# BANK CUSTOMER CHURN

EXPLORATORY DATA ANALYSIS & PREDICTION STUDY

## GOAL

- Identify and visualize which factors contribute to customer churn;
- Build a prediction model to classify if a customer will churn or not.

## DATA

From the data collected, we question:

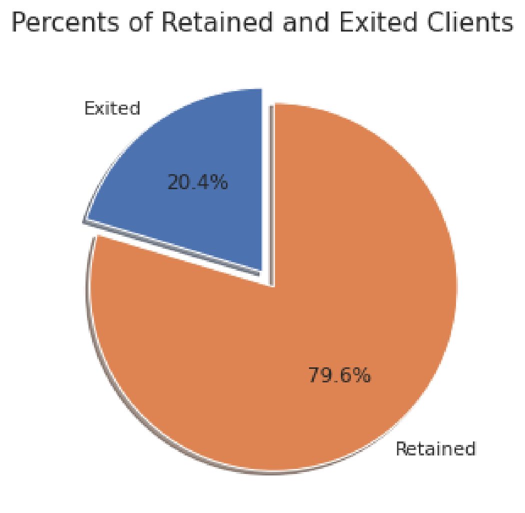
- Data is not a time series and shows only a specific point in time.The balance feature is from a given date and leaves questions:
  - What is the date, and what relevant event happened on it?
  - Would it be possible to obtain balances over time for better analysis?
  - Some clients have exited but still, have a positive balance. Could they have exited from a product and not the bank?

**Note:**

For this study, data were analyzed without context. In this field, context is fundamental to better understanding and modeling a more precise prediction system.

