Nombre del trabajo:

Proyecto DMD Etapa 1

Materia:

Datawharehouse y Minería de Datos

Docente:

Ingeniera Karens Medrano

Estudiantes:

Sermeño Zetino José Alexander SZ202008

María Dolores Martínez León ML20227

Fecha de entrega: 21-11-2021

Objetivos

- Utilizar técnicas de minería de datos para analizar una cantidad enorme de información, describir y llegar a conclusiones útiles sobre el comportamiento de los datos y explicar lo que representan en el mundo real.

- Interpretar la realidad de un fenómeno determinado a partir de información recopilada en un tiempo extenso.

- Simular una consultoría sobre temas relevantes para el MOP y generar un reporte escrito que describa los análisis realizados y las conclusiones encontradas.

- Identificar cuales son las estrategias de minería más adecuadas para resolver las consultas requeridas o que se ajusten mejor al conjunto de datos provisto para la investigación.

Marco Teórico

¿Qué es la minería de datos?

La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más.

En la última década, los avances en el poder y la velocidad de procesamiento nos han permitido llegar más allá de las prácticas manuales, tediosas y que toman mucho tiempo al análisis de datos rápido, fácil y automatizado. Cuanto más complejos son los conjuntos de datos recopilados, mayor es el potencial que hay para descubrir insights relevantes. Los comerciantes detallistas, bancos, fabricantes, proveedores de telecomunicaciones y aseguradoras, entre otros, utilizan la minería de datos para descubrir relaciones entre todas las cosas, desde precios, promociones y demografía hasta la forma en que la economía, el riesgo, la competencia y los medios sociales afectan sus modelos de negocios, ingresos, operaciones y relaciones con clientes.

¿Por qué es importante la minería de datos?

La minería de datos le permite:

- Filtrar todo el ruido caótico y repetitivo en sus datos.
- Entender qué es relevante y luego hacer un buen uso de esa información para evaluar resultados probables.
- Acelerar el ritmo de la toma de decisiones informadas.

La utilidad de Data Mining se puede dar dentro de los siguientes aspectos:

- **Sistemas parcialmente desconocidos**: Si el modelo del sistema que produce los datos es bien conocido, entonces no necesitamos de la minería de datos ya que todas las variables son de alguna manera predecibles.
- **Enorme cantidad de datos**: Al contar con mucha información en algunas bases de datos es importante para una empresa encontrar la forma de analizar "montañas" de información (lo que para un humano sería imposible) y que ello le produzca algún tipo de beneficio.
- **Potente hardware y software**: Muchas de las herramientas presentes en la minería de datos están basadas en el uso intensivo de la computación, en consecuencia, un equipo conveniente y un software eficiente, con el cual cuente una compañía, aumentará el desempeño del proceso de buscar y analizar información.

Proceso de la minería de datos

“Data Mining es una parte de un proceso de rango superior: el descubrimiento del conocimiento. Sin embargo, Data Mining es un proceso en sí mismo, que a su vez consta de varias fases.”

1. **Comprensión del negocio. Esta es la fase con la que se abre el proceso. Se encuentra enfocada en la comprensión de los objetivos y exigencias de proyecto partiendo desde la perspectiva del negocio.**
2. **Comprensión de los datos.** La fase de comprensión de datos comienza con la colección de datos inicial para continuar con las actividades que permiten alcanzar una familiaridad con ellos que permita identificar los problemas de calidad de datos.
3. **Preparación de datos.** En esta fase de preparación de datos se quieren cubrir todas las actividades necesarias para adaptar los datos origen en bruto y aproximarlos al conjunto de datos final (los datos que serán fuente de las herramientas de modelado).
4. **Modelado.** Como veremos en el próximo apartado, existen múltiples técnicas de modelado de datos, siendo en esta fase del proceso cuando, tras el conocimiento adquirido, se seleccionan las adecuadas (siempre de acuerdo a los objetivos de negocio y del proyecto) y se aplican.

En esta fase se buscan los siguientes **cuatro tipos de relaciones**:

- **Clases**: las observaciones se asignan a grupos predeterminados.
- **Clúster**: se construyen grupos de observaciones similares según un criterio prefijado.
- **Asociaciones**: las observaciones son usadas para identificar asociaciones entre variables.
- **Patrones secuenciales**: se trata de identificar patrones de comportamiento y tendencias.

1. **Evaluación.** Como resultado de la fase anterior, en esta etapa en el proyecto ya se ha construido un modelo. Para asegurarnos de que se cumple con los estándares de calidad propuestos para el proyecto es necesario evaluarlo desde una perspectiva de análisis de datos. Es decir, antes del proceder al despliegue final y su puesta en producción, es importante realizar una batería de pruebas junto con la revisión de cada paso ejecutados en la creación del modelo, que ayude a comparar el modelo obtenido con los objetivos de negocio.
2. **Despliegue o Explotación.** En esta fase se realiza la explotación y uso de los resultados del proceso de Data Mining lo que, dependiendo de los requerimientos, puede ser tan sencillo como la generación de un informe o tan complejo como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. Por lo que, en muchos casos, es el propio cliente y no el analista de datos, quien realiza la explotación.

Técnicas de minería de datos.

A continuación se describen algunas de las técnicas usadas para realizar minería de datos.

Clústeres K-means.

El **clustering** es una técnica para encontrar y clasificar K grupos de datos (clúster). Así, los elementos que comparten características semejantes estarán juntos en un mismo grupo, separados de los otros grupos con los que no comparten características.

