instalar calentadores solares en las viviendas de las y los poblanos que habitan en zonas vulnerables, fomentando el uso de energías alternas y cuidado del medio ambiente.
- Calidez sustentable: Contribuir al bienestar social mediante estrategias orientadas a disminuir las carencias sociales, a través de la participación social, corresponsabilidad y coinversión de los diferentes sectores, mediante la modalidad de acciones otorgadas.
- Captadores de agua de lluvia: Contribuir a mejorar el nivel de vida y bienestar de las familias poblanas en situación de pobreza y/o vulnerabilidad, a través del suministro e instalación de un sistema de captación y almacenamiento de agua pluvial.
- Centros preventivos de bienestar: Los Centros Preventivos de Bienestar, han sido creados para contribuir con una cultura en Salud Preventiva, a través de detecciones, seguimiento y tratamiento oportuno de enfermedades
- Coinversión social: Contribuir a fortalecer la cohesión social y mejorar el acceso a los derechos sociales de la población del Estado de Puebla mediante proyectos de las Organizaciones de la Sociedad Civil.
- Construcción y colocación de techos: Contribuir a mejorar la calidad de vida de las familias vulnerables poblanas en situación de pobreza, mediante la mejora de la calidad y espacios de la vivienda.
- Cuartos dormitorios: Mejorar la Calidad y Espacios de la Vivienda de la población de los Municipios del Estado de Puebla, mediante la entrega de un cuarto dormitorio, un cuarto para baño y un cuarto para cocina.
- Cuartos y baños para el sismo.
- Electrificación no convencional: Contribuir a mejorar el nivel de vida y bienestar de las familias poblanas en situación de pobreza y/o vulnerabilidad, a través del suministro e instalación de sistemas fotovoltaicos para dotar de energía eléctrica no convencional, en sus viviendas.
- Estufas ecológicas: Mejorar el acceso a los servicios básicos en la vivienda de las familias poblanas que habitan en viviendas que no disponen de estufa con chimenea cuando utilizan leña o carbón para cocinar.
- Fondo de apoyo a migrantes (fam): El Fondo de Apoyo a Migrantes tiene el objetivo de apoyar a los migrantes mexicanos en el retorno a sus lugares de origen, ayudarles a encontrar una ocupación dentro del mercado formal, incrementar sus opciones de autoempleo, así como fomentar la operación de albergues que los atiendan.
- Migrantes poblanos: Hacer frente a las necesidades derivadas del fenómeno migratorio; para brindar protección a las y los migrantes poblanos que viven o presentan

dificultades en el extranjero, aquellos que retornan al Estado y de sus familias en su lugar de origen conforme a la Política Exterior de México

- Otro programa.
- Pisos firmes: Mejorar el acceso y cobertura de la calidad y espacios en la vivienda de las familias del Estado de Puebla, a través del servicio de entrega de un techo firme y un piso firme.
- Programa de esquemas de financiamiento y subsidio federal para vivienda: Otorgar subsidios económicos a familias mexicanas que presentan alguna necesidad en su vivienda, pero que por sus bajos ingresos (sin llegar a la pobreza extrema), están excluidos del sistema de financiero, incluido el mercado hipotecario.
- Programa de infraestructura indígena: Contribuir a la disminución de rezagos, preferentemente en agua potable, electrificación, caminos, carreteras y alcantarillado de los pueblos y comunidades indígenas, a través de proyectos y obras de infraestructura básica, a partir del reconocimiento de sus demandas.
- Programa de organización productiva para mujeres indígenas (popmi): Contribuir a mejorar las condiciones de vida y posición social de las mujeres indígenas, fortaleciendo su participación como impulsoras de su propio desarrollo, mediante la ejecución de Proyectos de Organización Productiva, con perspectiva de equidad de género, sustentabilidad, multiculturalidad y derechos.
- Programa integral alimentario: Contribuir a mejorar el acceso a la alimentación de mujeres, de personas de municipios de alto y muy alto rezago social o de personas en zonas indígenas.
- Programa para el mejoramiento de la producción y la productividad indígena (proin): Contribuir al fortalecimiento de las economías de los pueblos y comunidades indígenas y afromexicanas, con la implementación de proyectos productivos y turísticos.
- Programa vivienda digna: otorga subsidios a los hogares mexicanos en situación de pobreza con ingresos por debajo de la línea de bienestar, con carencia de calidad y espacios de la vivienda, para que construyan, amplíen o mejoren sus viviendas.
- Proyectos productivos: Fomentar el trabajo digno mediante el otorgamiento de apoyos para el impulso de proyectos de emprendimiento para la generación o consolidación de empleo por cuenta propia, preferentemente innovadores, sostenibles y sustentables.
- Sanitarios con biodigestores: Consiste en mejorar la condición y la calidad de vida de las familias poblanas en situación de pobreza que presenten carencia en los servicios básicos de su vivienda, cuyos hogares se ubiquen en localidades de los municipios con muy alto rezago social, a través de la realización, acondicionamiento, suministro e instalación en el inmueble respectivo de un módulo sanitario con biodigestor.
- Tanques de agua potable: Prevenir enfermedades en la población del municipio, producidas por el almacenamiento incorrecto del agua potable en sus viviendas, y mejorar el acceso a los servicios de agua mediante la entrega de un tanque seguro y con tapa para el almacenamiento de agua potable, el cual, será instalado por los/las beneficiarios/as en los lugares que consideren convenientes en sus viviendas.

