UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Bases de Datos

Minería de Datos

Margarita Ordaz Ruiz

Maestra: Mayra Cristina Berrones Reyes

Grupo: 002

Matricula: 1802473

Google Play Store

Nombre de la base de datos: Google Play Store Apps (*googleplaystore_user_reviews.csv*)

Objetivo: Hacer a las aplicaciones bancarias más eficientes, es decir, que sean más rápidas, más amigables a la vista y más fácil de entender para todo tipo de usuarios.

Problema Planteado: Durante la pandemia, muchas de las actividades que se realizaban a diario tuvieron que adaptarse a la nueva realidad, una de las actividades que se vio muy afectada es la de hacer compras, dado que en la contingencia las personas tenían que quedarse en casa, las compras tuvieron que dejar de ser en tiendas físicas y comenzaron a ser en línea, por ende, se tuvo un aumento en el uso de aplicaciones bancarias.

Sin embargo, muchas de estas aplicaciones son muy lentas o difíciles de entender para el público, en especial para las personas mayores, quienes son los más vulnerables en esta pandemia, por lo que se debería de ver en que es lo que más están fallando para poderlas hacer más eficientes.

Solución: Utilizando la base de datos, hacer una clasificación de los de reviews negativos y una de los positivos de todas las aplicaciones bancarias, puestos en la columna *"Sentiment"*, y dados los reviews negativos, hacer un clustering para ver cuáles son las características que más le desagrada a las personas de las aplicaciones bancarias y trabajar para mejorarlas y dados los reviews positivos, igual hacer un clustering para ver cuáles son las características que más les gusta a las personas e implementarlas en su aplicación para que sean más eficientes el uso de estas aplicaciones y las compras en línea.

Coronavirus

Nombre de base de datos: Novel Corona Virus 2019 Dataset (*COVID19_open_line_list.csv*)

Objetivo: Identificar si las manifestaciones en los Estados Unidos a causa de la violencia racial causaron que el número de casos de Covid-19 aumentaran.

Problema Planteado: El 25 de mayo del 2020, falleció George Floyd, tras ser sometido brutalmente por el policía Derek Chauvin, esto conmocionó por completo a Estados Unidos, al grado que varias entidades de esa nación se levantaron en protestas por la violencia racial que se ha ido agudizando en dicho país durante los últimos años.

Dichas protestas iniciaron el 27 de mayo del presente año, sin embargo, como sabemos esto ocurrió dentro de la pandemia, y estaría interesante ver si esto tuvo un impacto en el incremento de los casos de Covid-19 en los Estados Unidos.

Solución: El primer paso es filtrar la información y seleccionar solo la que ocupamos que en este caso son los casos de Covid-19 en los Estados Unidos, después de haber segmentado la información utilizaremos la técnica de visualización para realizar gráficas como histogramas y ver si hay un aumento en los casos a partir de esas fechas, al igual podemos con otra gráfica ver el rango de edades en esas fechas y ver si coinciden con el promedio de edad de los manifestantes y ver si afectó más a los hombres o a las mujeres

Criticas de vinos

Nombre de base de datos: Wine Reviews (*winemag-data-130k-v2.csv)*

Objetivo: Identificar cual es la mejor región para construir un viñedo y asignarle un precio competitivo en el mercado.

Problema Planteado: Suponiendo que somos una empresa nueva, que tienen en mente el proyecto de construir un viñedo para vender vinos de alta calidad a un precio que sea competitivo en el mercado, queremos conocer cuál es la mejor región para construir el viñedo y cual es el rango de precios que es atractivo para el mercado, dadas las características del vino.

Solución: Para el precio utilizamos la técnica de clustering, y agrupamos los vinos de acuerdo con ciertos rangos puntuación en los vinos, y dado que nosotros queremos un vino de alta calidad vemos cual es el precio promedio en estos tipos de vinos y se lo asignamos al nuestro, teniendo en cuenta que el precio nos genere utilidades.

Con respecto a donde construir el viñedo, vemos la agrupación de vinos con mayor puntuación y con la técnica de visualización vemos en que regiones es donde se encuentran, para poder investigar, que condiciones climatológicas son las que hay en esas regiones y poder construir nuestro viñedo en una región que tenga características similares a esas.

Clasificación de plantas

Nombre de base de datos: Iris Species (*iris.csv)*

Objetivo: Construir un clasificador para las flores tipo Iris.

Problema Planteado: Queremos crear un algoritmo que clasifique a las flores tipo Iris, de acuerdo con las características de sus pétalos y sépalos, como lo son su largo y su ancho, y además queremos determinar cual de ellos tiene una mejor precisión.

Solución: Aplicamos los algoritmos de Machine Learning comenzando con Regresión Logística, definimos el algoritmo, LogisticRegression, seguidamente lo entrenamos utilizando la instrucción fit y realizamos la una predicción. Para determinar la precisión o confianza del algoritmo utilizamos la instrucción score para calcularla. Hacemos lo mismo con otros algoritmos o modelos como lo son Máquinas de Vectores de Soporte, Vecinos más cercanos y árboles de decisión, vemos cual de esos es el que tiene una mayor precisión y aplicamos ese.

Shows de Netflix

Nombre de base de datos: Netflix Movies and TV Shows (*netflix_titles.csv)*

Objetivo: Predecir cuales son los tipos de series que más se deberían de crear en un futuro para que la gente siga con sus subscripciones en Netflix.

Problema Planteado: Sabemos que últimamente se han empezado a crear muchas plataformas de streaming, tenemos Amazon Prime, HBO que ha estado lanzando series exclusivas que han tenido gran impacto, y el 17 de noviembre del presente año va llegar a Latinoamérica la nueva plataforma "Disney plus", esto le genera una necesidad a Netflix de producir contenido exclusivo para que las personas sigan con Netflix, el problema está en ver que tipo de series deberían de crear para que la gente siga con Netflix.

Solución: Primero debemos de segmentar los datos, elegimos en "type" las series, y las clasificamos de acuerdo el rating y a partir de ahí podemos identificar cuales son las que tienen un mayor número de temporadas y con la técnica de clustering agrupamos al cast y a los directores, para ver si hay alguna relación en que las series duren más de acuerdo al reparto y/o al director, así vemos que a que segmento ponerle cierto tipo de serie, con cierto reparto para que tenga éxito.