Para saber si los datos son parecidos o diferentes el algoritmo K-medias utiliza la distancia entre los datos. Las observaciones que se parecen tendrán una menor distancia entre ellas. En general, como medida se utiliza la distancia euclidiana aunque también se pueden utilizar otras funciones.

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

1. **Inicialización:** una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
2. **Asignación objetos a los centroides:** cada objeto de los datos es asignado a su centroide más cercano.
3. **Actualización centroides:** se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo *k-means* resuelve un **problema de optimización**, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su clúster.

Las principales ventajas del método *k-means* son que es un método sencillo y rápido. Pero es necesario decidir el valor de k y el resultado final depende de la inicialización de los centroides. En principio no converge al mínimo global sino a un mínimo local.

Árboles de decisión

Los árboles de decisión son un tipo de algoritmo que clasifica la información de forma que, como resultado, se genere un modelo en forma de árbol. Se trata de un modelo esquematizado de la información que representa las diferentes alternativas junto con los posibles resultados para cada alternativa elegida. Los árboles de decisión son un tipo de modelo muy utilizado debido a que facilita mucho la comprensión de las diferentes opciones.

El árbol se compone de nodos y ramas. A su vez, existen distintos tipos de nodos y ramas en función de lo que se quiera representar. Los nodos de decisión representan una decisión que se tomará, los nodos de probabilidad representan los posibles resultados inciertos y los nodos terminales son aquellos nodos que representan el resultado definitivo.

Por otro lado, las ramas se diferencian en ramificaciones alternativas, donde cada rama lleva a un tipo de resultado y, las ramas “rechazadas” , que representan los resultados que se rechazan. El modelo se caracteriza porque un mismo problema puede ser representado con diferentes árboles.

En minería de datos, un árbol de decisión sirve para abordar problemas tales como la clasificación, la predicción y la segmentación de datos con la finalidad de obtener información que pueda ser analizada para tomar decisiones futuras.

Si trasladamos el concepto al área de análisis de negocios, los árboles de decisión se utilizan mayoritariamente para predecir las probabilidades de alcanzar un resultado en función de unas

variables de entrada tales como edad, sexo, demografía o ingresos que indicarán, por ejemplo, si el cliente es apto o no para recibir un préstamo.

Reglas de asociación.

La minería de reglas de asociación es una técnica importante en la minería de datos y consiste en encontrar las asociaciones interesantes en forma de relaciones de implicación entre los valores de los atributos de los objetos de un conjunto de datos. Numerosos y recientes estudios avalan su actualidad e importancia y su aplicación en áreas como mercadeo, bioinformática, medicina y seguridad de redes entre otras.

Esta técnica emergió en la década de los 90 con una aplicación práctica, el análisis de información de ventas para el mercadeo. Mediante ella se descubrían las relaciones entre los datos recopilados a gran escala por los sistemas de terminales de punto de venta de supermercados. Los datos consistían en colecciones de transacciones, también conocidas como bases de datos transaccionales, donde cada transacción expresa qué productos compró un cliente.

Las reglas de asociación son declaraciones de “if-then”, que ayudan a mostrar la probabilidad de las relaciones entre los elementos de datos, dentro de grandes conjuntos de datos en diversos tipos de bases de datos. La minería de reglas de asociación tiene varias aplicaciones y se utiliza ampliamente para ayudar a descubrir correlaciones de ventas en datos de transacciones o en conjuntos de datos médicos.

Las reglas no extraen la preferencia de un individuo, sino que encuentran relaciones entre un conjunto de elementos de cada transacción distinta. Esto es lo que las hace diferentes del filtrado colaborativo.

Antecedentes.

El MOP es una institución gubernamental con una larga historia y que ha ido a través del tiempo ganando una gran cantidad de funciones, fue creada en 1905 bajo el nombre de Cuerpo de Ingenieros Oficiales. A esta oficina le correspondía la Dirección General de Obras Públicas como dependencia directa del Ministerio de Fomento, con la salvedad de que los trabajos de caminos eran realizados por el Ministerio de Gobernación.

En 1916 el Poder Ejecutivo considerando la necesidad urgente de poseer buenas vías de comunicación, emitió el Decreto de creación de la Dirección General de Caminos, la cual funcionaría como una entidad técnica – consultiva, anexa al Ministerio de Gobernación y Fomento, la cual tendría a su cargo todo lo relacionado con las vías de comunicación de la República, puentes y obras que tengan relación con éstas.

Fue hasta en 1917, que se emite un Decreto Legislativo de creación del Ministerio de Fomento y Obras Públicas, la cual posteriormente asumiría todas las funciones encomendadas a las anteriores oficinas de regulación vial.

En 1920, la Dirección General de Obras Públicas dentro del ramo de Fomento contaba con una Sección de Caminos, así como una Sección de Arquitectura, Saneamiento y Aguas y una Sección de Caminos, Puentes y Calzadas.

En 1948, El Ministerio de Fomento y Obras Públicas contaba con la Dirección General de Carreteras.

En 1954, la Dirección de Urbanismo y Arquitectura y la Dirección de Caminos, se convierte en Direcciones Generales dentro del Ramo de Obras Públicas. Todos estos cambios son producto de la necesidad de ordenar el crecimiento de las ciudades, tanto en su parte arquitectónica como en infraestructura, por lo cual se le encomiendan las funciones específicas de construir, mantener y rehabilitar la infraestructura urbana y vial del país, en esta última se incluyen las carreteras interurbanas, rurales y urbanas; las cuales se constituyen en uno de los pilares que sostiene la economía nacional.

