MINERÍA DE DATOS

Evidencia de Aprendizaje:

Ejercicio práctico bases de datos

Profra. Mayra Cristina Berrones

Alumno: VICTOR HUGO ALANIS MARES

Matrícula: 1821920

L.A. Licenciatura en Actuaría

FECHA: 13 de octubre del 2020

Nombre de la base de datos: Google Play Store

Objetivo: maximizar la capacidad de alcance de las aplicaciones

Ploblema planteado: Cientos (si no es que miles) de aplicaciones se publican en la playstore día a día, esto conlleva un problema gigante inherente a la propia play store, cuando yo publico una aplicación es como agregar una gota al océano ¿Cómo hago que mi aplicación se destaque? ¿Existe alguna formula especial para maximizar mis probabilidades de éxito? E incluso podemos plantear el problema que más se le adelanta, una vez que tenga audiencia ¿Cómo maximizo las ganancias?

Solución: Existen diferentes datos que podríamos utilizar en este nivel de problemática, sin ir muy lejos debemos evaluar la máxima cantidad de aspectos significativos que podamos, por ejemplo podemos primero intentar hacer un estudio de correlación entre nombre y éxito, este estudio se puede basar en diversas cosas tal como:

1: estudiar si existen palabras especificas que están en la mayoría de apps famosas

2: Estudiar si la longitud del nombre implica un cambio en la cantidad de descargas

Podríamos hacer esto ultimo con un estudio de regresión.

Posteriormente podríamos estudiar la relación entre categoría y éxito ¿Existen categorías especificas donde es más fácil triunfar? Esto lo podríamos hacer con un sistema de clasificación. Incluso si nuestra aplicación tiene micro-transacciones dentro del juego podríamos hacer un estudio de secuencias para determinar si hay categorías donde las personas este más dispuestas a gastar dinero que en otras para así dependiendo de la categoría en donde tengamos pensado lanzar nuestra aplicación determinar si nos conviene más un sistema de micro-transacciones o un sistema de negocio basado en publicidad dentro del juego. Por ultimo podríamos estudiar la relación entre categoría y numero de descargas (en vez de numero de gasto) para también evaluar la posibilidad de que el beneficio venga con la mayor audiencia o si nos conviene más hacer énfasis en un mercado pequeño pero con mucho potencial económico.

Nombre de la base de datos: Coronavirus

Objetivo: Estudiar la evolución, propagación y mitigación del virus del covid a lo largo del mundo.

Ploblema planteado: Esta es la primera pandemia a la que nos enfrentamos a nivel global en un grado tan alto por lo tanto no estamos seguros de como actuar al respecto, dígase que tenemos la teoría de como afrontar una pandemia gracias a los epidemiólogos sin embargo desconocemos tanto el impacto como el alcance que tendrán la enfermedad y las acciones para contrarrestar la enfermedad, por lo tanto deberíamos estudiar la evolución de la misma con diversos puntos de vista.

Solución: Dado que contamos con datos de diferentes partes del mundo podemos estudiar el impacto que tienen las medidas sanitarias en países que la aplican vs países donde no se aplicaron en tanto grado (esto gracias a un algoritmo de búsqueda de secuencias) También podríamos estudiar la evolución del virus en diferentes pacientes dependiendo de su nivel demográfico (sexo,posición económica, nacionalidad) y así poder intentar crear un modelo que determine la posible evolución del virus en zonas donde aun no se llega al punto más alto (otra vez gracias a búsqueda de secuencias (y un poco de algoritmos de predicción)). Por ultimo podríamos intentar hacer un análisis que no vimos en clase pero considero seria de gran ayuda para este data set considerando la información con la que contamos, podríamos hacer un intento de análisis de redes basado en posición demográfica del paciente, y así intentar determinar que paciente es más probable que contagie a otras personas, esto incluso se podría expandir tratando como pacientes a ciudades enteras y así determinar el impacto que una ciudad tiene para que así por ejemplo ONU pueda determinar en que país es más relevante mandar ayuda para así evitar una propagación más grande a largo plazo o en nuestro caso determinar en que ciudades se debió de haber mandado ayuda de manera prioritaria y que ciudades pudieron haber esperado un poco.

Nombre de la base de datos: Vinos

Objetivo: Determinar la manera optima de ponerle precio a un vino / determinar la manera optima de comprar un vino

Ploblema planteado: Esta base de datos se puede abordar desde dos puntos de vista diferentes, el que vende el vino y el que lo compra ¿Por qué? Por que el objetivo del que vende el vino no necesariamente es vender el mejor vino si no maximizar sus ganancias,

mientras que el objetivo del que lo compra es uno un poco complejo el no quiere comprar un mal vino pero tampoco se ve en la necesidad de comprar el mejor vino (al menos en la mayoría de los casos) por lo que el objetivo del cliente es “Con la cantidad de dinero que tengo disponible ¿Cómo se cual es el mejor vino a mi alcance?”