- Unidades móviles alimentarias (uma´s): Contribuir en la disminución del rezago social mediante el acercamiento y gestión de programas y servicios integrales en materia de alimentación y desarrollo social.
- Unidades móviles de desarrollo (umd´s): Ser el vínculo para contribuir a mejorar el bienestar social de la población en situación de pobreza, de comunidades rurales y áreas marginadas, a través de servicios de, salud física y emocional; deporte y activación física; cultura civilidad y desarrollo de capacidades.
- Yo si voy al preescolar: Ampliar la cobertura en educación preescolar incorporando a niños de tres a cinco años de edad que no asistían a la escuela, a fin de fortalecer sus competencias e iniciar su educación formal.

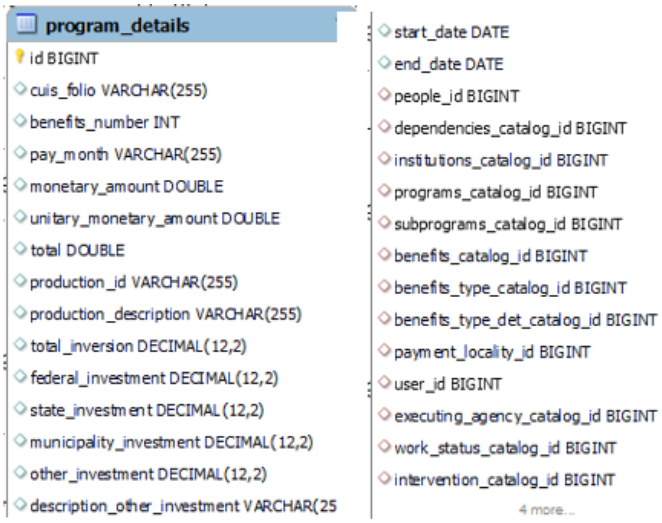


Figura 3: Tabla *program_details*.

La tabla 3 contiene toda la información referente a los programas ocupada en el análisis, como lo son los nombres, cantidad monetaria del programa por mes, cantidad unitaria y cantidad total, estos datos son ocupados para hacer menos dispersa la base de datos, ya que en ella se presentan instancias nulas o vacías.

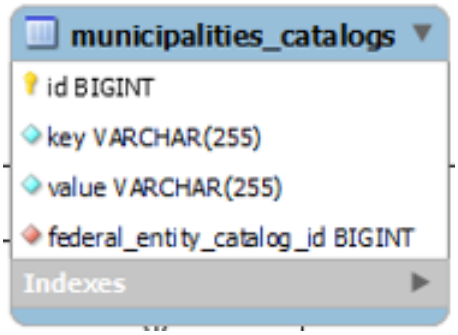


Figura 4: Tabla *municipalities_catalogs*.

Por otra parte la tabla de municipios nos ofrece información referente a la zona geográfica en donde se aplican los programas de bienestar como se muestra en la figura 4, consta con 217 municipios pertenecientes al estado de Puebla.

Cabe destacar que la base de datos ocupada presenta un alto grado de dispersión, por lo que se optó por ignorar los registros nulos. Además existen registros de programas en la base de datos que no cuentan con una cantidad de dinero asignado, sin embargo otros registros del mismo programa cuentan con montos variables, por lo que esos datos restantes se completan utilizando el método de la media.

3.2.2 Construcción del *data warehouse*

Los almacenes de datos (*data warehouses*) son infraestructuras diseñadas para almacenar grandes volúmenes de datos de diversas fuentes en un formato estructurado y optimizado para consultas y análisis. Su principal objetivo es brindar una vista unificada y coherente de los datos de una organización, lo que facilita la toma de decisiones basada en datos.

Para construir el *data warehouse* se consideran 4 dimensiones, localidad, beneficiarios, programa y tiempo. El hecho medible es la cantidad de dinero que corresponde a cada programa. Se puede apreciar el diseño del almacén de datos en la figura 5. La dimensión de beneficiario obtienen sus elementos de la tabla *people* los elementos que se recuperan son: fecha de nacimiento, sexo, edad y dirección (la cual esta representada por los atributos road_name, value, value1 y cp). La dimensión de programa se recupera de la tabla *program_details* y se conforma por los atributos nombre del programa, el id de la persona, el monto en efectivo que otorga dicho programa y el numero de meses que dura el programa. También podemos observar de la figura 5 que la dimensión de localidad tiene como atributo principal el nombre del municipio.

Figura 5: Data Werehouse.