Actualmente el ministerio de obras públicas, dentro de su organización cuenta con dos viceministerios: de Transporte, el cual se encarga de la reglamentación del tráfico, tanto rural como urbano, así como de los transportes aéreos, terrestre y marítimos; y de obras públicas, que es el encargado de dirigir la planificación, construcción, rehabilitación, reconstrucción, ampliación, expansión y mantenimiento de la infraestructura vial del país.

A partir de esto podemos considerar que conocer el tráfico vehicular y el numero de infractores es importante para el MOP para poder decidir o buscar soluciones estratégicas para reducir el número de accidentes de transito, decidir horarios de mantenimiento vial, prioridad en el uso o disponibilidad de vías y carreteras entre otros.

Situación Actual

En estos momentos el MOP requiere un análisis de los datos de 2018 sobre el parque vehicular y las infracciones, por lo que para entrar en el contexto de la consultoría debemos investigar cuales son los conocimientos actuales sobre las variables a investigar.

Aumento del parque vehicular en el año 2017.

Hasta el 20 de diciembre de 2017, el Registro Público de Vehículos Automotores del Viceministerio de Transporte (VMT) reportó 1,091,027 unidades en todo el país, lo que significa que el parque vehicular aumentó en casi 83,000 automotores durante el año, de los cuales el 46.42 % fueron motocicletas.

El año 2016 cerró con 1,008,078 unidades registradas, que fueron 82,630 vehículos más que los registrados en 2015. Hasta el 20 de diciembre, el VMT ya reportaba un incremento de 82,949 vehículos, que equivalen a un promedio diario de 234.

En el caso de las motocicletas, en 2016 había registradas 210,030 y para 2017 llegaron a 248,542, es decir, un aumento de 38,512 en comparación a 2016.

Por lo que a partir de los datos entregados por el MOP podríamos establecer el aumento en cantidad y porcentaje del aumento de vehículos en el año 2018 además de verificar la calidad de datos obtenidos, ya que si el total es menor a lo reportado en el 2017 podríamos estar lidiando con datos de baja calidad o con un decremento de parque vehicular lo cual no resulto lógico.

El conocimiento del parque vehicular nos permitirá hacer recomendaciones y proyecciones de crecimiento con las cuales será posible asignar el presupuesto para proyectos de construcción y mantenimiento de vial a las vías que lo requieren según la prioridad de transito y densidad vehicular de la zona.

Esquelas en el año 2017

Según los datos recopilados en el año 2017 se asignaron 375,001 esquelas por infracciones viales en los periodos entre enero y diciembre, dicha información puede observarse en la tabla siguiente.

La tabla contiene además variables importantes que deberán ser consideradas dentro de los análisis para el negocio, la clasificación del nivel de gravedad de las esquelas, las causas de la accidentalidad, la cantidad de esquelas y los periodos de tiempo en el que las esquelas son asignadas con mayor frecuencia.

A partir de la información recopilada será posible saber si hay algún tipo de disminución en las faltas viales, accidentes o de emisión de esquelas en general o si existe un aumento. Así como sí hay aumentos en el nivel de gravedad de las faltas incurridas por los conductores y finalmente concluir en recomendaciones o medidas necesarias para reducir el número de faltas.

ACCIDENTABILIDAD A NIVEL NACIONAL

PERIODO DEL 01 DE ENERO AL 31 DE DICIEMBRE DE 2017

RUBROS / AÑOS	2017
ACCIDENTES DE TTO	21582
LESIONADOS	9462
FALLECIDOS	1245

CAUSAS PRINCIPALES QUE PROVOCAN ACCIDENTES DE TTO

1ª	DISTRACCION DEL CONDUCTOR
2ª	INVADIR CARRIL
3ª	NO RESPETAR SELALES DE PRIORIDAD
4ª	NO GUARDAR DISTANCIA REGLAMENTARIA
5ª	VELOCIDAD INADECUADA
8º	ESTADO DE EBRIEDAD

ESQUELAS 2017

DETALLES	2017
LEVES	172,708
GRAVES	75,313
MUY GRAVES	126,980
TOTAL	375,001

Metodología de minería de datos.

La metodología de minería de datos que será utilizada es la conocida como KDD. KDD es una metodología propuesta por Fayyad et al. en 1996, que propone 5 fases: selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.

El Descubrimiento de conocimiento en bases de datos (kdd, del inglés Knowledge Discovery in Databases) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente procesar los datos, hacer minería de datos (data mining) y presentar resultados.