Solución: Abordare primero el punto de vista del vendedor, el vendedor al desear maximizar ganancias podría hacer un estudio de búsqueda de secuencias multinivel, primero podría intentar encontrar una correlación entre el tipo de vino y la cantidad de compras para así decidir que vino le conviene más comprar, esto debería venir acompañado con el segundo nivel, encontrar un patron en el que la gente compra más vino para así maximizar el inventario en esas épocas. Por ultimo el vendedor podría intentar encontrar si el lugar de procedencia del vino influye en como lo ve el cliente para así importar vinos que en principio no tendría pero a cambio el cliente esta dispuesto a pagar más por ellos. Ahora desde el punto de vista del cliente, el cliente al intentar maximizar calidad podría intentar hacer un estudio que identifique la procedencia de los vinos con mayor puntaje vs precio (probablemente haciendo un cociente) para así intentar determinar de que lugar le conviene más comprar (o en su debido caso importar) después debería intentar hacer un estudio de si realmente influye el precio vs la cantidad de puntos que el vino tiene (esto seria ideal si se tuvieran los datos de las personas que tomaron vino sin saber cuanto costaba pero desconozco si en la base de datos viene así o si las personas fueron sesgadas con la información del costo) una vez que tiene la información de que lugar proviene el mejor vino y cual es la ventana de costos optima donde no sacrifica tanto sabor a cambio de obtener mucho ahorro ahí es donde el cliente debe comprar.

Nombre de la base de datos: Iris

Objetivo: Determinar los cambios evolutivos que ha tenido una misma especie subdivida en 3 “sub-especies”

Ploblema planteado: A veces es un poco difícil entender el concepto de evolución ¿Cómo es que venimos del mono si somos tan diferentes a el? Esto es un dato incorrecto dado que no provenimos del mono si no que compartimos un ancestro con el, y es ahí donde entra nuestro data-set, en el encontraremos diversos datos de 3 tipos de planta con el

mismo ancestro en común, sin embargo lo suficientemente diversos como para ser considerados sube-especies distintas, con el debido estudio podríamos intentar entender un poco más de lo que ya lo hacemos ahora el proceso de la evolución.
Solución: Podemos primero hacer un análisis exploratorios de los datos segmentado por especie para así determinar las variables donde se encuentre la mayor variabilidad dependiendo de la especie, posteriormente de eso debemos hacer un estudio de medida de esa variabilidad por especie ¿Qué tanto determina la especie el cambio en esa variable? Esto se puede hacer por muchos métodos ya sea por arboles de regresión, regresiones simples, sistema de clasificación. Posteriormente a eso debemos aplicar un meta análisis de los datos obtenidos ¿Cómo es que la variable que obtuvimos obtuvo tanto cambio en la especie que encontramos? ¿Qué ventaja obtiene esa especie por tener ese cambio tan brusco y por que no es relevante para las otras dos especies? ¿Qué tanto afecta la zona geográfica donde encontramos a esa especie en particular?.
Nombre de la base de datos: Netflix
Objetivo: Determinar horas y géneros específicos para publicar nuevo contenido.
Problema planteado: Netflix hace poco se aventuro al área de creación de contenido, esto es bueno dado que representa un factor diferenciador a netflix pero malo dado que hace nuevas preguntas que son necesarias de responder si se desea tener éxito comercial, tales como ¿En que horario la gente esta más dispuesta a que géneros? ¿Cuál es el genero que tiene más probabilidades de ser un éxito?
Solución: En este data set básicamente es búsqueda de secuencias aplicada de manera iterativa para determinar las respuestas necesarias, podríamos primero buscar el horario donde la gente ve más contenido (de manera general y a gran escala (día de la semana por ejemplo)) para así sacar el nuevo contenido por genero y día de la semana dependiendo de los días donde la gente este más activa, después de eso podríamos hacer un estudio a nivel personalizado donde a las personas se les recomiendan cosas más arriesgadas en los horarios que sabemos con anterioridad que estarán más activos (esto con el fin de conocer más a nuestros usuarios sin la necesidad de sacarlos de sus zonas de confort) y apuestas más seguras en los días donde el usuario este menos tiempo en la plataforma (con el objetivo de que el poco tiempo que tenga para aprovechar en Netflix esos días nos aseguremos que si lo gaste en Netflix) Por ultimo podríamos hacer un macro estudio de los géneros con mejores calificaciones por temporada para así lanzar contenido que vaya de acuerdo a la estación sin ser muy predecible e intentar buscar combinaciones que aunque tal vez suenen un poco locas al principio tengamos la certeza que a la larga saldrán rentables para la compañía.