Para construir la dimensión de tiempo mediante la fecha de creación de un usuario presente en la tabla *people* se ocupó el siguiente algoritmo:

```
SET Language 'Spanish';

DECLARE @StartDate  date = '20180101';
DECLARE @CutoffDate date = DATEADD(DAY, -1, DATEADD(YEAR, 10, @StartDate));

;WITH seq(n) AS
(
  SELECT 0 UNION ALL SELECT n + 1 FROM seq
  WHERE n < DATEDIFF(DAY, @StartDate, @CutoffDate)
),
d(d) AS
(
  SELECT DATEADD(DAY, n, @StartDate) FROM seq
),
src AS
(
  SELECT
DateKey = CONVERT(VARCHAR(20),CONVERT(date, d),112),
    TheDate         = CONVERT(date, d),
    TheDay          = DATEPART(DAY,       d),
    TheDayName      = DATENAME(WEEKDAY,   d),
```

```
    TheWeek         = DATEPART(WEEK,      d),
    TheISOWeek      = DATEPART(ISO_WEEK,  d),
    TheDayOfWeek    = DATEPART(WEEKDAY,   d),
    TheMonth        = DATEPART(MONTH,     d),
    TheMonthName    = DATENAME(MONTH,     d),
    TheQuarter      = DATEPART(Quarter,   d),
    TheYear         = DATEPART(YEAR,      d),
    TheFirstOfMonth = DATEFROMPARTS(YEAR(d), MONTH(d), 1),
    TheLastOfYear   = DATEFROMPARTS(YEAR(d), 12, 31),
    TheDayOfYear    = DATEPART(DAYOFYEAR, d)
  FROM d
)
insert into dbo.dim_time
SELECT DateKey, TheDate, TheDay, TheDayName, TheWeek, TheDayOfWeek,
        TheMonth, TheMonthName, TheQuarter, TheYear, TheFirstOfMonth,
        TheLastOfYear, TheDayOfYear
FROM src
ORDER BY TheDate
OPTION (MAXRECURSION 0);
```

Este algoritmo genera registros desde el primero de enero de 2018, los registros contienen los datos de número del día, nombre del día, semana del año, mes, trimestre y año, así cuando un día es seleccionado de la base de datos, todos estos elementos son asociados a este día y se guardan en la dimensión de tiempo.

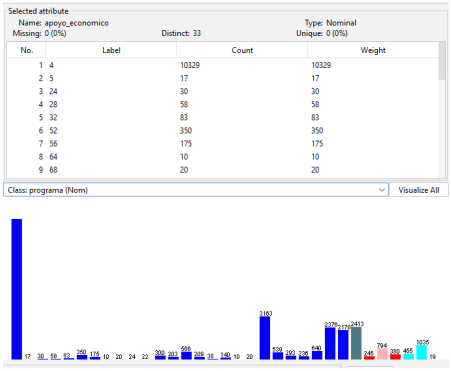
4 Resultados

Para realizar la conexión del data werehouse construido en *sql server* con Weka se ocupa la siguiente consulta:

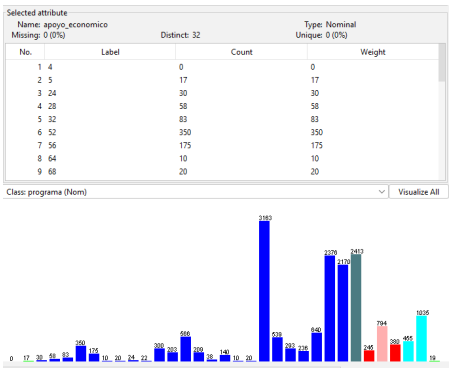
```
    select facts.monetary_amount as apoyo_economico,
            dim_beneficiary.sex as sexo,
            dim_beneficiary.cp as codigo_postal,
            dim_beneficiary.value as localidad,
            dim_locality.value as municipio,
            DATEDIFF(YEAR,dim_beneficiary.birthday,GETDATE()) as edad,
            dim_program.value as programa
    from facts inner join
     dim_program on facts.id_program = dim_program.id inner join
     dim_time on facts.id_time = dim_time.DateKey inner join
     dim_locality on facts.id_locality = dim_locality.id inner join
     dim_beneficiary on facts.id_beneficiary = dim_beneficiary.id
    where not dim_program.value = 'CENTROS PREVENTIVOS DE BIENESTAR' and
            dim_program.monetary_amount is not null
```