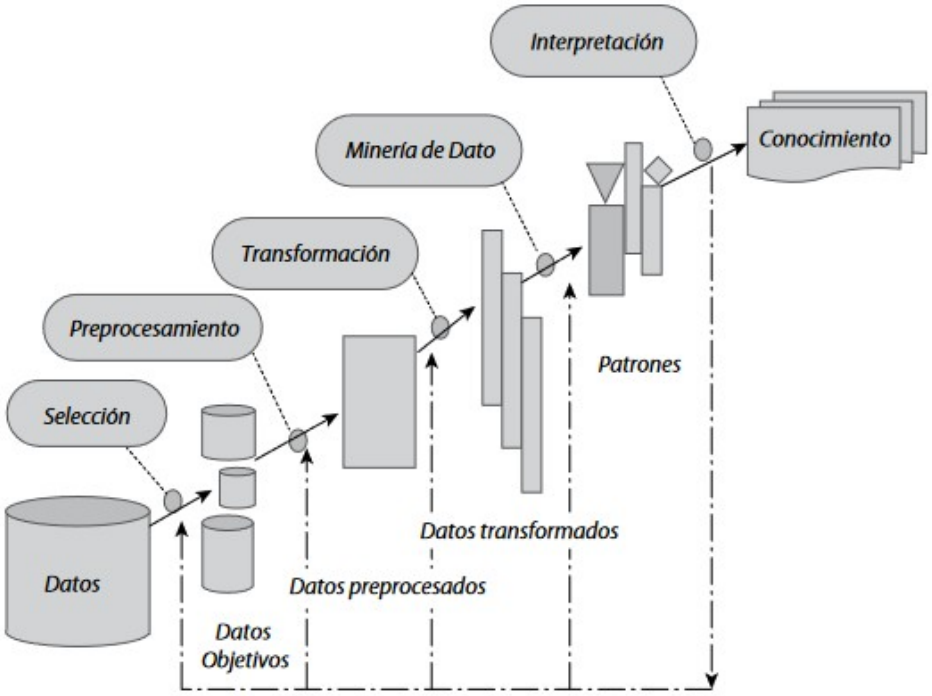
KDD se puede aplicar en diferentes dominios, por ejemplo, para determinar perfiles de clientes fraudulentos (evasión de impuestos), para descubrir relaciones implícitas existentes entre síntomas y enfermedades, entre características técnicas y diagnóstico del estado de equipos y máquinas, para determinar perfiles de estudiantes "académicamente exitosos" en términos de sus características socioeconómicas y para determinar patrones de compra de los clientes en sus canastas de mercado.

Etapas del proceso KDD

El proceso kdd que se muestra en la figura 1 es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones.

Se resume en las siguientes etapas:

- Selección. Una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso kdd, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo.
- Preprocesamiento/limpieza. Se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo.
- Transformación/reducción. Se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos
- Minería de datos (data mining). El objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación, patrones secuenciales, asociaciones y otras.
- Interpretación/evaluación. En la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones.


Interpretación
Conocimiento
Minería de Dato
Transformación
Preprocesamiento
Selección
Patrones
Datos transformados
Datos
Datos preprocesados
Datos
Objetivos

Formulación del problema.

En esta investigación se formularan algunos problemas relacionados con el origen de los datos y las variables que se requieren conocer para obtener las conclusiones deseadas y más útiles para los interés del MOP.

Con respecto a los datos de esquelas o faltas de trafico cometidas en enunciado del problemas es el siguiente:

El estudio pretende sintetizar y determinar cuales han sido las 10 faltas viales más frecuentes durante el año 2018, así determinar clasificar las faltas documentadas según su grado de gravedad, y de poseer la información necesaria el monto monetario recolectado con el pago de las esquelas así como la accidentalidad del reportada para dicho año.

A mismo se debe plantear un problema para el parque vehicular, por lo que el enunciado es el siguiente:

Se verificará el crecimiento del parque vehicular durante el año 2018, se estimara el porcentaje de aumento o decrecimiento con respecto a las estadísticas disponibles del año 2017, se fuera posible se segmentarán los vehículos por tipo y marca y se hará un clasificación según estas variables.

Factibilidad

El proyecto consiste en encontrar patrones en la información obtenida por el MOP respecto a las esquelas y el parque vehicular en El Salvador en 2018, utilizando técnicas de minería de datos y presentación de informes de negocios sobre las variables de interés.

Se prevé caracterizar la cantidad de esquelas anuales por categorías según la gravedad de las mismas con el propósito de detectar incrementos o decrementos así como buscar maneras de disminuir el número de faltas reportadas. También se busca encontrar un patrón en el crecimiento del parque vehicular.

A partir de esto se harán las recomendaciones y conclusiones pertinentes para que el MOP sea capaz de tomar las mejores decisiones posibles.

Los datos a utilizar abarcan todo El Salvador, por lo que son bases de datos grandísimas, se limitará a los datos entregados en los archivos de texto, y el trabajo será finalizado en diciembre de 2021.

Se requiere con equipos computaciones que cumplan los requerimientos de los software a utilizar, se requiere mano de obra con suficiente conocimiento para generar los modelos de minería requeridos, y finalmente se requiere acceso a los software a utilizar, se prevé utilizar solamente software gratuito o con licencias de pruebas.

Se considera que el proyecto tendrá un impacto positivo para la sociedad y la administración de las vías terrestres y carreteras, ya que las autoridades tendrán información de buena calidad que garantice la planificación de proyectos, mantenimiento y construcciones en el futuro.

Por estos factores se considera que llevar a acabo el proyecto es factible, ya que se tiene la mano de obra necesaria, los equipos, tiempo y recursos económicos necesarios para el proyecto.

Justificación

La consultoría pretende estudiar el comportamiento de faltas viables y crecimiento del parque vehicular en El Salvador, por medio de la utilización de técnicas de minería de datos, bases de datos e inteligencia de mercado para generar recomendaciones y conclusiones útiles que sirvan de guía para la planificación estratégica de obras, mantenimiento, políticas y construcciones viales.

Con el objetivo de comenzar planes de reordenamiento urbano y buen convivencia social, es necesario generar una imagen realista y objetiva de las variables pertinentes en el comportamiento de la población automovilista, por un lado es necesario encontrar un patrón en el crecimiento para estimar hasta que punto este podrá ser sostenible o deberá ser desacelerado por medio de políticas publicas que desincentiven la compra de vehículos y por otra parte debemos conocer cuales son las principales faltas en las que incurren los conductores para generar programas efectivos de educación vial o penas juridicas para tratar de reducir la incidencia de estas faltas.

