### **Integrantes:**

* **Marcelo Médez**
* **Juan Pablo González.**
* **Michelle Amaro**
* **Gustavo Garcia**
* **Christiane Voigt**

### 1- Realizar un análisis descriptivo utilizando SQL. Aplique medidas de centralidad y dispersión para cada una de las variables.

| SELECT  race,sex,  AVG(age) as AVR\_age,  min(age) as MIN\_age,  max(age) as MAX\_age,  stddev(age) as STD\_age,  AVG(education\_num) as AVR\_education\_num,  min(education\_num) as MIN\_education\_num,  max(education\_num) as MAX\_education\_num,  stddev(education\_num) as STD\_education\_num,  AVG(capital\_gain) as AVR\_capital\_gain,  min(capital\_gain) as MIN\_capital\_gain,  max(capital\_gain) as MAX\_capital\_gain,  stddev(capital\_gain) as STD\_capital\_gain,  AVG(capital\_loss) as AVR\_capital\_loss,  min(capital\_loss) as MIN\_capital\_loss,  max(capital\_loss) as MAX\_capital\_loss,  stddev(capital\_loss) as STD\_capital\_loss  from `bigquery-public-data.ml\_datasets.census\_adult\_income`  group by race,sex |
| --- |

#### resultado

<https://docs.google.com/spreadsheets/d/1jL5GBKJVo6pcvOFBsC9Cp4sOgv6hGoHufwGpm9ob_54/edit?usp=sharing>

Análisis

Para analizar la información se separó en raza y sexo, determinando que no existe diferencias significativas en las edades, sin embargo, en los años máximos de estudio se observa una diferencia entre razas y sexo. Por otro lado, se determina que existe que existe una diferencia importante en el capital por rza y sexo, donde detacan los hombres de raza asiatica, superendo en mas de tres veces al menor valor.

### 2 - Genere nuevas features, puede utilizar alguna transformación definida por el estudiante o ocupar funciones ya implementadas en BQ (FEATURE\_CROSS, QUANTILE\_BUCKETIZE, entre otras).

| SELECT ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  from `bigquery-public-data.ml\_datasets.census\_adult\_income` |
| --- |

#### resultado

<https://docs.google.com/spreadsheets/d/10RY2YJc_SaiM_NT_SrVkEs7kNHWUlthEYBDhdkxxPyU/edit?usp=sharing>

Análisis

Se generó una nueva variableque combinó las clases de sexo y raza, que se justifica con los datos anteriormente determinados.

### 3-Cree un modelo utilizando una regresión logística y muestre los pesos de la regresión.

| CREATE OR REPLACE MODEL `tarea1.modelo\_regre\_log`  OPTIONS ( MODEL\_TYPE='LOGISTIC\_REG',  L1\_REG=5,  MAX\_ITERATIONS=5,  input\_label\_cols=['income\_bracket']) AS  SELECT  age,  education\_num,  capital\_gain,  functional\_weight,  capital\_loss,  ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  income\_bracket  FROM `bigquery-public-data.ml\_datasets.census\_adult\_income` |
| --- |

#### métrica



#### Peso

| SELECT  \*  FROM  ML.WEIGHTS(MODEL `tarea1.modelo\_regre\_log`) |
| --- |

resultado

<https://docs.google.com/spreadsheets/d/14DCPAYyCsgBVta6ra5h8SqFCKNOwGoihQJLakmnaf5c/edit?usp=sharing>

Análisis

Del resultado de las métricas que el modelo presenta una precisión de 0,72, una exactitud de 0,80, que representa un modelo que puede predecir con un nivel adecuado el Target.

De los resultados de los pesos de las variables, las variables de tipos de empleo presenta pesos negativos con diferencias importantes en sus valores que muestra su peso en la determinación de los ingresos, La variable que presenta el mayor peso es la variable años de estudio.

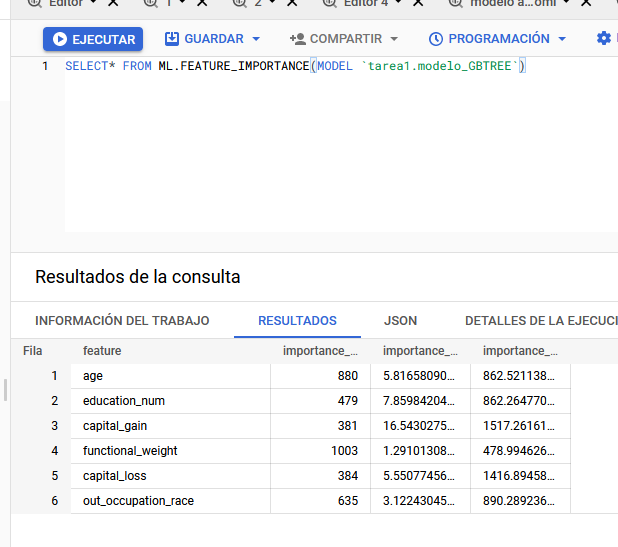
### 4- Crear un modelo utilizando Boosting y ver la feature importance ordenando por ganancia.

