Challenge Option

Por Alexander González Ilufi

Contexto

- Para abordar el ejercicio se seguirán métricas del proceso KDD que implica etapas bien definidas que por objetivo apuntan a mejorar la calidad y valor de los datos.
- Esto mediante herramientas de Google Cloud Platform.
- A priori se sabe el objetivo del ejercicio y el dataset está definido, lo único adicional será recolectar metadata para el entendimiento del nombre de las columnas y significado de nomenclatura de su contenido.

Seleccionado los datos

Dataset:

- titanic.csv

Metadata:

- titanic_cheatsheet.csv
  - Posee información sobre el nombre de las columnas y el significado de valores para el caso de variables discretas.

**Se almacenarán en Google Storage*

Limpieza

- Para ello usaremos en primera instancia DataPrep.
- DataPrep entrega una visión bastante gráfica de como es el comportamiento de los datos, de lo cual inicialmente realizamos la siguiente estadística.
- Pero para efectos aplicativos, usaré BigTable.

Estadística inicial

Column	DataType	MissingValues	MismatchedValues	UniquesValues	Count
Survived	int	0	0	2	891
Pclass	int	0	0	3	891
Name	string	0	0	891	891
Sex	gender	0	0	2	891
Age	int	177	25	89	714
SibSp	int	0	0	7	891
Parch	int	0	0	7	891
Ticket	int	0	230	681	891
Fare	decimal	0	0	248	891
Cabin	string	687	0	148	204
Embarked	string	2	0	4	889

Limpieza

- Convertimos el data **type** del campo ticket en string.
- Estandarizamos la columna **sex**: male = 0 ; female = 1
- Para el completado de los datos nulos, en el caso del campo **Age,** usaremos la mediana (claramente existen mejores formas de abordar este punto, como relacionando titulo nombre, clase…, usando regresiones lineales, KNN, entro otras).
- El campo **Cabin** contiene un 77,1% de *missingValues*, por lo que se completará con 0 y se dejará fuera para efectos de análisis.
- El campo **Embarked** se reemplazará con la mediana.

Estructuración y despliegue

- Usando BigQuery generamos diferentes vistas sobre nuestra sabanas de datos inicial, en este caso, tendremos una con los datos, otra que explican las columnas, y la última con la estadística valiosa para análisis…no terminado

No terminado

- Distribución por sexo

sex	not_survived	survived
male	468	109
female	81	233

- Distribución por clase

Pclass	survived	not_Survived
1	136	80
2	87	97
3	119	372

Análisis descriptivo

- Sobrevivieron más mujeres que hombres.
- La distribución del dataset no es simétrica y la mayoría de los pasajeros correspondía al grupo entre 20-40 años.
- El 47,2% de la clase 2 sobrevivió; El 62,9% de la clase 1 sobrevivió; el 24,2% de la clase 3 sobrevivió.
- …..