Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

MINERÍA DE DATOS

Avance I – Proyecto Integrador

Licenciatura en Actuaría

Equipo 10

Nombre	Matrícula
Magdaly Rodríguez Ortiz	1815330
Evelyn Lizbeth Trejo Rodríguez	1811917
Alexis Hernández Morales	1887948

Profesor: Mayra Cristina Berrones Reyes

Grupo: 002 (miércoles de 7:00 p.m. a 10:00 p.m.)

Ciudad San Nicolás de los Garza – 27 de octubre del 2020.

1) Título de la base de datos

a) Nombre con el que se encuentra en la página encontrada.

Sanfranciso Crime Dataset

b) URL de la página.

https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset

2) Descripción de los datos

a) Qué tipo de datos son.

Es un conjunto de datos organizados por tablas que contienen información sobre los incidentes del departamento de policía de la ciudad de San Francisco en el año 2016. El archivo contiene la información sobre las tasas de criminalidad en diferentes regiones de San Francisco con algunos otros aspectos importantes relacionados con la delincuencia.

b) Descripción de las columnas.

IncidntNum: número del incidente, siendo un dato numérico.

Category: en esta columna se describe la categoría del delito, que en este caso su escala de medición es de tipo nominal, en donde ya sea no criminal, asalto, robo, persona desaparecida, etcétera se puede ir clasificando.

Descript: en este apartado se encuentra la descripción del crimen, ya sea un robo, asalto, fraude, etcétera, te describe como aconteció o de qué manera sucedió el crimen siendo de tipo texto largo.

DayOfWeek: día de la semana en el que ocurrió el delito, siendo categórico de tipo texto.

Date: es la fecha precisa (dd/mm/aaaa) en la que sucedió el crimen, siendo de tipo tiempo.

Time: hora exacta en la que se reportó el crimen, su medición es de tiempo.

PdDistrict: distrito de la ciudad donde ocurrió el crimen, el cual su escala de medición es de tipo nominal.

Resolution: en este apartado encontramos el tipo de castigo que recibe el autor del crimen para resolver el caso, ya sea arresto, nada, u otro, siendo este apartado de tipo nominal.

Address: en dicha columna se encuentra la dirección precisa donde ocurrió la escena del crimen, siendo una variable de texto.

X: latitud de la locación del crimen, siendo una variable categórica.

Y: longitud de la locación del crimen, siendo una variable categórica.

Location: coordenadas exactas de la ubicación, siendo una variable categórica.

PdId: Pd ID, siendo una variable de tipo numérica.

3) Justificación del uso de datos.

a) ¿Cuáles fueron las características que les llamó la atención de los datos? ¿Qué les hizo querer trabajar con ellos?

Lo que consideramos interesante al encontrarnos con dicha base de datos es que nos presentaba distintas columnas de información valiosa que podemos trabajar con distintas técnicas, para así erradicar alguna problemática; lo que termino de convencernos a trabajar con dichos datos fue su tema central que son los crímenes en la ciudad de San Francisco, ya que esta información sería de gran utilidad tanto para la policía de la ciudad, como para los ciudadanos de cada uno de los condados.

b) ¿Qué beneficio encuentran de trabajar con estos datos?

Tenemos el beneficio de que, a través de los datos proporcionados por el archivo, al ser muy precisos y completos, podemos encontrar relaciones entre ellos y plantearnos diferentes problemáticas a las cuales les podremos dar solución utilizando diferentes técnicas ya sea visualización, clustering, etcétera, ya que tenemos suficiente información de donde obtener.

4) Planteamiento del problema.

Tomando en cuenta las características utilizadas para la tarea de análisis de bases de datos, elabora una problemática que te gustaría resolver con tu investigación.

Somos una empresa asociada al gobierno de San Francisco encargada de brindar información estadística significativa, enfocada en el área de seguridad; debido a que San Francisco a lo largo de estos años ha tenido una tasa de criminalidad general de 151% más alta que el promedio nacional y el número de crímenes anuales se han ido incrementando, las autoridades estatales se han alarmado debido a que la manera en la han distribuido sus unidades no ha sido eficaz, por lo nos interesa realizar un estudio que nos brinde información útil para asistir al gobierno a disminuir esa tasa de criminalidad.

