Materia: Minería de Datos

Ejercicio práctico de bases de datos

Profesora Mayra Cristina Berrones Reyes

Alumno:

Alfonso Llanos Morales

1887939

7° Semestre

Licenciatura en Actuaría

14/10/2020

Google Play Store

Nombre de la Base de Datos: googleplaystore.csv

Objetivo: Predecir el éxito que tendrá una determinada aplicación dependiendo de las características de esta, como el género, si es de paga o gratis o la calificación de contenido.

Problema planteado: Los accionistas de una nueva aplicación para Android desean saber si en base a la información del mercado actual en la Play Store se espera que la nueva aplicación sea exitosa o no, y así tomar la decisión de invertir en ella.

Solución: Utilizar la técnica de clasificación para predecir el éxito de la aplicación en base a características que comparte o no con otras aplicaciones exitosas del mercado. Se utilizará la base de datos "googleplaystore.csv" de la cual tomaremos para realizar la clasificación las columnas de *Category*, porque hay categorías más populares como videojuegos; *Size*, porque si la aplicación es muy pesada puede que no sea agradable para todo público; *Type*, debido a que el ser gratis o de paga es un factor determinante en su uso y en la estrategia de ingresos de la empresa dueña de la app; *Price*, por las mismas razones del anterior; *Content Rating*, ya que la calificación del contenido también segmenta de manera considerable al mercado y *Genres*, que es una clasificación más especifica que complementa a la columna de *Category*.

Coronavirus

Nombre de la Base de Datos: covid_19_data.csv

Objetivo: Analizar el desempeño que han tenido los países y sus regiones durante la pandemia del Covid-19 para tomar las mejores decisiones en la distribución de vacunas dando prioridad a las zonas que más lo necesitan.

Problema planteado: La Organización Mundial de la Salud (OMS) en coordinación con diferentes empresas de biotecnología y farmacéutica necesita desarrollar un plan para una distribución óptima de las próximas vacunas para la enfermedad Covid-19, ya que en un principio la cantidad de vacunas y su producción serán limitadas, por lo que deberán decidir el orden de las zonas en las que se entregarán y aplicarán las vacunas.

Solución: Se desarrollará un algoritmo que ordene los países y sus zonas geográficas dependiendo de las características que los hagan más prioritarios para la aplicación de las vacunas. De la base de datos "covid_19_data.csv" se utilizarán las columnas *Province/State*, *Country/Region, Confirmed*, *Deaths* y *Recovered*.

Crítica de vinos

Nombre de la Base de Datos: winemag-data-130k-v2.csv

Objetivo: Encontrar las características que sean determinantes en el éxito en las críticas para una clase de vino en específico.

Problema planteado: Una empresa de vinos está desarrollando un nuevo producto y desea conocer las características que necesita para ser un vino amado por la crítica.

Solución: Desarrollar un programa que nos muestre las características que son más relevantes para los críticos al momento de realizar su review. De la base de datos "winemag-data-130k-v2.csv" utilizaremos las columnas de *country*, *province*, *región_1*, *región_2*, porque el lugar de origen de un vino es muy importante porque las condiciones geográficas, climáticas y la materia prima afectan en el sabor, aroma y apariencia del vino; *variety*, porque el tipo de uvas es un factor determinante y más si se toma en cuenta que en determinados países se produce un determinado tipo de uva de mayor calidad; *description* y *points*, para encontrar los atributos que más son mencionados por los críticos y el número de puntos como una respuesta generalizada.

Clasificación de plantas

Nombre de la Base de Datos: Iris.csv

Objetivo: Determinar el tipo de Iris que es una planta dependiendo de las dimensiones de sus pétalos y de sus sépalos.

Problema planteado: Un estudiante de biología necesita clasificar una serie de plantas iris dependiendo si son de la especie setosa, virginica o versicolor en el menor tiempo posible y con el menor margen de error, ya que analizar y comparar cada una de las plantas por si solo puede tomarle mucho tiempo.

Solución: Desarrollar una herramienta de aprendizaje de máquina que pueda determinar la especie de Iris dependiendo de la longitud del sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo de una planta dada.

Shows de Netflix

Nombre de la Base de Datos: netflix_titles.csv

Objetivo: Predecir si una película o serie de televisión será incorporada al catalogo de Netflix en un futuro y cuando.

Problema planteado: Una persona por razones varias no pudo ver una película que estaba esperando con mucha emoción cuando fue estrenada en cines, y las fechas de emisión ya pasaron, o se quedó sin ver el final de su serie favorita en TV, por lo que está buscando una alternativa para poder verla y no quiere entrar a una página en internet de dudosa procedencia donde puede estar en peligro de descargar un virus, además no se verá con buena calidad y él está en contra de la piratería, por lo que decide esperarse a que sea incorporada al catalogo de Netflix y poder disfrutarla con toda la tranquilidad. Por todo lo anterior desea saber si existe la posibilidad de que la película o el programa de TV sea agregado al catalogo y cuando, sino para buscar otra alternativa.

Solución: Con base en la información proporcionada en la base de datos “netflix_titles.csv” se creará un algoritmo que nos diga si una película o serie de TV se espera que sea agregada al catalogo de Netflix y que pueda predecir cuando será agregada, dependiendo de características como el director, el país de procedencia, la fecha de incorporación, los actores, la clasificación y el género de las películas y series que ya han sido agregadas anteriormente.