4.1 Algoritmo Apriori

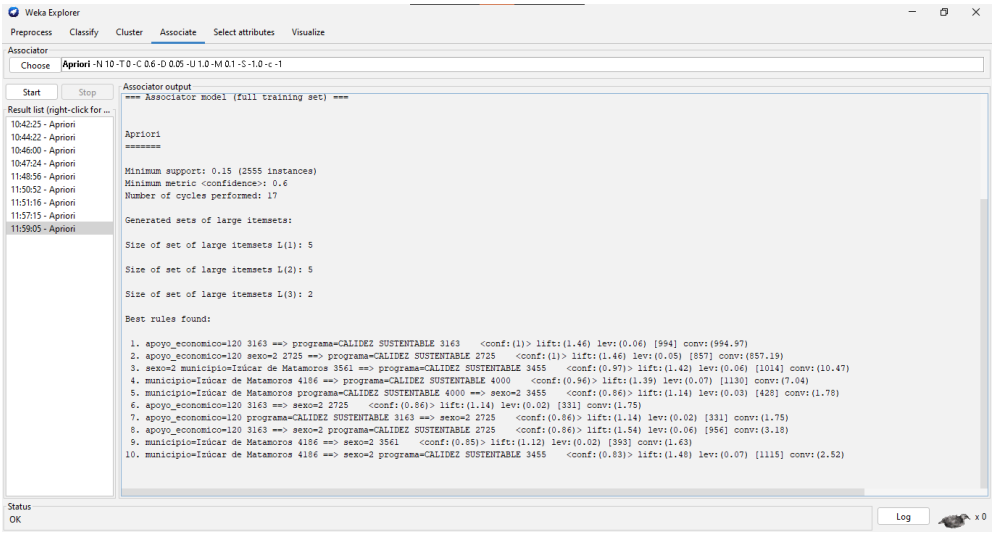
El objetivo del algoritmo Apriori en Weka es descubrir conjuntos de elementos frecuentes en un conjunto de datos y utilizar estos conjuntos para la generación de reglas de asociación y otras tareas de minería de datos. Después de analizar los datos se establece en weka que el atributo nominal que se manejará en esta sección es la dimensión de programa.



(a) Datos antes de remover el apoyo económico de 4 pesos.



(b) Datos después de remover apoyo económico de 4 pesos.

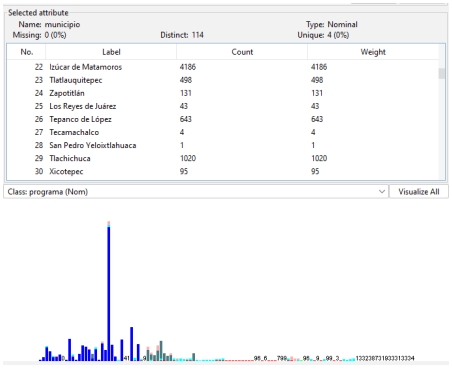


(c) Resultados del algoritmo apriorí removiendo el apoyo económico de 4 pesos.

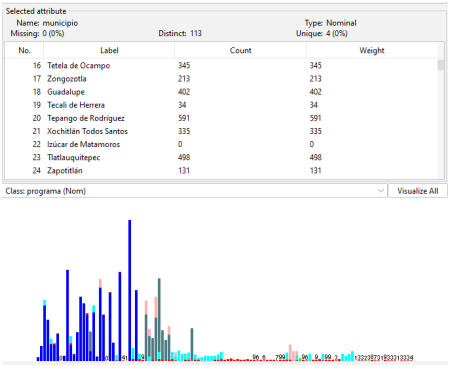
Figura 6: Experimento: Remover el programa con apoyo económico de 4 pesos.

En la figura 6 se muestra el preprocesamiento aplicado al conjunto de datos que se encuentra en el data werehouse, este preprocesamiento consiste en eliminar el apoyo

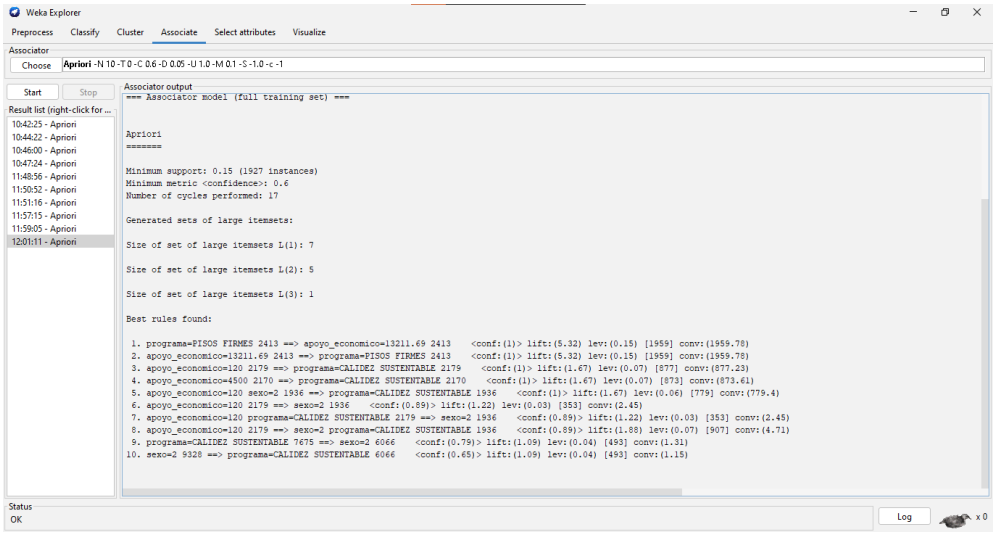
monetario de 4 pesos (6a), ya que estas instancias son considerados *outliers* al ser elementos que se encuentran fuera del rango común de los apoyos gubernamentales. Se puede apreciar en la figura 6b como quedan los datos al aplicar el método *remove with values* en Weka. Observamos que las regla de asociación muestran datos obvios, como el echo que de 3163 personas que reciben el programa calidez sustentable, reciben el monto de 120 pesos o que en Izúcar de Matamoros de 4186 personas 4000 reciben el apoyo de calidez sustentable, se pude ver que que de las 4000 el 85% corresponden al genero femenino, es decir 3561 mujeres reciben este apoyo. Estos resultados se muestran en la figura 6c.



