Project Title: Bank Churn Prediction
Author: James Imbuido

Project Overview:

- Customer churn in the banking industry poses significant financial and reputational risks. Acquiring new customers is often more expensive than retaining existing ones. Predicting customer churn with high accuracy allows banks to proactively address customer dissatisfaction and implement targeted retention strategies.
- Traditional churn prediction models in banking typically rely on customer demographics, account activity, product usage, and financial behavior. While these factors provide valuable insights, they may overlook the potential influence of geographic and spatial patterns on customer churn.
- The machine learning approach used for this project was supervised as it was deemed most appropriate with the dataset.

Data:

- The dataset chosen for this project is titled 'Bank Customer Churn Dataset' from Kaggle (https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset). This was a dataset supplied by Gaurav Topre, which is best suited for classification applications.
- The variables involved were 'customer_id', 'credit_score', 'country', 'gender', 'age', 'tenure', 'balance', 'products_number', 'credit_card', 'active_member', 'estimated_salary', and 'churn'.
- 'customer_id' was distinguished to be insignificant in the analysis and hence it was omitted from the final list of feature variables. The feature variables used included: 'credit_score', 'country', 'gender', 'age', 'tenure', 'balance', 'products_number', 'credit_card', 'active_member', and 'estimated_salary'. On the other hand, the chosen response variable was 'churn'.
- The feature variables are essentially relevant bank characteristics of the customers that would better inform a pattern and/or reason for their churn status, hence why the variables were allocated as so.
- Upon checking in on the raw data, there was no need for further pre-processing of the dataset.

Model Development:

- The classification algorithms used in this project were Logistic Regression, Support Vector Machines (SVM), Random Forests (RF), and Naïve Bayes.

Evaluation:

- The metrics used to evaluate these models involved a coefficient plot, a confusion matrix, and ROC curve.