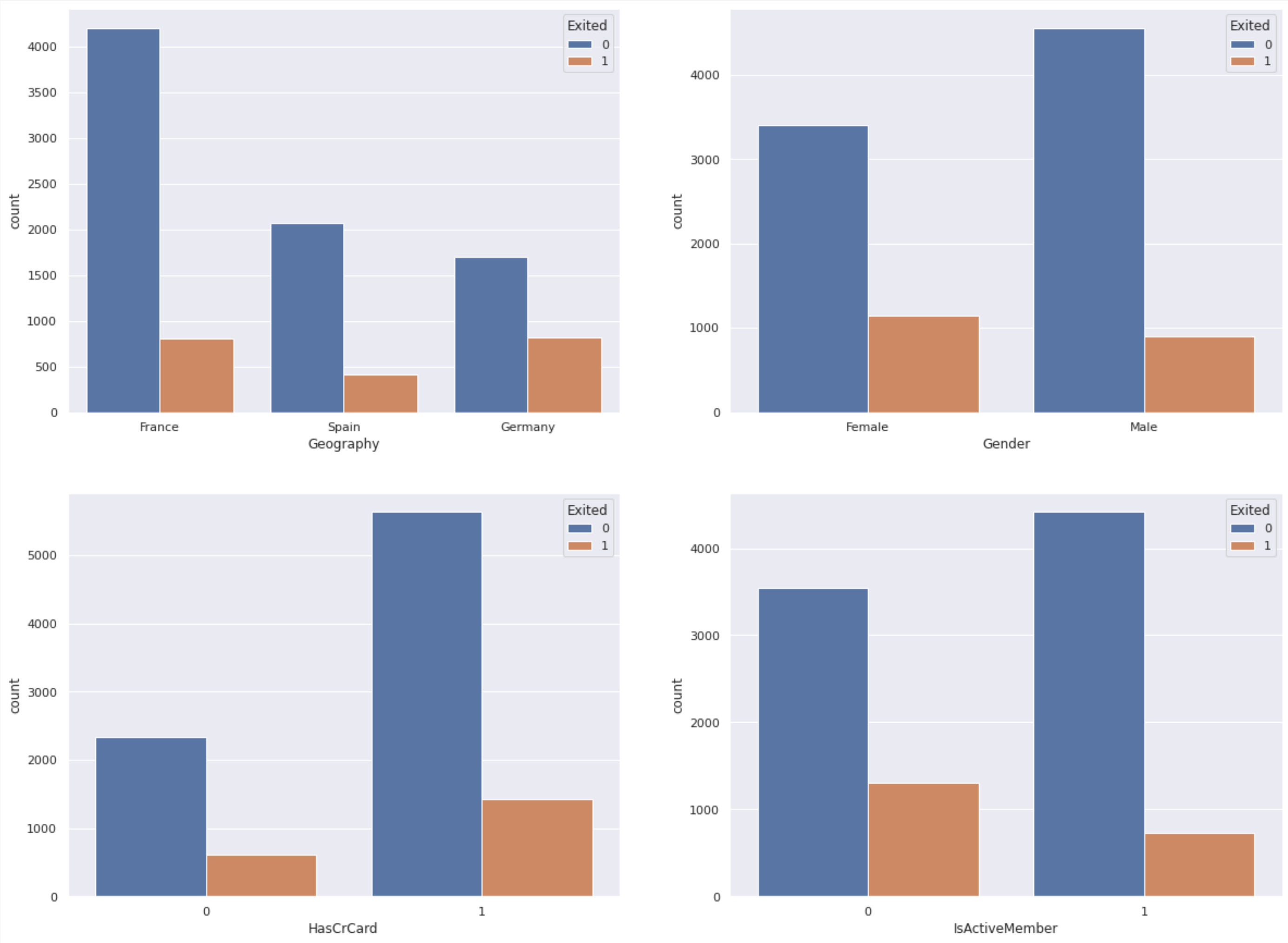
## EXPLORATORY DATA ANALYSIS

This section's goal is to understand how the