(a) Datos antes de remover la localidad Izúcar de Matamoros.



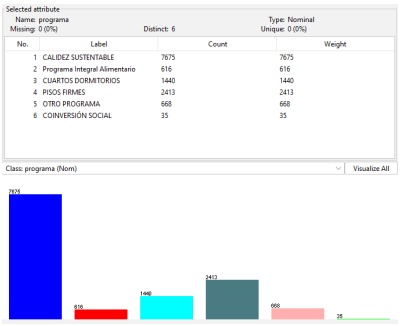
(b) Datos después de remover la localidad Izúcar de Matamoros.



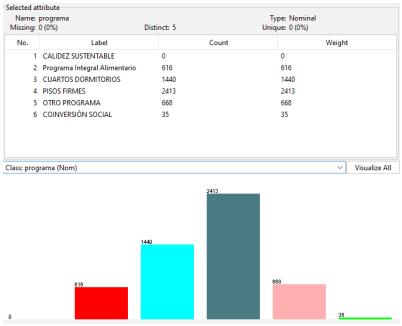
(c) Resultados al remover la localidad Izúcar de Matamoros.

Figura 7: Experimento: Remover localidad Izúcar de Matamoros.

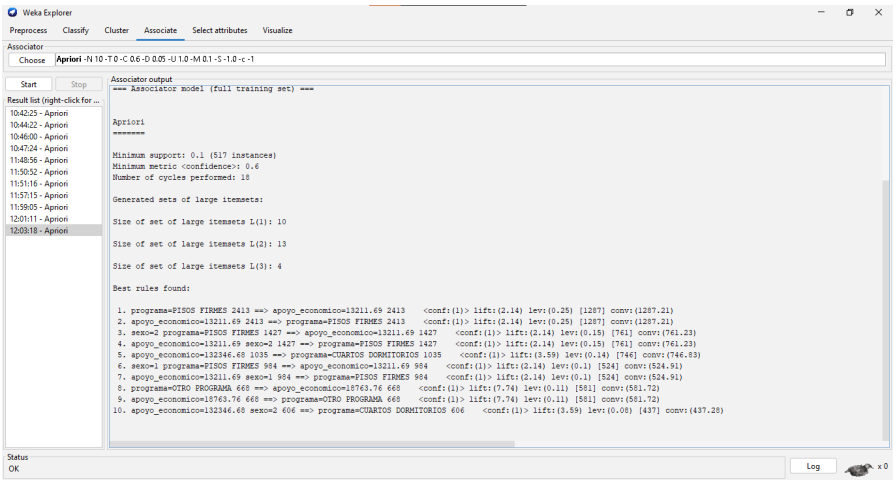
En la figura 7 podemos apreciar el segundo experimento realizado, en el se elimina el programa que da un apoyo económico de 4 pesos también se remueve el municipio de Izúcar de Matamóros, esto debido a que tiene una gran cantidad de instancias y hace que los datos no estén balanceados 7a. En la figura 7b los datos de localidad se encuentran mejor balanceados, este ajuste da como resultado reglas de asociación que relacionan un programa y el apoyo económico recibido, por ejemplo la regla de asociación 8 de la figura 7c nos indica que de 2179 de personas que reciben un apoyo económico de 120 pesos, el 79% porciento son mujeres y se encuentran en el programa calidez sustentable. Podemos ver que estas reglas de asociación no arrojan resultados novedosos pues estos datos pueden ser obtenidos mediante consultas sql.



(a) Datos de la dimensión de programa antes de remover la instancia calidez sustentable.



(b) Datos de la dimensión de programa después de remover la instancia calidez sustentable.



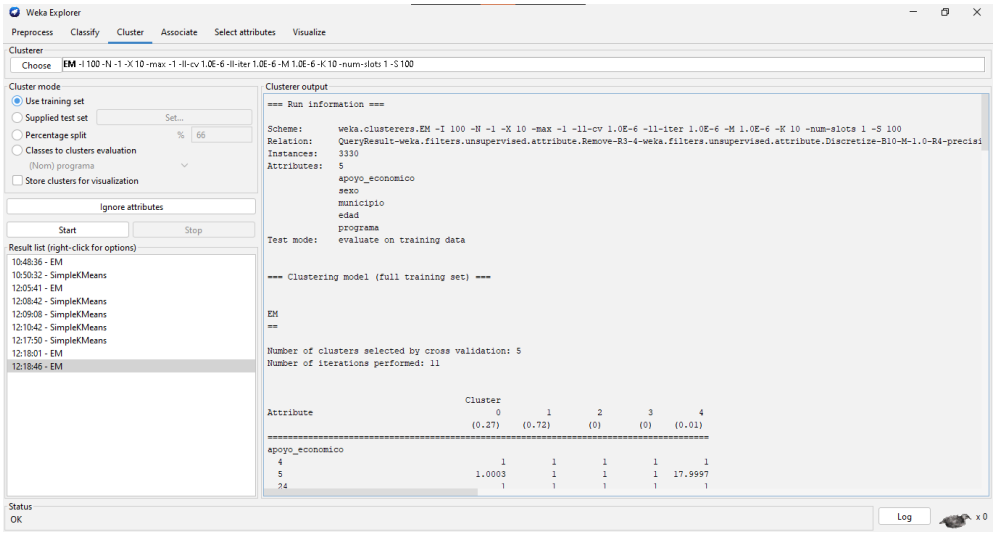
(c) Reglas de asociación al remover el programa calidez sustentable.