| CREATE OR REPLACE MODEL `tarea1.modelo\_GBTREE`  OPTIONS(MODEL\_TYPE='BOOSTED\_TREE\_CLASSIFIER',  BOOSTER\_TYPE = 'GBTREE',  NUM\_PARALLEL\_TREE = 1,  MAX\_ITERATIONS = 50,  TREE\_METHOD = 'HIST',  EARLY\_STOP = FALSE,  SUBSAMPLE = 0.85,  INPUT\_LABEL\_COLS = ['income\_bracket'])  AS  SELECT  age,  education\_num,  capital\_gain,  functional\_weight,  capital\_loss,  ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  income\_bracket  FROM `bigquery-public-data.ml\_datasets.census\_adult\_income` |
| --- |

#### métrica



#### feature importance



Análisis

1. functional\_weight
2. age
3. out\_occupation\_race
4. education\_num
5. capital\_loss
6. capital\_gain

### 5- Repita el paso 3 y 4 tratando el problema de desbalance de clase.

#### Solución desbalance regresión logística

| CREATE OR REPLACE MODEL `tarea1.modelo\_regre\_log\_2`  OPTIONS ( MODEL\_TYPE='LOGISTIC\_REG',  L1\_REG=5,  MAX\_ITERATIONS=5,  input\_label\_cols=['income\_bracket']) AS  SELECT age,  education\_num,  capital\_gain,  functional\_weight,  capital\_loss,  ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  income\_bracket  from `bigquery-public-data.ml\_datasets.census\_adult\_income`  WHERE income\_bracket LIKE '%<%' AND rand() < 0.3  UNION ALL  SELECT age,  education\_num,  capital\_gain,  functional\_weight,  capital\_loss,  ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  income\_bracket  from `bigquery-public-data.ml\_datasets.census\_adult\_income`  WHERE income\_bracket LIKE '%>%' |
| --- |

Código:

Se determina que la clase <=50K representa aproximadamente un 70% del total del universo de información , por tanto se selecciona en forma aleatoria el 30% de esta clase, Se selecciona toda la información de la clase >50K y finalmente se unen y con ello se equilibra la base.

#### Comparación

| métrica regre log antes | métrica regre log solución |
| --- | --- |
|  |  |

Del resultado de las métricas se determina que el equilibrar las clases impacto en el resultado de la métrica,destacando la recuperación el cual se duplicó, la puntuación F1 que pasa de un valor de 0,44 a 0,76 y la pérdida logísticas pasa de un valor de 0.43 a 0,52, por lo cual el modelo posee una mejor capacidad de predicción sobre todo en la determinación de la clase >50K

#### Solución desbalance boosting

| CREATE OR REPLACE MODEL `tarea1.modelo\_GBTREE\_2`  OPTIONS(MODEL\_TYPE='BOOSTED\_TREE\_CLASSIFIER',  BOOSTER\_TYPE = 'GBTREE',  NUM\_PARALLEL\_TREE = 1,  MAX\_ITERATIONS = 50,  TREE\_METHOD = 'HIST',  EARLY\_STOP = FALSE,  SUBSAMPLE = 0.85,  INPUT\_LABEL\_COLS = ['income\_bracket'])  AS  SELECT  age,  education\_num,  capital\_gain,  functional\_weight,  capital\_loss,  ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  income\_bracket  from `bigquery-public-data.ml\_datasets.census\_adult\_income`  WHERE income\_bracket LIKE '%<%' AND rand() < 0.3  UNION ALL  SELECT  age,  education\_num,  capital\_gain,  functional\_weight,  capital\_loss,  ML.FEATURE\_CROSS(STRUCT(occupation, race)) AS out,  income\_bracket  from `bigquery-public-data.ml\_datasets.census\_adult\_income`  WHERE income\_bracket LIKE '%>%' |
| --- |

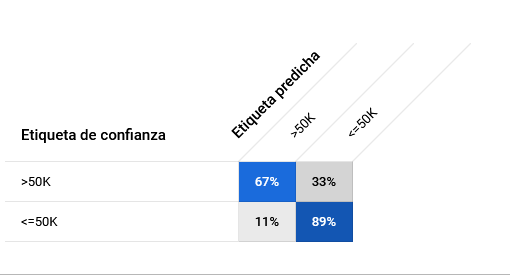
#### Comparación

| métrica boosting antes | métrica boosting solución |
| --- | --- |
|  |  |

Para el módelo Boosting el equilibrar la base impacta positivamente en la métrica,destacando la recuperación el cual se duplicó, la puntuación F1 que pasa de un valor de 0,61 a 0,79 y la pérdida logísticas pasa de un valor de 0.33 a 0,42, por lo cual el modelo posee una mejor capacidad de predicción sobre todo en la determinación de la clase >50K

### 6- Compare los 4 modelos creados a través de una matriz de confusión, vea cómo cambian los resultados del modelo de Boosting cambiando los hyper-parámetros.

|  | Antes | Solución |
| --- | --- | --- |
| REGRE LOG |  |  |
| BOOSTING |  |  |



De los resultados de las matrices de confusión se infiere que el equilibrar la base para la determinación del modelo impacta en la predicción de la clase >50 disminuyendo los valores falsos negativos, sin embargo, el valor de los falsos positivos aumenta.