features relate to the 'Exit' status.



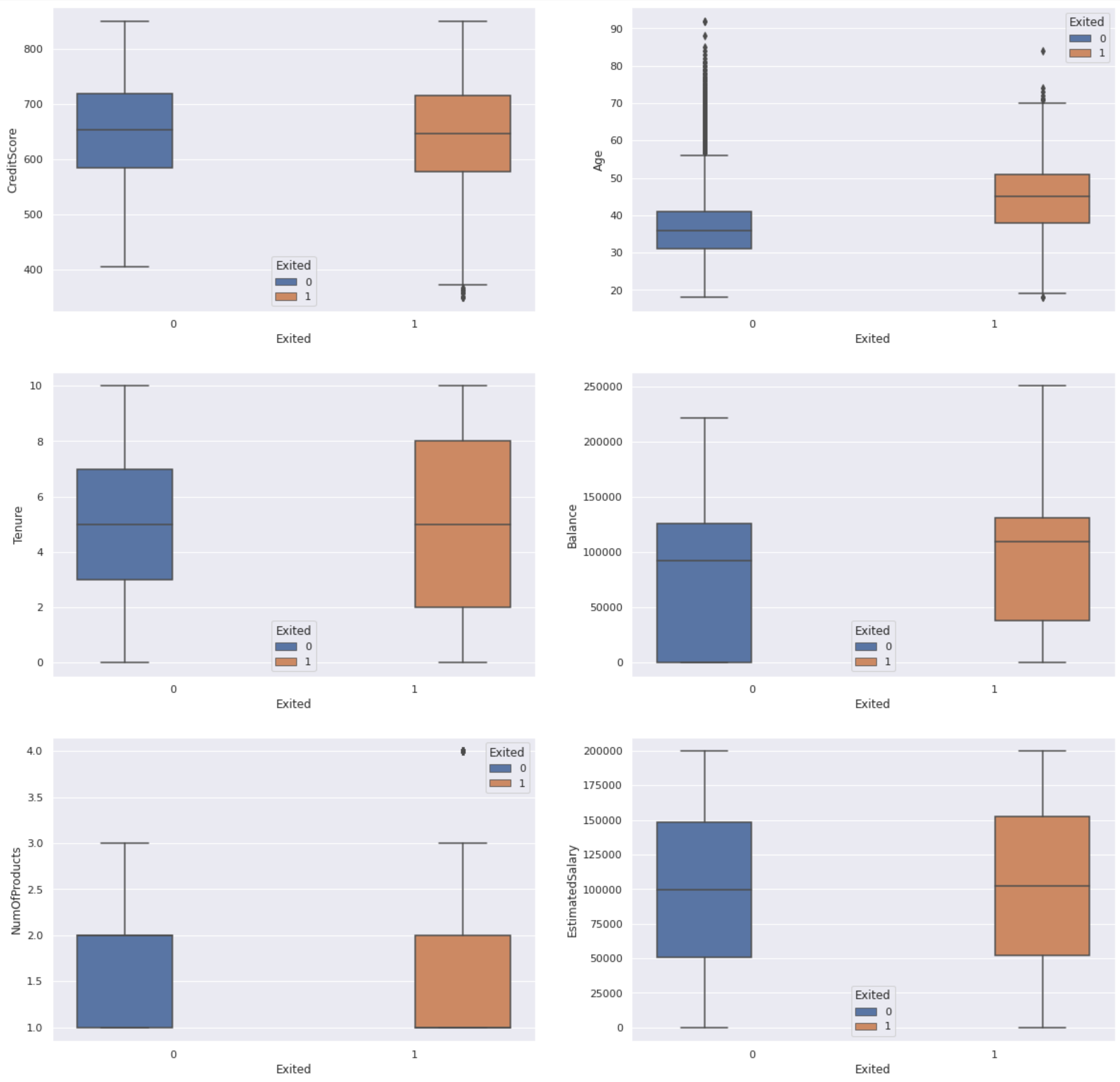
About 20% of the clients churned.