Figura 8: Experimento: Remover el programa calidez sustentable.

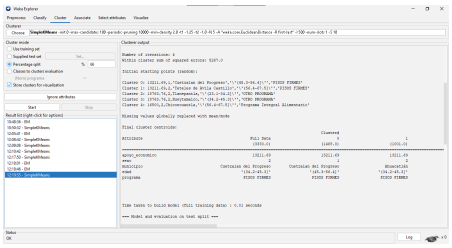
Como tercer experimento se elimina el programa calidez sustentable (ver figura 8), este programa ofrece un monto de 250,000 pesos y consiste en coinversión social por lo que no solo se le otorga a una persona si no más bien a una asociación, debido a que se busca encontrar patrones entre personas y programas, se optó por remover esta instancia en la figura 8a y la figura 8b se observa el antes y después de remover el programa calidez sustentable. Los resultados son similares al experimento 2 y se observa que las reglas obtenidas puden ser obtenidads con simples consultas sql.

4.2 Algoritmo Simple k-means

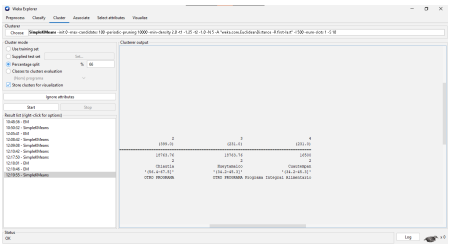
El algoritmo k-means en Weka tiene por objetivo agrupar los datos en k clústeres, donde k es un número predefinido por el usuario. El algoritmo busca minimizar la varianza dentro de cada clúster y maximizar la distancia entre los clústeres.



(a) Algoritmo EM



(b) Resultado del algoritmo simple k-means.



(c) Resultado del algoritmo simple k-means.

Figura 9: Experimento: Algoritmo simple k-means.

Aplicamos este algoritmo utilizando el siguiente preprocesamiento: se discretizan en 10 rangos las edades de las personas, se eliminaron los programas que ofrecen un apoyo económico de 4 y los que ofrecen apoyos mayores a 135,00 pesos, se eliminó el municipio de Izúcar de Matamóros ya que muchas personas que reciben algún programa pertenecen a esta localidad, se elimina el programa calidez sustentable pues existe un alto numero de personas en este programa y desbalancean los datos.

En la figura 9 se observa el experimento realizado para obtener clusters mediante el algoritmo simple k-means, para ello primero se utilizó el algoritmo EM, el cual nos da una idea del número de clusters razonables que se pueden obtener. Se puede apreciar en la figura 9a que el algoritmo arrojo 5 clusters. Este número es utilizado como parámetro en el algoritmo simple k-means. En la figura 9b se muestran los puntos iniciales de los 5 clusters y la descripcion de los primeros dos. Los restantes 3 clusters se muestran en la figura 9c, en ella se aprecian los centroides para los primeros 3 clusters, podemos apreciar que la edad promedio de los clusters se encuentra entre los 34 y 45 años y el programa sobresaliente es pisos firmes.

