# UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

# FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS

## MINERÍA DE DATOS

### EVIDENCIA DE APRENDIZAJE
### EJERCICIO BASES DE DATOS

MAESTRA: MAYRA CRISTINA BERRONES REYES

ALUMNA: PATRICIA SARAHÍ ARVIZU RIVERA

MATRÍCULA: 1823604

GRUPO: 002

SAN NICOLÁS DE LOS GARZA A 14 DE OCTUBRE DE 2020

**GOOGLE PLAY STORE APPS**

**Objetivo:** Determinar las principales y/o mejores características que una aplicación debe cumplir para brindar mayores probabilidades de obtener un gran número de instalaciones y rating alto.

**Problema planteado:** Las aplicaciones tienen muchas características, tales como categoría, tamaño, de paga o gratuita, público a quien va dirigido, precio y género. Sin embargo, como compañía de desarrollo, por ejemplo, es importante conocer cuáles son las mejores características que nuestra aplicación debe cumplir para que así nuestro próximo lanzamiento tenga más oportunidades de éxito; además se podrán realizar proyecciones sobre la aplicación en proyecto para así reducir la probabilidad de que nuestra aplicación fracase conllevando pérdidas económicas y tiempo perdido para la empresa.

**Solución:** Desarrollar una herramienta que estime el número de instalaciones y el rating a obtener para clasificarlas y así tener una calificación similar a la que usen las calificadoras de riesgos y para poder tener una idea sobre cómo se comportaría la aplicación a lanzar. Para esto, se podrían usar dos regresiones múltiples. Una usando el número de instalaciones como variable de respuesta y otra donde lo sea el rating, usaríamos como variables regresoras las distintas características que nos da la base de datos y con las estimaciones que nos arroje la regresión realizaríamos la clasificación.

**NOVEL CORONA VIRUS 2019 DATASET**

**Objetivo:** Determinar el porcentaje de población infectada por ciudad en el mundo en un tiempo determinado para visualizarlo en un mapa y conocer el riesgo que tiene una población.

**Problema planteado:** El número de contagios se ha multiplicado en cantidades significativas en todas las ciudades. Es necesario conocer o determinar el riesgo que presenta una región de acuerdo con el porcentaje de población infectada. Además, las migraciones juegan un papel muy importante y por temas económicos y/o sociales estas pueden ser más difíciles de detener. Es importante reactivar la economía, un buen comienzo sería determinar zonas “seguras” en donde las personas puedan transitar y el riesgo o probabilidad de contagio no aumente.

**Solución:** Desarrollar una herramienta de visualización, como un mapa dividido por ciudades donde cada una tenga un color de acuerdo con porcentaje de población infectada. Para esto, agruparíamos a los individuos que tenemos en la base de datos de acuerdo con la ciudad de registro, con esto tendríamos la cantidad de personas infectadas por ciudad y necesitaríamos otra base de datos con la población actual de la ciudad para así obtener el porcentaje total por ciudad y determinar el color que tendrá la ciudad en el mapa; de esta misma manera se pudieran poner como filtros para tener información más detallada, por ejemplo un mapa con el porcentaje de población infectada de las personas mayores a 68

año. Ahora, podemos desarrollar una técnica de clustering para hacer regiones que compartan el mismo riesgo y así las zonas que tienen el menor porcentaje de población infectada podrán transitar dentro de estas.

**WINE REVIEWS**

**Objetivo:** Determinar si existen vinos que se produzcan sólo en un país debido al tipo de uva, así como los países que fabrican los mejores vinos de acuerdo con el rating obtenido en una escala del 1 al 100.

**Problema planteado:** Es atractivo para los turistas, sobre todo los amantes del vino, conocer aquellos que son únicos es decir aquellos que sólo se producen en un país. Darles difusión a estos vinos ayudaría a incrementar el turismo y la economía de dicho país, así como para los países que se caractericen por tener los vinos con mejores ratings. Además, serviría para darse a conocer y poder distribuirse en las cadenas de restaurantes más importantes y las mejores vinotecas o viceversa, si un inversionista busca invertir en un restaurante con un concepto elegante y único, esta información sería de gran ayuda.

**Solución:** Desarrollar una herramienta de asociación para asociar los vinos a su país, aquellos que sean asociación única serán los vinos que estamos buscando. Para clasificar a los países con los mejores vinos, podemos clasificar a los vinos en distintas categorías según su rating y después nuevamente asociarlos para saber cuáles son los países que tienen más vinos de la mejor categoría. Sería una buena decisión, por ejemplo, importar vinos de dicho país ya que por compartir el país de origen algunos costos de importación y traslado se podrían reducir.

**IRIS SPECIES**

**Objetivo:** Determinar el tipo de especie de Iris al que corresponde una flor de acuerdo con sus características tales como longitud y anchura del sépalo y del pétalo para facilitar el estudio e investigación en ámbitos como la biología.

**Problema planteado:** Tenemos tres tipos de la planta de especie Iris y de acuerdo con sus características tales como longitud y anchura del sépalo y pétalo podemos determinar si la planta seleccionada pertenece a la especie de *setosa, versicolor o virginica*. Esto tiene una gran relevancia en el ámbito de la botánica y la biología en general ya que el mismo algoritmo se puede replicar para otras especies tanto de plantas como de animales y ayudaría a tener una mejor clasificación para realizar más estudios con gran relevancia en la investigación.

**Solución:** Desarrollar un algoritmo de predicción de árbol de clasificación, donde por medio de las características que tenga la planta, tales como la longitud y anchura del sépalo y del pétalo podamos determinar el tipo de especie de Iris a la cual corresponde la planta que hemos seleccionado. Este mismo tipo se puede replicar para otras especies tanto de plantas como de animales y así tener clasificadas las especies para un mejor estudio o investigación.

**NETFLIX MOVIES AND TV SHOWS**

**Objetivo:** Ofrecer un servicio más personalizado para los usuarios basado en el contenido más visto dentro de la plataforma para así aumentar la probabilidad de que el usuario renueve su suscripción.

**Problema planteado:** La competencia dentro de los medios digitales para entretenimiento ha ido en aumento; asegurarse de ofrecer la mejor experiencia para el usuario se ha convertido en un aspecto esencial. Contamos con diversos datos sobre las películas y los shows en Netflix, quien también se ha dedicado a crear su propio contenido pero al incluir en su catálogo también películas creadas por otras instancias o simplemente tener programas con diferente duración nos da una situación para analizar y trabajar para que el contenido sea el mejor, tomando en cuenta también el tiempo que pasa desde que la película o programa fue lanzado a nivel mundial y el tiempo en que Netflix lo agrega a su catálogo.

**Solución:** Obtener una base de datos donde tengamos el número de veces que se ha visto un programa o película para realizar un algoritmo de regresión donde usemos este número como variable de respuesta y como variable regresora diversos datos como el tiempo que ha transcurrido entre la fecha de lanzamiento internacional y la fecha en fue agregada al catálogo de la plataforma o la categoría a la que pertenece. Con los datos de directores y protagonistas podemos usar relaciones o asociaciones y ver cuáles son las que nos da un mejor número de vistos. Además, podemos clasificar los programas según sus categorías o actores que tienen en común y así recomendar al usuario estas opciones para ver después y crear una secuencia que sea personalizada gracias a las herramientas de minería de datos.

SAN NICOLÁS DE LOS GARZA A 14 DE OCTUBRE DE 2020