



Universidad Autónoma de Nuevo  
León



Facultad de Ciencias Físico  
Matemáticas

Minería de datos

Docente: Mayra Cristina Berrones Reyes

Ejercicio bases de datos

Grupo 002

Sofía Pamela Rosales Garza 1799219

San Nicolás de los Garza, 14 de octubre del 2020

## Ejercicio bases de datos

### *Google Play Store*

Nombre de la base de datos: googleplaystore

Objetivo: Mejorar la calidad de las recomendaciones que Google Play Store hace a sus usuarios al momento de ingresar a la plataforma y después de descargar una aplicación.

Problema planteado: Las recomendaciones que hace Google Play Store son más enfocadas a lo que actualmente es popular (para el público en general), muchas veces esto no coincide con las preferencias de los usuarios y aunque es lo primero que les aparece al ingresar al sitio, no descargan esas aplicaciones porque no son de su interés.

Solución: Para mejorar esto, podríamos usar la técnica de predicción y crear árboles de decisión según las aplicaciones que el usuario ha instalado en su dispositivo, para que lo primero que vea al momento de entrar a Google Play Store sean recomendaciones más precisas de aplicaciones que de alguna manera estén relacionadas (podría ser por la categoría, el desarrollador, etc.). También, con ayuda de reglas de asociación y patrones secuenciales, después de que un usuario descargue cualquier app, podemos mostrarle aplicaciones que también le podrían interesar, en base a un estudio de las apps que otros usuarios descargaron en común y que están relacionadas. Esto también ayudaría a posicionar y dar a conocer apps nuevas que están relacionadas a los gustos de los usuarios.

### *Coronavirus*

Nombre de la base de datos: COVID19\_line\_list\_data

Objetivo: Determinar si el nivel socioeconómico de un país o una región está relacionado con qué tantos contagios hay y también que tan rápida es la recuperación.

Problema planteado: Al inicio de la pandemia se “buscaba echar culpas” o responsables sobre quién había propagado el virus, algunos decían que las personas con un nivel socioeconómico más alto eran los responsables porque son quienes tienen mayor y más fácil acceso a viajar internacionalmente y se decía que traían el virus de otras partes. También, al momento de que el gobierno recomendó la cuarentena, se “responsabilizaba” a la clase obrera del país pues eran quienes seguían con sus actividades normales, la mayoría por motivos de trabajo. Un estudio nos ayudaría a rechazar alguna de estas dos hipótesis y también a saber si el nivel socioeconómico influye en la pronta o lenta recuperación de las personas.

Solución: Con la técnica de reglas de asociación podemos ayudarnos para relacionar a las personas contagiadas con su nivel socioeconómico y encontrar una relación o no respecto a su recuperación.

## ***Críticas de vinos***

Nombre de la base de datos: winemag-data-130k-v2

Objetivo: Abrir un viñedo que cumpla con las mismas o el mayor número de características de los viñedos que tienen las mejores críticas de sus vinos, pero siendo también específicos en qué tipo de vino queremos producir (no porque sea un buen terreno quiere decir que todo tipo de uva se da en esa región).

Problema planteado: Queremos abrir un viñedo pero no sabemos en qué ubicación sería adecuado y tampoco sabemos qué tipo de uva queremos o convendría cosechar según el clima, características del suelo, etc. de diferentes y potenciales regiones en las que podríamos abrirlo. Al tener los datos de los vinos mejor calificados, es más fácil tomar decisiones del viñedo basándonos en los que recibieron mejores críticas.

Solución: Vamos a estudiar los vinos que tuvieron mejores críticas clasificándolos también según sus descripciones, el lugar de dónde provienen y así analizar las características de las regiones donde las uvas de esos vinos fueron cosechadas con el objetivo de buscar un lugar que sea similar a los terrenos de los otros viñedos y saber cuál tipo de uva sería óptimo cosechar ahí. Con ayuda de la técnica de minería de clasificación, primero ordenamos los vinos según las descripciones que los críticos dieron (su tipo) y dentro de esa clasificación hacemos otra de las características que tienen los viñedos (el terreno, el tipo de suelo, el clima de la región, qué uva crece en esas condiciones, entre otras cosas).

## ***Clasificación de plantas***

Nombre de la base de datos: Iris

Objetivo: Facilitar y minimizar el tiempo de búsqueda de plantas dentro de una base de datos según las características que presenta.

Problema planteado: Buscar dentro de una base de datos grande puede ser muy tedioso y también puede alentar el proceso de búsqueda, además de que si se hace manualmente podríamos pasar por alto alguna planta o característica según sea el caso lo que llevaría a un error o una omisión.

Solución: Clasificar las plantas de acuerdo con sus características, esto podría ser el tipo de planta que es, el lugar en donde crece, su tamaño, si tiene pétalos o frutos, si es venenosa o no, entre muchas otras. Considero que además de la técnica de minería de clasificación, también podría usarse el clustering, pues crea grupos según características similares entre sus elementos y ayudaría a diferenciar entre los tipos de plantas según la característica que quiera tomarse como diferenciadora.

## ***Shows de Netflix***

Nombre de la base de datos: netflix\_titles

Objetivo: Extender el tiempo que el usuario pasa en Netflix viendo una película/serie.

Problema planteado: Muchas veces los usuarios pasan mucho tiempo buscando qué ver en Netflix, pero lo que verdaderamente se busca es que ese tiempo lo pasen viendo algún show y no sólo en búsquedas, por lo que deberían mejorarse las recomendaciones que la plataforma hace. Muchas veces aparece “porque viste x show”, pero no siempre terminaste de ver ese programa, algunas veces sólo viste uno o dos episodios y después dejó de interesarte. De esta manera si hace recomendaciones relacionadas pero de algo que anteriormente no te atrapó por completo, teniendo como consecuencia una lista de series y películas que es probable que no vayas a ver.

Solución: Monitorear qué películas/series se vieron completamente o poner un mínimo de episodios en el caso de los shows (según el número total de capítulos que tenga) y en base a eso y con ayuda de la técnica de predicción (haciendo árboles, ya que ayudan analizando datos del pasado para obtener información no conocida) hacer recomendaciones más acertadas para el usuario, que sean programas o películas que verdaderamente sean de su interés para que de esta manera quiera verlos y pase más tiempo dentro de la plataforma.