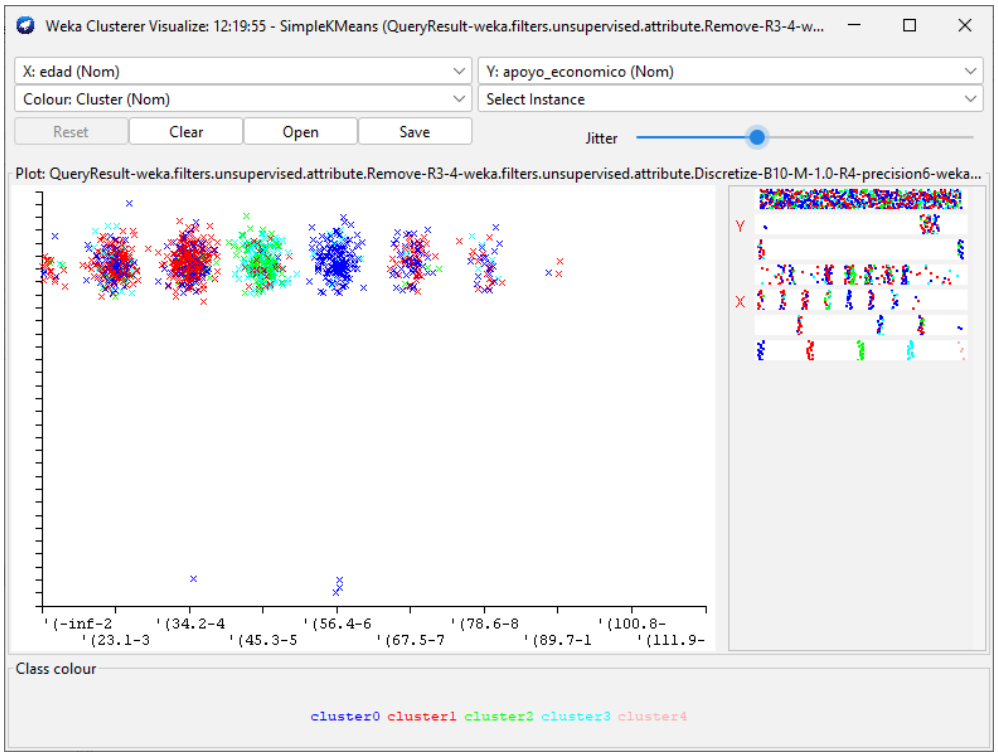


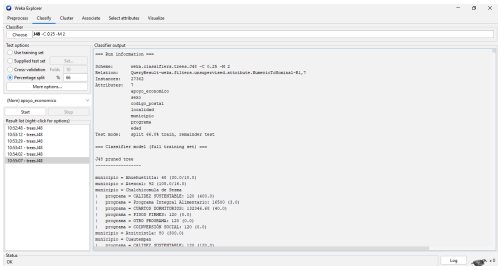
Figura 10: Visualización de los clusters

En la figura 10 se aprecian los como los datos son divididos por sus rangos de edad, el cluster 1 se presenta con mayor presencia en el rango de 34 a 45 años y el cluster 0 se presenta en un rango de edad de 56 a 67 años, revisando los datos de la

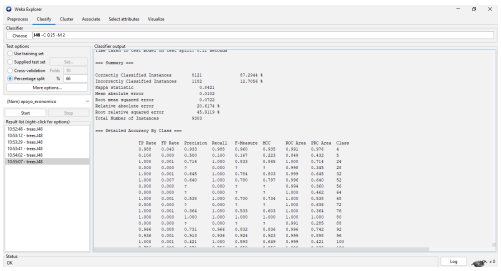
figura 9 se observa que el cluster 1 tiene en promedio un apoyo económico de 13,211.69 pesos entregados en su mayoria a hombres pertenecientes a la localidad de Cuetzala del Progreso y que se encuentran en el programa pisos firmes. Por otra parte el cluster 1 se centra en las mujeres que reciben el apoyo de pisos firmes en el municipio de Ahuacatlán. Tambien se puede observar que los centroides de los cluster tienen a ser personas femeninas, esto puede indicar que las mujeres tienden a ser más responsables al momento de buscar apoyos sociales para el beneficio de su familia.

4.3 Algoritmo J48

El algoritmo J48 genera un modelo de árbol de decisión capaz de clasificar con precisión nuevas instancias desconocidas. Este modelo puede proporcionar comprensión y explicabilidad en el proceso de toma de decisiones, ya que permite seguir las reglas y caminos de decisión definidos por el árbol.



(a) Parámetros utilizados para generar el modelo del algoritmo J48



(b) Resultados del algoritmo J48

Figura 11: Experimento: Algoritmo J48

El experimento que se realizó utilizando el algoritmo J48 mostrado en la figura 11 tomo como argumentos las 7 atributos del *data werehouse*: el apoyo económico; el sexo de un beneficiario, la dirección del beneficiario conformado por código postal, localidad y municipio; el programa al que perteneces el beneficiario y su edad. El preprocesamiento realizado es el siguiente: se eliminaron las instancias que desbalancean

los datos como el programa calidez sustentable; los apoyos de 4 pesos y los que otorgan más de 135000 pesos además se eliminó la localidad de Izúcar de Matamoros el cual tienen una densa población. Estos elementos pueden apreciarse en la figura 11a utilizando un 66% de porcentaje de split para dividir el dataset en entrenamiento y prueba, con estos valores se obtiene un 87 por ciento de datos correctamente clasificados con un error cuadrático medio de 45 por ciento, estos resultados se observan en la figura 11b. Por lo que se puede afirmar que el modelo genera adecuadas predicciones para clasificar a un nuevo beneficiario con respecto al apoyo económico (clase nominal) que puede llegar a recibir de algún programa de bienestar.

5 Conclusiones

Weka ofrece una combinación de facilidad de uso, flexibilidad, variedad de algoritmos y evaluación de modelos, lo que lo convierte en una herramienta valiosa para la minería de datos en una amplia gama de aplicaciones y entornos, por lo que es ideal para realizar trabajos sobre minería de datos.

Debido a la dispersión presentada en la base de datos, el algoritmo apriori el cual se encarga de proporcionar reglas de inferencia, esta parte suele ser de gran importancia pues encuentra asociaciones entre los datos que debido a la cantidad masiva de información es difícil de inferir para un sistema normal.

Se pude apreciar que al aplicar un preprocesamiento, los datos pueden ser divididos en 5 cluster, en donde predomina el programa piso firme, esto quiere decir que una vez quitado los programas que son dirigidos a asociaciones civiles o publicas, los beneficiarios suelen recibir mayor apoyo económico de este programa, además la edad de las personas que reciben este beneficio se encuentra cerca de los 40 años.

