UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO

MATEMÁTICAS

Unidad de Aprendizaje

Minería de Datos

“Ejercicio práctico base de datos”

Profesor: Mayra Cristina Berrones Reyes

Alumna: Valeria Solis Agundis

Matricula: 1815413

Carrera: LA

Semestre: 7mo

Grupo: 002

14/10/2020

Clasificación de plantas

Base de datos: “Especies de Iris”

Objetivo: Los datos recopilados incluyen las tres especies de iris con 50 muestras cada una, así como algunas propiedades de cada flor. Donde una especie de flor es linealmente separable de las otras dos, pero las otras dos no son linealmente separables entre sí.

En las columnas podemos encontrar información sobre la longitud y ancho del sépalo, también encontramos la longitud y ancho del pétalo, y así mismo las especies que conforman esta base de datos, las cuales son setosa, versicolor, y virginica.

Es importante mencionar que todos los datos obtenidos son en centímetros.

Queremos hacer diferentes análisis a esta base de datos, sobre cómo se comportan los datos, o si queremos hacer un análisis más a fondo sobre un caso en específico para aprovechar la información recopilada.

Problema planteado: Queramos analizar los datos para las variables ancho del sépalo vs longitud del sépalo.

Solución: Una técnica de la minería de datos que nos podría ayudar a la solución de esta propuesta sería la visualización, ya que esta mediante la representación gráfica nos ayuda a interpretar los datos de una manera visual y sencilla, en este ejemplo podríamos realizar una gráfica de dispersión.

También podríamos realizar histogramas para ver e interpretar como se han manejado las variables que tenemos en la base de datos, en fin, la visualización puede ser una gran técnica para usar en esta base de datos puesto que es muy sencilla, pero a la vez muy práctica.

Google play store

Base de datos: Aplicaciones de Google Play Store

Objetivo:

Este conjunto de datos prácticamente recopila información de la Play Store, sobre las aplicaciones que las personas descargan, donde la información recopilada de cada aplicación muestra lo siguiente: su categoría, el nombre de la aplicación, el número de descargas, el tamaño de la aplicación, si es gratuita o de paga, entre muchos más datos interesantes.

Algo interesante de esta base de datos es que tiene mucho potencial para la creación y análisis de aplicaciones, ya que con los datos recabados se pueden hacer observaciones y análisis interesantes para que los expertos sepan hacía que ámbito buscar e ir.

Problema Planteado:

Queremos saber que aplicación será la más descargada en el futuro a través de los datos observados. O así mismo que cuando el usuario entre a la Play Store de acuerdo a la

información obtenida del anteriormente, como sus búsquedas, descargas se pueda ser capaz de hacer recomendaciones basadas en sus gustos y en sus búsquedas más recientes

Solución:

Esto es posible de predecir con alguna técnica de minería, como lo es la regresión, ya que esta se basa en predecir valores basándonos en los datos anteriores, y aquí si contamos con este tipo de datos, ya seria cuestión de hacer el análisis de regresión adecuado para obtener una predicción a esta situación.

Y lo segundo planteado puede lograrse haciendo uso de las reglas de asociación, para hacer las recomendaciones adecuadas de acuerdo con los gustos del usuario.

Críticas de vino

Reseñas de vinos

Objetivo:

La siguiente base de datos cuenta con 130k reseñas de vinos con variedad, ubicación, bodega, precio y descripción. Donde los datos se extrajeron de WineEnthusiast durante la semana del 15 de junio de 2017.

A través de toda la información recopilada el autor busca que con esta se pueda analizar y vincular que tipos de vinos se tiene de manera rápida y confiable.

Problema Planteado:

El autor de esta base de datos comenta una problemática que le gustaría resolver, la cual plantea que le gustaría crear un modelo que pueda identificar la variedad, la bodega y la ubicación de un vino en función de una descripción.

La base de datos con la que contamos recopila información que nos da características muy importantes para poder crear un modelo, tenemos desde el tipo de vino, el viñedo del que provino, su precio, entre muchos datos más.

El autor comentaba que le gustaría que a través de pocas palabras se pueda saber del tipo de vino que se está hablando.

Solución del Problema:

Tomando en cuenta todas las técnicas que hemos visto y analizado, una solución a esto puede ser la técnica de “Predicción” ya que a través de los árboles de decisión podemos durante el proceso de identificación ir descartando los tipos de vinos mediante las características obtenidas, ya que irían pasando estas por “filtros” hasta llegar a la que se asemeje a la descripción dada y predecir el vino del cual se hablaba desde un principio.

Coronavirus

Nuevo conjunto de datos del virus Corona 2019

Objetivo:

Debido a la situación que estamos pasando actualmente, la siguiente base de datos es un recopilatorio de datos de acuerdo con cómo se ha ido comportando el virus con las personas.

El nuevo coronavirus 2019 (2019-nCoV) es un virus (más específicamente, un coronavirus) identificado como la causa de un brote de enfermedad respiratoria detectado por primera vez en Wuhan, China. Este conjunto de datos tiene información de nivel diario sobre el número de casos afectados, muertes y recuperación del nuevo coronavirus de 2019.

Al analizar el tipo de datos, los cuales son a través del tiempo, podemos realizar grandes análisis, si bien no para predecir el virus como tal, podemos analizar y crear estadísticas.

Problema Planteado:

En el conjunto de datos tenemos, cuando fue hospitalizado, si es que falleció o no, para analizarlos en casos específicos, con estos datos se quiera predecir un estimado de muertos o recuperados de acuerdo con los datos recopilados.

Solución del Problema:

Esto se podría hacer con la ayuda de las técnicas de minería de datos, en este caso "Regresión" puede ser de gran ayuda para poder predecir el número de muertos, el número de recuperados, entre otras situaciones ya que tomaría los datos anteriores y con el análisis adecuado de regresión se podría llegar a una estimación acertada.

Shows de Netflix

Base de datos: "Películas y programas de televisión de Netflix"

Objetivo:

La base de datos consta de programas de televisión y películas disponibles en Netflix a partir de 2019, al estar observando esta base de datos algo que toma importancia y se menciona es el que Netflix está apostando más por contenido original.

Las principales columnas con información recabada que podemos encontrar son el saber si lo visto es una película o serie, los actores que participan, su director, el país de origen y su durabilidad, esta es una información muy valiosa a la que le podemos dar diferentes usos.

Como asumir y recopilar de manera rápida y eficaz los gustos del usuario para hacer recomendaciones.

Problema Planteado:

Ya vimos que, a través de los años, Netflix ha agregado más contenido original como series, un uso para esta base de datos es ver cuántos usuarios disfrutan este contenido y si es de su agrado.

Solución:

Con la ayuda de una técnica de la minería de datos, la cual es predicción podremos saber a través de la información recopilada, primero que nada, si se detecta que el usuario tiene cierta atracción por este contenido, predecir si el usuario seguirá consumiendo este tipo y así mismo, predecir y recomendar contenido de su agrado.

También a través de patrones secuenciales, por temporadas podemos analizar para así mismo recomendar al usuario contenido de esa temporada para así llamar su atención, a que de click sobre el título, por ejemplo, si es temporada navideña, mostrar al principio mediante los patrones secuenciales una agrupación de películas y series con la temática.