Importancia

Existen muchos problemas relacionadas al crecimiento vehicular y las faltas al volante, pero el más visible de todos y el que afecta a la mayoría son las congestiones viales, este es un problema que continua empeorando a medida que pasa el tiempo y que la población y le numero de vehículos crece, este disminuye la calidad de vida y tiene impactos económicos directos en los sectores tanto comerciales, económicos y civiles.

Los tiempos de viajes incrementan así como los costos de consumo de combustible y otros costos operativos y sumando al negativo impacto ambiental.

Debemos sumar factores externos como el mal estado de las calles y caminos que agravan considerablemente la situación, la mala administración publica que descuida y permite un deterioro avanzado de la infraestructura.

Por estos y otros factores se vuelve importante estudiar y buscar patrones de comportamiento en los hábitos de los conductores y el incremento en la cantidad de vehículos, ya que con la cantidad correcta de información es posible generar proyecciones y planes más certeros con los que sea posible mitigar los efectos negativos del trafico vehicular y mejorar la calidad de vida de los habitantes.

Alcances

- El proyecto de minería de datos está limitado a los datos entregados por el MOP para llevar a cabo la consultoría, por lo que no se investigará más ni se buscará información complementaria que pudiera sesgar las conclusiones alcanzadas por el estudio.
- Dado que el software es gratuito no se podrán implementar técnicas más avanzadas que las permitidas por las limitaciones de la licencia en los paquetes de software usados, ya que muchas empresas ofrecen paquetes de pruebas gratuitos con algunas restricciones técnicas.
- Los limites de la investigación están definidos por el territorio de El Salvador y los datos que fueron recabados en el año 2018.
- Se espera que las recomendaciones y conclusiones generadas con el estudio sean suficientemente certeras para detectar patrones en los datos analizados y que dichos patrones se apeguen a la realidad de El Salvador y que las proyecciones o decisiones que sean tomadas con esta información sean las más adecuadas, sin embargo pudieran haber otros eventos o datos no relacionados con las fatas viales o el crecimiento vehicular que limiten la aplicabilidad de las conclusiones, como por ejemplo crisis económicas, catástrofes naturales y otro tipo de eventos impredecibles que modifiquen bruscamente el comportamiento de la población.

Limitaciones

- El tiempo para realizar los análisis está limitado con la duración del ciclo II del año 2021 y limitado por las demás actividades académicas a realizar.
- Los equipos de computación personales podrían no estar a la altura en desempeño de lo sutilizados por empresas y profesionales dedicados a la minería de datos.
- El software utilizado es de licencia libre o gratuita por lo que podría no contar con todas las características o soporte con el que cuentan las versiones de pago.
- Los recursos económicos son limitados y escasos.
- Si los datos recopilados por el MOP son erróneos o presentan inconsistencias pueda que no sea posible generar recomendaciones acertadas.

Planificación de recursos

- Están disponibles las computadoras personales de cada uno de los integrantes del grupo.
- Los software a utilizar son RapidMiner, Visual Studio, SQL Server, Power BI, todos en su versión gratuita o de prueba.
- EL grupo de consultores consiste de tres personas que realizaran los análisis que sean requeridos de manera coordinada y en grupo.
- Software de video conferencias como Meets, Teams, etc. estará disponible para coordinar los trabajos en grupo y realizar reuniones virtuales con el cliente y asesores.
- GitHub será la herramienta de versiones utilizada durante el proyecto.
- Se planea que cada integrante pueda al menos emplear tres horas diaria para la realización del proyecto, esto incluye tareas de redacción, investigación o análisis de minería.

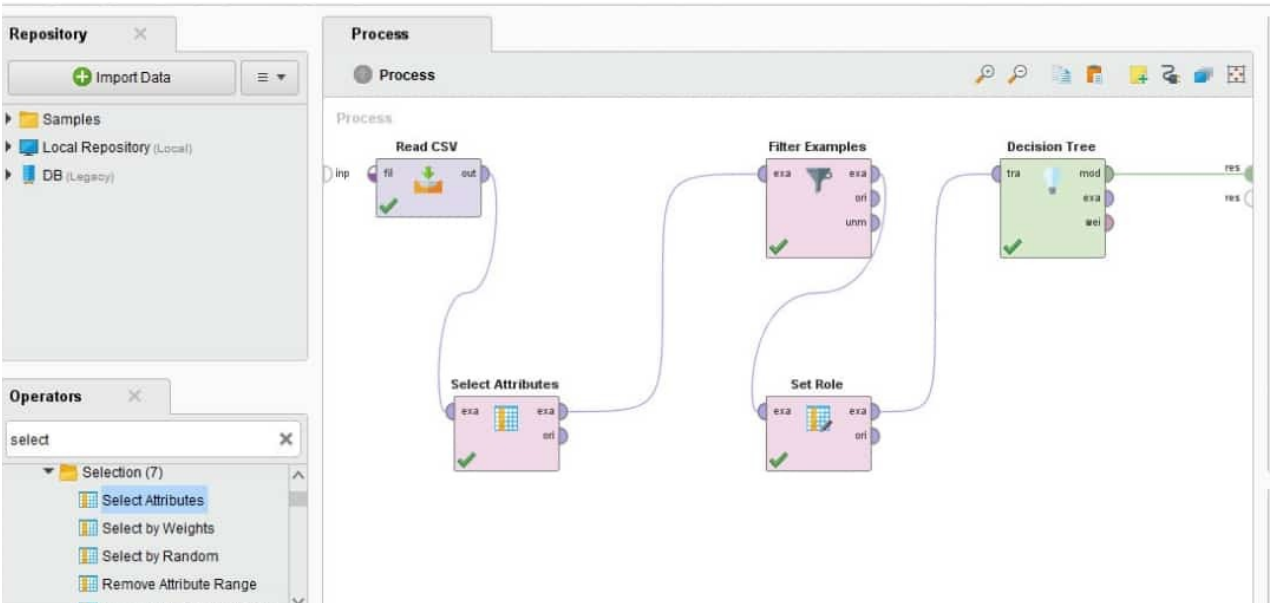
Desarrollo de los análisis Parque vehículo.

