



UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICO



Asignatura:

Minería de Datos

Tema:

Ejercicio – Bases de Datos

Datos.

Docente: Mayra Cristina Berrones Reyes

Alumno: Saúl Arath Hernández Hernández

Matrícula: 1815642

Fecha: 11/10/20

1. Google Play Store Apps.

Datos extraídos de la web de las aplicaciones de Play Store de 10k para analizar el mercado de Android.

Los datos que presentan las aplicaciones de Play Store nos dan la oportunidad de identificar patrones para impulsar el éxito de las empresas creadoras de aplicaciones.

Tenemos datos como la calificación de la app, comentarios de la app, categoría, número de descargas, precio, género y tamaño.

Con las columnas podemos revisar patrones secuenciales, clasificación o reglas de asociación en donde podamos establecer reglas en las que, si cumple con determinados criterios, podríamos decir que cualquier app que cumpla con esos criterios, tendrá asegurado el éxito en el mercado.

Por otro lado, de la base de datos podemos establecer criterios de relación entre categorías y ratings en donde podríamos mostrar un algoritmo de recomendación a los usuarios determinado por estos datos.

2. Novel Corona Virus 2019 Dataset.

Información a nivel de día sobre casos afectados por covid-19.

La base de datos es una serie de tiempo, por lo que los datos sobre muertes, recuperados son acumulados.

La base de datos trata información como región, estado, día de observación, casos confirmados, muertes y recuperados.

Con estos datos podemos plantear modelos de regresión en donde podamos predecir a futuro el comportamiento de las muertes, con un modelo de regresión podría preverse el tiempo en el cual el número de muertes baje y así implementar medidas para la recuperación de la economía.

Por otro lado, con los datos presentados, podrían crearse patrones que a futuro podríamos identificar ante una nueva amenaza de virus y con esto, estar mejor preparados para no volver a pasar por las repercusiones que pasamos con la pandemia actual. Además, se nos permite realizar inferencia sobre estos conjuntos de datos en donde podemos ver la distribución de ellas y aplicar distintas pruebas de hipótesis para su análisis estadístico.

3. Crítica de Vinos.

130k reseñas de vinos con variedad, ubicación, bodega, precio y descripción.

Tenemos datos de críticas de consumidores en donde puntuaron la calidad del vino, una de las problemáticas que podríamos solucionar viéndolo desde el lado de un productor o un restaurante, sería tomar en cuenta los patrones que hicieron que los consumidores puntuaron de manera alta un vino, usualmente la calidad depende del tipo de vino y de la región de donde pertenece la materia prima.

Como productor o dueño de un restaurante, nos interesa buscar las características que más les gusta a los consumidores, por lo que, mediante reglas de asociación, podemos identificar esos patrones con los cuales tenemos una crítica positiva con una puntuación alta en la calidad de vino. Una vez revisadas las preferencias, podemos tomar decisión administrativa sobre los precios, ventas, etc.

Desde el lado del consumidor, con los mismos patrones de asociación, podemos identificar los vinos de mejor calidad para así poder tomarlos en cuenta para su consumo personal.

4. Clasificación de Plantas.

Clasificación de las plantas de iris en tres especies en este conjunto de datos.

Con esta base de datos, vemos información en variable de medida, tenemos a la anchura y longitud del tallo del sépalo. Viendo el conjunto de datos por orden, pareciera ver que la relación entre ancho y longitud tiene una relación directa positiva por lo que podríamos aplicar un modelo de regresión en el cual relacionemos estas dos variables.

Con ello, podemos tener un modelo de comportamiento por especie de Iris en donde podamos ver la relación entre ancho y longitud.

Con esta información, tendríamos ya las características por especie, por lo que, para su estudio en campo, con esta información tendríamos ya la manera de identificar por sus características qué tipo de especie es la que estamos analizando. Esto puede ser aplicado por árbol de decisión en conjunto con un análisis de regresión lineal o del modelo al que mejor se relacione.

5. Shows y Películas de Netflix.

Listados de películas y programas de televisión en Netflix.

Para los usuarios, podríamos implementar un algoritmo en el que por técnica de clustering identificar lo que ven los usuarios en base a ciertos patrones, con ello al momento de terminar una serie o una película, el algoritmo despliegue una lista rápida en la que podrá elegir entre ciertas opciones para ahorrarse el tiempo de salir al menú o de usar el buscar para buscar una siguiente reproducción.

Para los socios o empresas, una manera de hacer eficaz la publicidad o de buscar más reproducciones o vistas en sus programas, tener como primera vista o tener una pestaña en la cuál se promocionen los programas con menor audiencia, esto para lograr salvar el proyecto.

Por región, la clasificación y tipo de contenido varía mucho; una manera de salvaguardar a los servidores e incluso reducir costos, podría ser segmentar el contenido por país de modo que no saturaríamos la biblioteca de contenido.