Nota: Como lo comentado en clase, traten de pensar como una empresa, y cuáles son los problemas que les gustaría mejorar al finalizar su proyecto.

5) Objetivo Final.

a) Explicar a detalle cual es el objetivo principal (y secundarios en el caso de existir) para trabajar con este tipo de datos.

Objetivo Principal:

La detección temprana de los lugares del delito es importante para que la ciudad de San Francisco pueda tomar decisiones preventivas que permitan aumentar la percepción de la seguridad pública.

Sabemos que las actividades criminales están distribuidas aleatoriamente sobre un espacio geográfico, sin embargo, tienden a concentrarse en ciertos distritos por razones como escasez de vigilancia, etcétera, es decir, tienden a existir puntos calientes de crimen en donde el número de incidentes de delito está por arriba del promedio, entonces nuestro objetivo principal será diseñar un método de clasificación para predecir de forma automática si un lugar específico en la ciudad de San Francisco será un centro de delincuencia o no, brindando al gobierno información útil para que tengan una mejor distribución de sus unidades policiacas, dado cierto día, hora y lugar.

Objetivo secundario:

Predecir la cantidad de crímenes a 1 año en los que se verán involucrados el robo de casa o auto, y negociar esta información a una aseguradora ya que, dada la seguridad en cierto distrito, podrán recalcular el deducible del seguro.

Nota: Los objetivos secundarios nos servirán para tener varias líneas de investigación abiertas, en caso de que el objetivo principal elegido no pueda completarse.

6) Planeación de la herramienta a utilizar.

a) Observando el tipo de datos que se tiene, describir cual es el tipo de técnica que se planea utilizar y dar una explicación concisa de por qué se va a trabajar con esa técnica.

Para nuestro objetivo principal la técnica que más se adapta a nuestra necesidad es Clustering ya que, es una herramienta fundamental del proceso de análisis avanzado de los datos porque nos va a permitir segmentar y crear particiones de nuestro gran volumen de información, es decir, de nuestra base de datos de San Francisco, crearemos grupos relacionando la cantidad de crimines por distritos, días y hora , es decir, serán grupos más pequeños, interpretables, y diferenciables entre sí, esto nos permitirá ver relaciones que a simple vista nos es difícil observar.

Para el objetivo secundario planeamos utilizar regresión ya que esta técnica de minería de datos es de la categoría predictiva, es decir nos predice el número de crímenes de las categorías que nos interesan, basándose en los datos anteriores de otros atributos.

b) Si ya se tiene observado algún algoritmo o herramienta dentro de esta técnica seleccionada, menciona cuál es, y por qué se desea trabajar con ella.

En el objetivo principal planeamos utilizar el algoritmo de K-medias ya que utiliza las distancias entre puntos y agrupa los clústeres según lo cerca que están de los centros de gravedad de los clústeres. A priori, tenemos que saber cuántos clústeres vamos a crear para que el algoritmo sepa dónde colocar cada punto, obteniendo ese valor de k contemplando todos y cada uno de nuestros distritos.

Pasos para el método K-medias:

1. Centroides: elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada clúster.
2. Distancias: analizamos la distancia de cada dato al centroide más cercano, perteneciendo a su clúster.
3. Media: obtener media de cada clúster y este será el nuevo centro.
4. Iterar: repetimos el proceso hasta que los clústeres no cambien.

Planeamos utilizar dicho algoritmo ya que es con el que se trabaja al utilizar la técnica de k-medias.

Para el objetivo secundario, se podría utilizar una regresión simple para obtener las estimaciones de cada uno de los incidentes en base al tiempo.

Algoritmo por utilizar:

La estimación de $y = \beta_o + \beta_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza:

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1} x$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

El algoritmo de estimación que consideramos utilizar se dio, ya que, elegimos una regresión simple.