El algoritmo J48 genera una predicción razonable con un 87% de instancias correctamente clasificadas, el error medio absoluto es del 23% por ciento por lo que la dispersión no es tan alta, este valor lo alcanza tomando como clase nominal el apoyo económico por lo que si se analiza las características de las personas que viven en distintas zonas vulnerables de puebla se pude mejorar los programas a los que pertenecen los beneficiarios de esas zonas. Así, las instancias correspondientes pueden tomar decisiones para que la población obtenga un mejor beneficio de los programas a los que están inscritos.

Las bases de datos son fundamentales en los análisis de minería de datos, si una base está correctamente creada puede arrojar resultados buenos, sin embargo, la base de datos utilizada para este estudio presentó una gran dispersión, además algunos programas de bienestar presentaban apoyos económicos incongruentes como un apoyo de 4 pesos mensuales, también en la base de datos existen beneficiarios que reciben apoyos de mas de 130,000 pesos, por lo que hace que el dataset no se encuentre balanceado, esto provocó grandes dificultades al momento de obtener resultados de los algoritmos. Después de un proceso de prueba y error bastante largo, no se obtuvieron los resultados deseados, esto nos indica que la base de datos ocupada no fue creada correctamente ya que de no ser así se pude inferir que existe mucha incongruencia entre los apoyos ofrecidos por la secretaria de bienestar y los apoyos recibidos por los beneficiarios.

Como trabajo futuro se considera realizar un convenio con la secretaria de bienestar para obtener una base de datos más completa y poder realizar un nuevo proceso de análisis, comparar los resultados con los obtenidos en esta investigación y obtener mejores modelos para la toma de decisiones que ayuden a medir el impacto social de estos programas.

Referencias

[1] Secretaría de Bienestar. https://sb.puebla.gob.mx/. Accedido el 13 de mayo de 2023

[2] Han, J., Pei, J., Tong, H.: Data Mining: Concepts and Techniques. Morgan kaufmann, California (2022)

[3] Weka. Wikipedia. Accessed May 13, 2023 (2021). https://en.wikipedia.org/wiki/Weka_(machine_learning)

[4] Fern'andez-Llamazares, A., Su'arez-Seoane, S., Crecente-Romero, F.: A review of decision-making methods based on multi-criteria analysis in sustainable energy planning. Journal of Cleaner Production **318**, 128487 (2021) https://doi.org/10.1016/j.jclepro.2021.128487

[5] Martínez-Rodríguez, J., Cerrada-Serra, P.: Minería de datos como herramienta estratégica para la toma de decisiones. Revista Internacional de Organizaciones (5), 63–72 (2009)

[6] López-Carrasco, E.A., Jiménez-González, J.L., Molina-Gómez, D.: Aplicación de la minería de datos en la toma de decisiones empresariales: una revisión de literatura. Innovaciones de Negocios **33**(81), 59–78 (2020)

[7] Ramos, H., Mata-López, V., Aguilar-Ramírez, M.: Analysis of the energy sector in mexico using data mining techniques. Journal of Energy Engineering **143**(3), 04017015 (2017)

[8] Banda, L.A., Pineda, A.M.: Trend analysis and patterns of homicides in mexico using data mining techniques. Applied Sciences **9**(13), 2728 (2019)

[9] López-Ramírez, A., Morales-Menéndez, R., García-Alcaraz, J.L., González-Potes, A.: Quality air analysis in mexico city using data mining techniques. Journal of Ambient Intelligence and Humanized Computing **9**(4), 1057–1069 (2018)

[10] García-Hernández, Y.V., Sánchez-Acuña, R.: Fraudulent taxpayer identification number detection in mexico using data mining techniques. Journal of Information Systems Engineering & Management **2**(3), 15 (2017)

[11] Vázquez-Rodríguez, J.A., Rodríguez-González, A.: Analysis of crime in mexico through data mining techniques. International Journal of Intelligent Systems and

Applications in Engineering **7**(2), 55–62 (2019)

[12] Berón, E.A., Mejía, D., Castrillón, O.D.: Principales causas de ausentismo laboral: una aplicación desde la minería de datos. Información tecnológica **32**(2), 11–18 (2021)

[13] Castrillón, O.D., Castillo, L.F., Castaño, C.E.: Minería de datos aplicada a la detección de cáncer gástrico. Información tecnológica **33**(4), 151–160 (2022)

[14] Spositto, O.M., Etcheverry, M.E., Ryckeboer, H.E.L., Bossero, J.C.: Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil (2022)

[15] Quezada, M.A., Tobón-Rivera, A., Castrillón-Gómez, O.D.: Minería de datos: una aplicación para determinar cuáles factores socio-económicos influyen en el embarazo adolescente. Información tecnológica **31**(6), 53–60 (2020)

[16] Franco, E.A., Martínez, R.E.L., Domínguez, V.H.M.: Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. Revista de Educación a Distancia (RED) **21**(66) (2021)

[17] Gobierno, S.G.: Orden Jurídico Poblano. https://ojp.puebla.gob.mx Accessed 2023-06-01