**Supervised Learning: Classification**

**Section 1.** **Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.**

Main objective of the analysis is prediction of Bank Customer Churn and finding the most important feature on determining the quality. This analysis will help to producing Customer Churn with fine quality.

**Section 2.** **Brief description of the data set you chose and a summary of its attributes.**

**Bank Customer Churn Dataset**

Dataset is related to ABC Multistate Bank. Dataset link from UCI machine learning repository:

<https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset>

Input Variables:

* customer\_id
* credit\_score
* country
* gender
* age
* tenure
* balance
* products\_number
* credit\_card
* active\_member
* estimated\_salary

Output

* churn

Text

Description automatically generated**Section 3.** **Data exploration and actions taken for data cleaning and feature engineering.**

Graphical user interface, application

Description automatically generatedText

Description automatically generatedChecking for missing value. Checking for duplicated column.

**# Dataset is not including any unique column. So, I don’t delete any column.**

Checking for value in column which is an object type.

Graphical user interface, application

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Graphical user interface, text, application

Description automatically generatedText

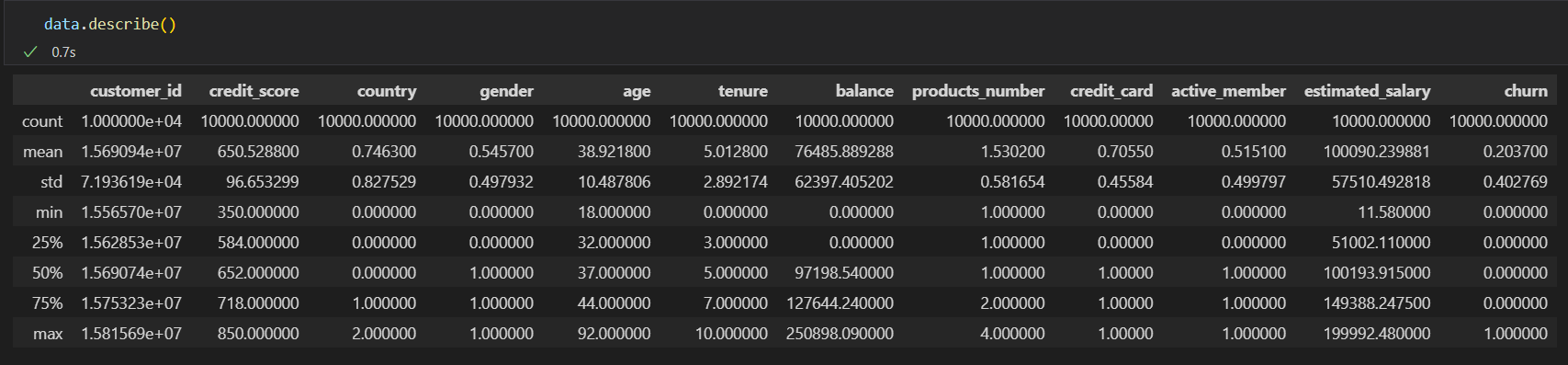
Description automatically generated

**# Encoded the object value with LabelEncoder from sklearn.preprocessing**

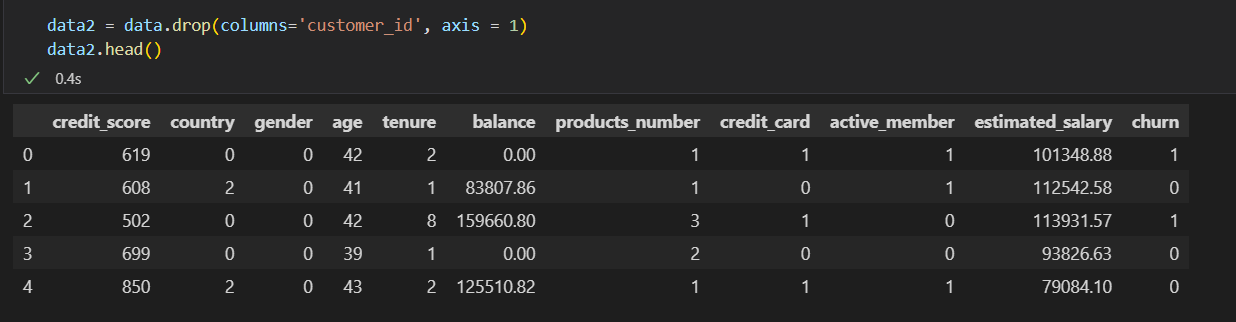
**Text

Description automatically generated**

Checking for statistics value of dataset.



Drop some useless column.



Analyzing features importance

Chart

Description automatically generatedText

Description automatically generated

**Summary of data exploration and actions taken for data cleaning and feature engineering.**

After the data exploration there’s no need to any action for data cleaning. The dataset has no missing value. Besides that, there are 2 columns which are object value, so I use a Label Encoder to encode this value to continuous value. I will do scaling the values.

**Section 4. Scaling, Train-Test Split and Classification**

Text

Description automatically generatedTrain-Test Split and Scaling

Classification

Text

Description automatically generated

The Accuracy and F1-score for each model

* Logistic Regression

Text

Description automatically generated

* Text

  Description automatically generatedDecision Tree
* K Nearest Neighbors

Text

Description automatically generated

* Text

  Description automatically generatedRandom Forest
* Gradient Boosting

Text

Description automatically generated

The Confusion Matrix

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

Chart, treemap chart

Description automatically generated

**Section 5.** **Final Model According to the results Random Forest Classifier is the best model for this dataset.**

According to the results Gradient Boosting Classifier is the best model for this dataset.

**Section 6. Results**

Considering correlations between features, age is the key features to determine customer churn. Accuracy is %88 and F1-Score is %62 for this classification. For better results, hyperparameter tuning to the models will increase accuracy.