1. **Selección de datos.**

   Los datos seleccionados fueron aquellos recopilados en el documento de parque vehicular provisto por el MOP, este documento presenta algunos campos vacíos y columnas irrelevantes para el proceso de extracción de datos.

2. **Preprocesamiento.** Las etapas del proceso ETL necesarias para preparar el modelo fueron la implementación de filtros para eliminar las filas con valores perdidos, así como la selección relevante para el análisis.

3. **Transformación.** Se utilizó un operador para fijar una columna como etiqueta para producir los nodos principales del árbol de decisión.

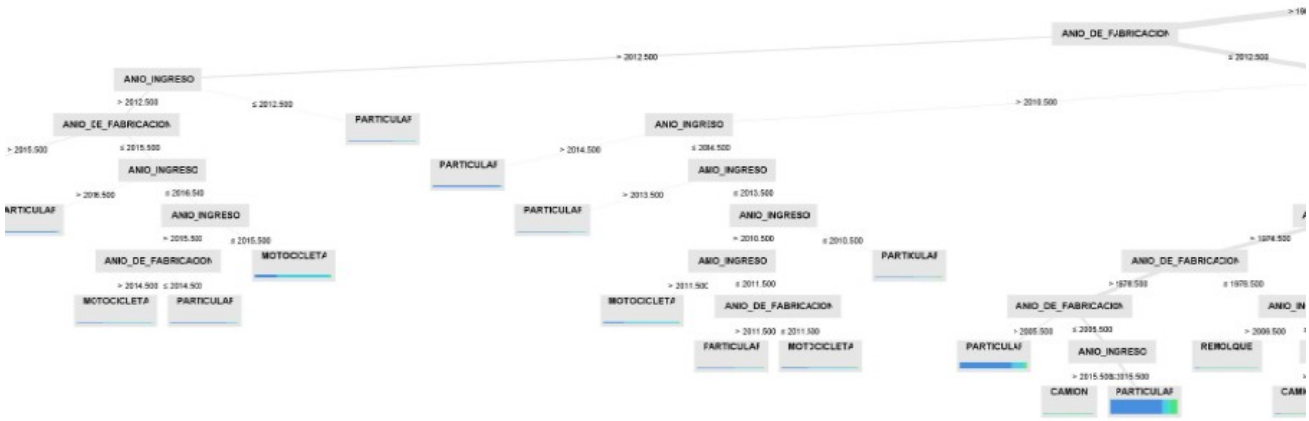


4. **Minería de datos.** Se analizó la implementación de varios modelos de minería de datos, sin embargo por el tipo de datos y por las limitaciones computacionales se optó por generar un modelo de arbol de decisión.

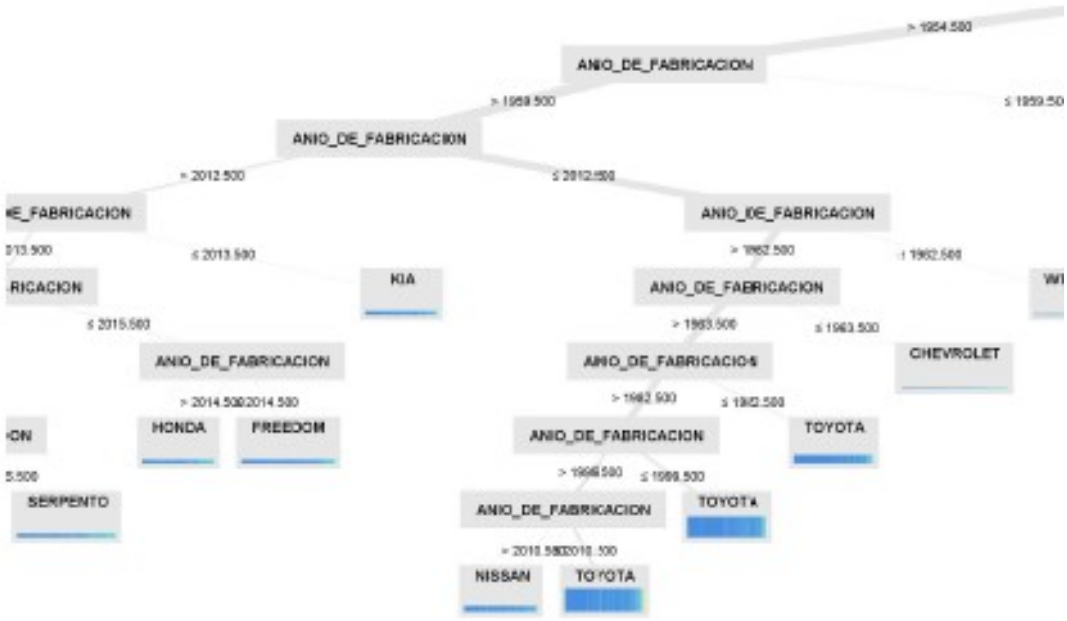
5. **Interpretación.** A continuación se presentan los resultados de los diferentes procesos de minería realizados.