The prediction model needs to ensure more than 80% accuracy to be efficient enough for the bank.



**Exploratory Categorical Features Graphs show:**

- The majority of the clients are from France and the proportion of churned is inversely related to the number of clients. This could indicate bank problems such as not having enough customer service resources in the areas where it has fewer clients.
- Even representing the less count proportion of clients, females churn more than male clients.
- The majority of the clients that churned have credit cards, but proportionally, it doesn't look significant.
- The inactive members have a greater churn. Surprisingly the proportion of inactive members is high. A good strategy for the bank would be a program to turn this group into active clients, as this could have a positive impact on client churn.

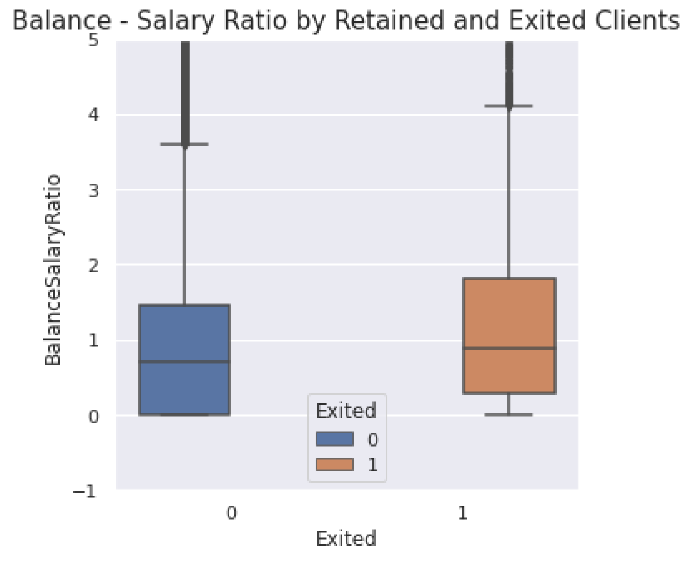


**Exploratory Continuous Features Graphs show:**

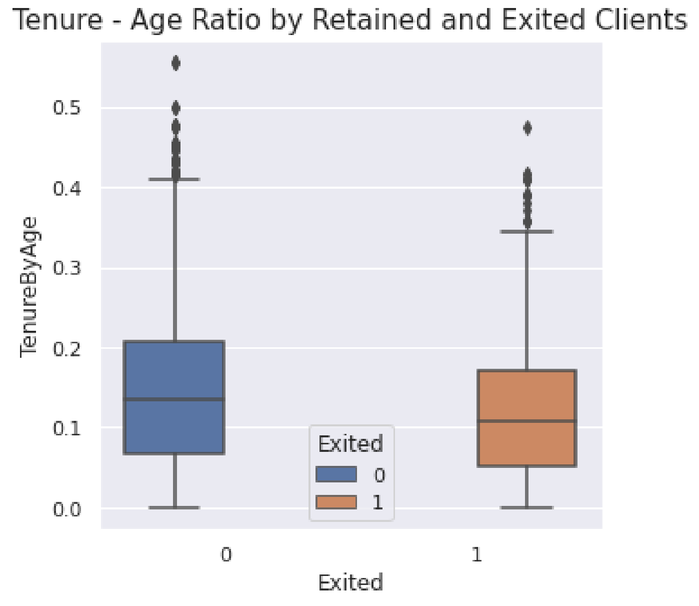
- There is no significant difference in credit score distribution between retained and exited clients.
- Older clients are churning more than younger ones. This could be due to differences in services by age. The bank should review its strategy of retention based on age groups.
- Clients on either extreme of tenure are more likely to churn compared to those on average.
- The bank is losing clients with better balances.
- Neither product nor salary has a significant effect on churn.

## GENERATED FEATURES

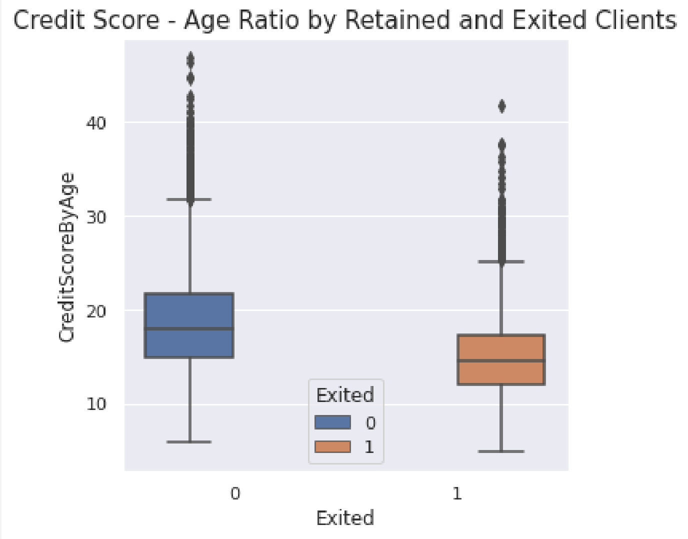
Generating features to better explore what impacts more on client churn



Salary has little effect on the chance of a client to churn. However, the ratio of bank balance and estimated salary, indicates that clients with a higher balance salary ratio churn more. Worrying information to the bank as this impacts their source capital.

