

ST3

Integrantes:

- Alejandro Hidalgo Badillo **A01423412**
- Emmanuel Prieto Caro **A01423341**
- Miguel Jimenez Padilla **A01423189**

Parte 2:

¿Qué datos hay que seleccionar? Por qué.

Los datos relevantes para nuestro análisis incluyen los nombres de los desarrolladores, así como también si ofrecen compras dentro de la app o anuncios. Nos son relevantes ya que pensamos compararlos e intersecciónarlos para determinar que desarrolladores ofrecen estas cosas con más frecuencia

¿Hay que eliminar o reemplazar valores en blanco? **Sí** / No / Por qué.

Existen columnas que contienen valores redundantes o que no aportan a nuestro análisis, siendo esta información de contacto o números de identificación para las aplicaciones, para lo cual ya existen sus nombres.

¿Es posible agregar más datos? **Sí** / No / Por qué.

Sí debido a que es una base de datos de las aplicaciones más utilizadas en Google Play, si incrementamos la base de datos incrementaremos la cantidad de usuarios de distintas categorías de aplicaciones, aumentando variables y generando diferencias (lo mismo con los distintos datos). Podríamos generar más estadísticas obteniendo más datos de las aplicaciones, como número de usuarios activos, usuarios diarios, descargas u otros datos similares.

¿Hay que integrar o fusionar datos de varias fuentes? **Sí** / No / Por qué.

Se deben realizar intersecciones entre datos existentes de la tabla para poner estadísticas utilizables en nuestro análisis. Estas intersecciones en realidad son las que serán más relevantes.

¿Es necesario ordenar los datos para el análisis? **Sí** / **No** / Por qué.

No es necesario, debido a que el orden de nuestros datos no cambia el resultado de nuestro análisis, al ser valores booleanos o escalas.

¿Tengo que hacer conjuntos de datos para entrenamiento y prueba? **Sí** / No / Por qué.

11 may 2022

Para un modelo entrenado, como de regresión lineal, si tienes que hacer conjuntos de datos, debido a que si los datos son muy distintos a la hora de analizar el resultado será incoherente.

¿Qué ajustes se tuvieron que hacer a los datos (agregar, integrar, modificar registros (filas), cambiar atributos (columnas))?

Se quitan las columnas 'App Id', 'Developer Website', 'Developer Email', debido a que, ya sea por su naturaleza de ser información de contacto o por redundancia, no aportan ninguna información útil a nuestro análisis.

Parte 3:

¿Hay alguna variable que no aporte información?

Sí, la de los nombres de aplicación ID, y las informaciones de contacto de los desarrolladores. no tiene sentido comparar ese dato.

Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué?

El nombre de la aplicación y el ide de esta misma. No tiene sentido comparar estos datos, como también es el caso con los datos de contacto de los desarrolladores.

¿Existen variables que tengan datos extraños?

El ID de las aplicaciones contiene datos numéricos, caracteres y puntos, por lo que es difícil compararlo con otros datos.

Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?

Sí y no, dependiendo la posición que se encuentre la app, es su nivel de popularidad.