En el árbol 1. se generó con el objetivo de conocer cual es el estado del automóviles ingresados al parque vehículo, en su mayoría se observó que la mayor parte de los vehículos fueron ingresados como usados. También se encontró que el tipo de vehículo de mayor circulación son los automóviles.

También se comparó el año de fabricación con el tipo de vehículos para encontrar algún patrón entre los ingresos, fecha de fabricación y el tipo de vehículo, se encontró de nuevo que el tipo de vehículo introducido con mayor frecuencia fue el automóvil y que la mayor cantidad de ingresos fue de carros fabricados en el 2005. Se observó una tendencia al aumento significativo de del ingreso de motocicletas a partir del año 2015. También se observó que muchos automóviles fabricados en 2015 fueron ingresados en 2015.



Por ultimo el último árbol generado con respecto a las marcas y los años de fabricación encontró que la marca preferida por los consumidores ha sido por mucho tiempo TOYOTA, siguiéndole otras marcas como NISSAN, de motocicletas se encontraron marcas más recientes como Serpento y clásicas como HONDA.



Desarrollo de los análisis Esquelas

2. **Selección de datos.**

   Del archivo recibido se ignoraron algunas columnas donde los valores de la mayoría eran ceros y provocaban errores de lectura.

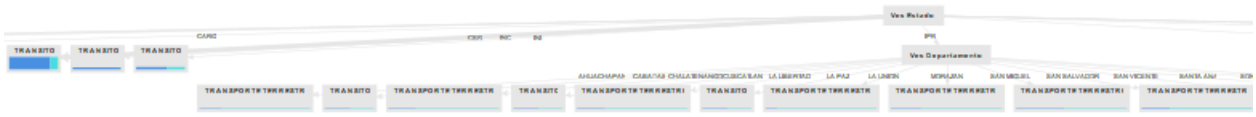
2. **Preprocesamiento.** AL utilizar el operador de lectura para el archivo CVS, se seleccionó la opción de sustituir los valores que provocan errores como valores perdidos, estos luego serían filtrados para poder utilizar los modelos de minería de datos.

3. **Transformación.** Se utilizó un operador para fijar una columna como etiqueta para producir los nodos principales del árbol de decisión.

4. **Minería de datos.** Con este modelo se trató de implementar K-means y clusters, pero solamente fue posible obtener un cluster, es decir el modelo no presenta tanta variabilidad para obtener información relevante de dicha técnica, por otra parta también se trato de realizar análisis por reglas de asociación, pero por la cantidad de datos utilizados no fue posible completar el análisis, ya que los recursos computacionales fueron insuficientes para completarlo. Por lo que se optó en aplicar de nuevo el método de arboles de decisión.

5. **Interpretación.** A continuación se presentan los resultados de los diferentes procesos de minería realizados.

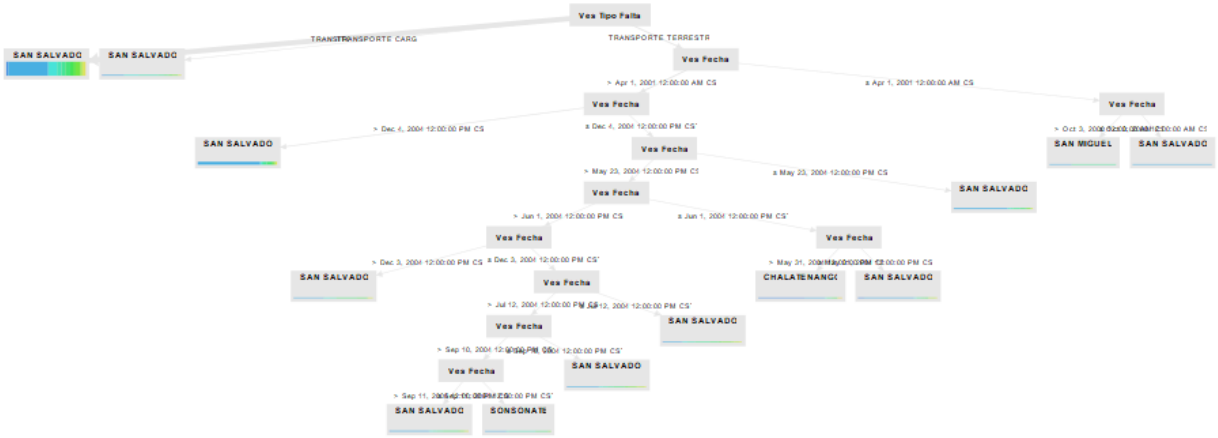
A partir de un árbol generado por todas las columnas se observó que el mayor numero de infracciones por tipo fue el de Transito, y usando el monto por esquela el mayor numero de faltas fue grave.



Según el estado de las esquelas y según el departamento que se impusieron la mayor parte de las esquelas fueron canceladas, un porcentaje mucho mejor y al menos un 10 por ciento de las mismas era invalida.

Los tipos de esquela variaban según el departamento, y se puede interpretar como que los departamentos con mayor transito pesado y de carga presentan mayores cantidades de esquelas de transporte terreste, este es el caso de departamentos como La Unión, Morazan y San Miguel.

