UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Minería de Datos

Fase 2: Análisis de Bases de Datos

Gilberto Noé López Ávila

1812678

14 de octubre de 2020

Aplicaciones de Google Play Store

Base de datos: Google Play Store Apps

Objetivo

Predecir el rating de una nueva aplicación a partir de sus características previas a lanzamiento.

Problema planteado

El rating de una aplicación determina su percepción ante el público, mostrándola como una buena y útil herramienta o como una mala alternativa a otra aplicación de mejor rating. El éxito de la aplicación dependerá de la calificación de los usuarios, por lo que es importante conocer las características que más pueden influir en estas calificaciones. Considerando las características de la base de datos que están disponibles al momento de lanzar la aplicación, como el precio, género, clasificación, etc., podremos predecir con cierto grado de exactitud el rating que la aplicación podrá obtener.

Solución

Necesitamos un modelo predictivo para generar una estimación del rating de una aplicación nueva tomando en cuenta las características listadas en la base de datos. Es posible realizar una regresión multivariable para poder determinar cuál característica tiene mayor influencia en la predicción del rating y seleccionar el conjunto de características que realicen la predicción más confiable a los ratings observados para las aplicaciones de la base de datos.

Casos de COVID-19

Base de datos: Novel Corona Virus 2019 Dataset

Objetivo

Comparar la evolución de la pandemia de COVID-19 a través del tiempo en diferentes países e identificar las regiones geográficas más afectadas y aquellas con pocos casos actualmente.

Problema planteado

La pandemia actual ha presentado un desafío histórico para todas las naciones del mundo. Comenzando en China, y con el tiempo trasladándose al resto del mundo, el COVID-19 ha entrado a distintos países en diferente tiempo y ha evolucionado en estos a diferente ritmo. La visualización de esta base de datos brindará una mayor comprensión del comportamiento de la enfermedad a nivel mundial y permitirá generar hipótesis sobre los datos a partir de las conclusiones del análisis visual.

Solución

Podemos realizar varias visualizaciones de los datos de acuerdo con el tema que se desee observar. Por ejemplo, para distintos periodos, es posible graficar un mapa de calor sobre el mapa del mundo, resaltando los países con mayor número de casos durante ese periodo, y así observar fácilmente las regiones más afectadas. También podemos comparar las series de tiempo de los casos y muertes acumulados para un conjunto de países, y reconocer aquellos que han manejado la pandemia de forma más controlada.

Críticas de Vinos

Base de datos: Wine Reviews

Objetivo

Clasificar distintos tipos de vino por su descripción al gusto y olfato para poder sugerir varias opciones de vinos con características similares.

Problema planteado

Existen varios factores que pueden influir en el sabor, aroma y consistencia de un tipo de vino, como el tipo de uvas, el país y la provincia de origen, etc. Distintas combinaciones de estas características podrían generar distintos sabores únicos. Por ejemplo, el mismo tipo de uva cultivado en distintos países puede tener cualidades suficientemente distintas para quien desee un tipo particular de vino. Podemos utilizar las características deseadas para seleccionar un grupo de vinos similares, considerando los distintos factores de su elaboración.

Solución

Una técnica de agrupación nos permitirá definir conjuntos de vinos de acuerdo con las características que produzcan un mismo sabor y aroma. Un algoritmo de clustering puede utilizarse después de convertir las descripciones para cada vino a un formato utilizable. Cada cluster representará un conjunto de descripciones deseadas, las cuales podremos mostrar como distintos tipos de vino por sus características de interés.

Especies de Iris

Base de datos: Iris Species

Objetivo

Obtener un método preciso de clasificación botánica de especies de iris en función de sus características observables y medibles.

Problema planteado

La clasificación de especies es comúnmente usada en la biología como herramienta de estudio e investigación. Dada su importancia en el campo de estudio, buscamos que esta clasificación dependa de características objetivas y mediciones, en vez de alguna posiblemente ambigua descripción del color o la forma de la flor. La base de datos proporciona la anchura y longitud del sépalo y los pétalos de la flor, medidas que pueden fácilmente registrarse para un nuevo iris por analizar.

Solución

La técnica de clasificación de árboles de decisión permite seguir una sencilla serie de preguntas acerca de las características de la flor para determinar su especie. Para cada clasificación podemos obtener los valores adecuados tales que la gran mayoría de las veces podamos predecir la especie a la que una nueva flor de iris pertenece a partir de sus características principales.

Películas y Series de Netflix

Base de datos: Netflix Movies and TV Shows

Objetivo: Sugerir recomendaciones de películas y series basadas en la dirección o el reparto de otras películas vistas por el usuario.

Problema planteado: Netflix cuenta con un algoritmo de recomendación basado en el historial del usuario para mostrar con mayor frecuencia contenido del mismo género o de misma audiencia que las series y películas vistas anteriormente. Sin embargo, este servicio de streaming no cuenta con categorías de sugerencia por directores o por actores y actrices de contenido previamente visto. Podemos construir un método de recomendación que muestre, para algún director o actor, otros directores y actores que han colaborado con frecuencia con esta persona, para recomendar las películas o series de la misma y las colaboraciones mencionadas.

Solución: Construir reglas de asociación, tratando cada película o serie como una transacción y considerando a su director y su reparto como los elementos de la lista. Las reglas de asociación más frecuentes indicarán las colaboraciones más comunes entre directores y actores o entre el mismo reparto. Así, las recomendaciones de este tipo consistirán en contenido con una dinámica similar entre la actuación y la dirección.