UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Licenciatura en Actuaría

Minería de datos

Grupo: 003

Mtra. MAYRA CRISTINA BERRONES REYES

Evidencia: Análisis de Bases de Datos

Alumno: Ulises Solis Moises

Matrícula: 1887850

15/10/2020

Nombre de base de datos: Google Play store Apps

Objetivo: Mejorar la manera en cómo se representan los ratings de las aplicaciones para que sean más fiables.

Problema planteado: El problema de los ratings es que no son muy representativos, puesto que el usuario nada más observa la cantidad de estrellas y solo si se pone a investigar puede observar cuántas personas fueron las que reseñaron. Esto es un problema puesto que una aplicación con únicamente 2 ratings de 5 estrellas y un total de descargas de 1 millón, tendrá un rating de 5 estrella cuando podría darse el caso de que esto no represente el sentimiento real que se tenga. O su contra parte, que tenga 100mil reseñas para 150 mil descargas habla de que su resultado es más real a lo que la gente piensa. Entonces ese es el inconveniente, que no se tenga una variable que te mida la relación entre una y otra.

Solución: Usando la técnica de regresión lineal se plantea la solución de crear una variable que nos pueda demostrar el nivel de confianza que se tiene y la correlación que tiene una cantidad de ratings comparada con su cantidad de descargas. Podemos utilizar el coeficiente de correlación y se podría generar una tabla en la cual se tengan unos intervalos en los cuales se pueda medir de manera más simple para el usuario.

Nombre de base de datos: Novel Corona Virus 2019 dataset

Objetivo: Revisar cómo fue la propagación del virus, para poder prepararnos para una futura pandemia.

Problema planteado: La propagación del virus es bastante complicada de seguir puesto que no se tienen registros al momento y pararlo es muy difícil. Por tanto, con la serie de tiempo de casos y como se fueron propagando podemos preparar nuestro país o en su caso preparar al mundo para una futura pandemia y saber cómo detectar la propagación y detenerla y así ejecutar un plan de manera mucho más ordenada y sin tantas improvisaciones.

Solución: La técnica que deberíamos revisar no sería solo una, puesto que primero tenemos que usar el clustering para dividir por lugar los contagios y fechas, posteriormente podemos tomar como un outlier el primer contagio de una persona en una ciudad posterior a la primera y que siga una continuidad de fecha; también deberíamos utilizar la detección de patrones secuenciales para poder revisar los factores que pueden servirnos para lo último que sería la predicción para que, basándonos en el tiempo y ubicación geográfica, podamos predecir la rapidez de propagación y los factores principales que harían que esta pase, para así deternerla de una manera mucho más efectiva.

Nombre de base de datos: Wine reviews

Objetivo: Generar un mapa del mundo, y por país, que muestre la relación calidad/región del vino, que también incluya el precio y descripciones de sabor.

Problema planteado: El problema que las personas viajeras tienen es que normalmente los mapas y guías de viaje son muy generales y no existe información específica para las personas que son entusiastas del vino. Por lo cual estas personas no pueden planificar de manera eficiente y precisa su visita por un lugar tomando en cuenta su presupuesto y su gusto del vino que ellos tengan.

Solución: En este problema se tendrían que usar solo dos técnicas, la de clasificación y la de visualización. Puesto que tenemos que primero clasificar por país, provincias, precios y aromas. Esta base de datos sería grande pero es muy importante poder clasificar de manera precisa todos estos con el fin de que la manera de aplicar la visualización sea la óptima y se trabaje eficazmente. Por tanto la manera de visualización también es fundamental, poner un mapa mundial de manera más general en el cual se pueda poner zoom (esto en una aplicación o una web) para poder ver por distintos países y ya de manera más específicas por país, con colores remarcando el tipo de sabor y con símbolos de pesos con el fin de que la persona pueda ver un mapa en el cual tenga los filtros necesarios para tomar decisiones de su viaje.

Nombre de base de datos: Iris

Objetivo: Revisar el promedio del largo y ancho del pétalo por especie. Para así poder predecir el cómo crecerá cierta especie de Iris solo brindando información del sépalo, puesto que este crece primero antes que el pétalo.

Problema planteado: El problema presentado aquí es que la base de datos tiene un tamaño considerable y solo da la información de los tamaños por separado de ancho y largo de sépalo y pétalo. Por tanto no tenemos una manera de poder predecir cómo crecerá la planta, de echo ni se tiene la relación entre el sépalo y el pétalo por tanto se necesita saber cómo predecirlo.

Solución: Usar la regresión lineal, primero deberíamos hacer la regresión líneal para tanto el largo del sépalo y el pétalo; y del ancho de igual forma. Para cada tipo de especie se haría una diferente y así ya podríamos saber una ecuación que nos permita predecir las medidas futuras del pétalo dadas las condiciones del sépalo original.

Nombre de base de datos: Netlfix shows

Objetivo: Identificar la mejor manera en la cual se tiene que crear un show para que sea “exitoso” y que hayan más series exitosas y por tanto que se gaste menos dinero haciendo series fracaso y se hagan menos pero que sean más exitosas.

Problema planteado: El problema se puede ver al tomar un factor en cuenta principal, el dinero. Por tanto, tenemos que identificar con la base de datos que factores influyen para que una serie sea exitosa. Cabe aclarar que la manera en cómo podemos medir si una serie fue exitosa, o no, es por la cantidad de temporadas, puesto que esto depende de los costos y de la cantidad de vistas, así como la cantidad de membresías nuevas que se vendan gracias a esta .

Solución: Utilizando la cantidad de temporadas como parte principal, podemos utilizar los patrones secuenciales utilizando la clasificación de edad que tenga el show, lugares de filmación, director y el año en el cuál se publicó. Estos son los factores que realmente influyen en nuestra base de datos y nos puede influir de la manera en la que revisemos la cantidad de temporadas de esto. Por tanto con los patrones podemos optimizar la manera de combinar esto para garantizar que una serie tenga al menos dos temporadas, y así nos estaría garantizando que la inversión de la primer temporada fue lo suficientemente buena para que se requiera sacar otra.