Por último el departamento donde han ocurrido la mayoría de faltas ha sido San Salvador, y esto es considerablemente notorio en el año 2018.


Vex Tipo Falta
SAN SALVADO
SAN SALVADO
TRANSTRANSPORTE CARG
TRANSPORTE TERRESTR
Vex Fecha
> Apr 1, 2001 12:00:00 AM CS
a Apr 1, 2001 12:00:00 AM CS
Vex Fecha
Vex Fecha
> Dec 4, 2004 12:00:00 PM CS
a Dec 4, 2004 12:00:00 PM CS'
SAN SALVADO
> Oct 3, 2005 12:00:00 AM CS
SAN MIGUEL
SAN SALVADO
Vex Fecha
> May 23, 2004 12:00:00 PM CS
a May 23, 2004 12:00:00 PM CS
Vex Fecha
SAN SALVADO
> Jun 1, 2004 12:00:00 PM CS
a Jun 1, 2004 12:00:00 PM CS'
Vex Fecha
Vex Fecha
> Dec 3, 2004 12:00:00 PM CS
a Dec 3, 2004 12:00:00 PM CS'
SAN SALVADO
Vex Fecha
> May 31, 2005 12:00:00 PM CS
CHALATENANGO
SAN SALVADO
> Jul 12, 2004 12:00:00 PM CS
a Jul 12, 2004 12:00:00 PM CS'
Vex Fecha
SAN SALVADO
> Sep 10, 2004 12:00:00 PM CS
a Sep 10, 2004 12:00:00 PM CS'
Vex Fecha
SAN SALVADO
> Sep 11, 2004 12:00:00 PM CS
a Sep 11, 2004 12:00:00 PM CS'
SAN SALVADO
SONSONATE

Recomendaciones.

Las recomendaciones que haremos después de las investigación son las siguientes:

1. Se debe de establecer una campaña de educación y consciencia vial, especialmente fuerte en las zonas urbanas como San Salvador ya que a medida que le parque vehículo incrementa siguen incrementando el número de faltas reportadas.
2. Es necesario generar un o mecanismo de almacenamiento de datos enfocado en el apredizaje y solución de las causas del problema del tránsito, ya que no hay un estándar en las causas o decripciones de las faltas, y esto podría mejorar el proceso de analisis.
3. Se recomienda regular o supervisar mejor las zonas geográficas con mayor transito pesado, ya que podría requerirse que ciertos sectores de transporte sean multados o requieran mayor cantidad de horas de entrenamiento antes de utilizar vehículos pesados.
4. Se necesitan leyes enfocadas en prevenir el abuso mientras se maneja.

Las recomendaciones que se hacen con respecto al incremento del parque vehicular son las siguientes:

1. De igual manera se debe estandarizar y simplificar la manera en que se toman los datos, ya que se observaron inconsistencias en algunos atributos como el numero de cilindros por ejemplo, por lo que es necesario implementar el uso de formularios que mejores la manera de captar los datos y que generen almacenes de datos más útiles.
2. Es necesario promover el uso de motocicletas, la población ha mostrado un interés en estas de manera natural, pero dado que el parque vehículo ha crecido especialmente en desde el año 2005, es necesario promover el uso de vehículos más compactos y que emitan menos gases contaminantes, esto podría lograrse por medio de un subsidio o excepción fiscal para los motociclistas.
3. Se debe de regular el año de entrada de los vehículos al país ya que ingresas vehículos con vidas útiles muy cortas, usados y que provocan el incremento desmedido del parque vehicular y de accidentes de trafico que como ya vimos se reportó una cantidad considerable en el año 2008.

Referencias.

1. *Marco Institucional*. (2020, diciembre 2). Gob.sv. https://www.mop.gob.sv/marco-institucional/
2. *¿Qué es la minería de datos?* (2021, noviembre 5). Sas.com. https://www.sas.com/es_mx/insights/analytics/data-mining.html
3. Benito, F. V. (s/f). *Minería de datos y aplicaciones*. Uc3m.es. Recuperado el 12 de noviembre de 2021, de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/22.pdf
4. de CEUPE, B. (2019, febrero 9). *Proceso del Data Mining*. Ceupe.com; CEUPE. https://www.ceupe.com/blog/proceso-del-data-mining.html
5. *kmeans*. (s/f). Unioviedo.es. Recuperado el 12 de noviembre de 2021, de https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html
6. Conecta Software. (2020, enero 14). *Arboles de decisiones en la minería de datos - Conecta Software*. Conectasoftware.com. https://conectasoftware.com/analytics/arboles-de-decisiones-en-la-mineria-de-datos/
7. González, A. Y. R., Trinidad, J. F. M., Ochoa, J. A. C., & Shulcloper, J. R. (s/f). *Minería de Reglas de Asociación sobre Datos Mezclados*. Inaoep.mx. Recuperado el 12 de noviembre de 2021, de https://ccc.inaoep.mx/portalfiles/file/CCC-09-001.pdf
8. El Salvador-La Prensa Gráfica, N. de. (2018, enero 1). *Parque vehicular aumentó casi 83,000 automotores*. Noticias de El Salvador - La Prensa Gráfica | Informate con la verdad. https://www.laprensagrafica.com/elsalvador/Parque-vehicular-aumento-casi-83000-automotores-20171231-0219.html
9. (S/f). Recuperado el 12 de noviembre de 2021, de http://file:///C:/Users/alexz/AppData/Local/Temp/ACCIDENTABILIDAD_GENERAL_.pdf
10. Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490