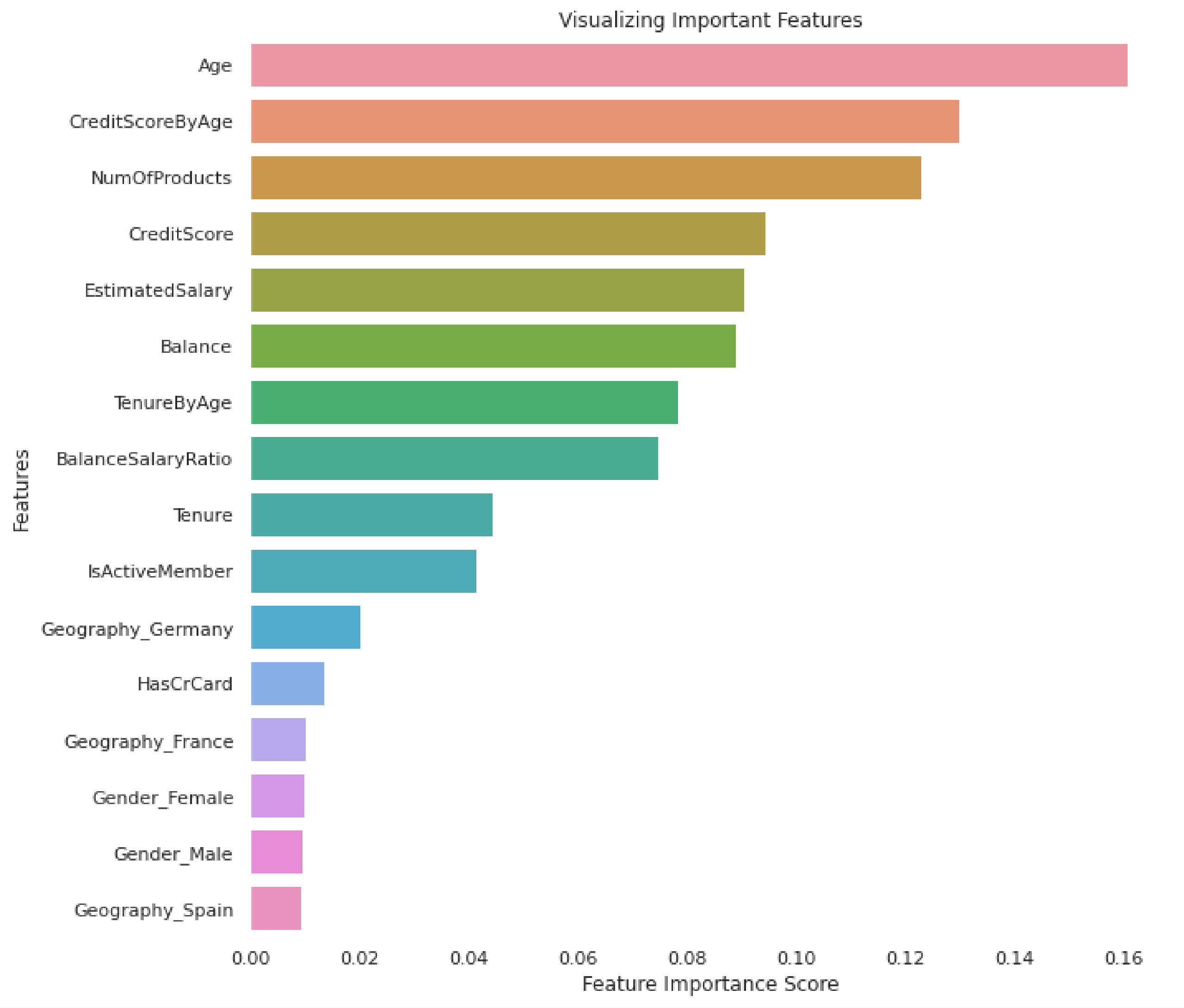


Tenure by age shows the effect of age on churn.



Credit Score by age shows the effect of age on churn.

## DATA PREPARATION FOR MODEL FIT



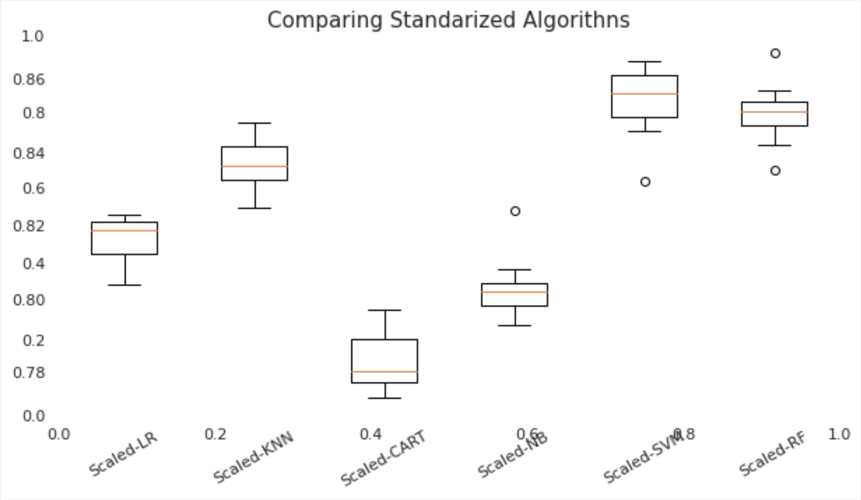
This graph list features that have more impact on churn

- Age, Credit Score by Age, and Number of Products are the three most influential features on clients' churn.
- For this study, we selected features that affect at least 30%.

## MODEL FITTING AND SELECTION

We tested the following models:

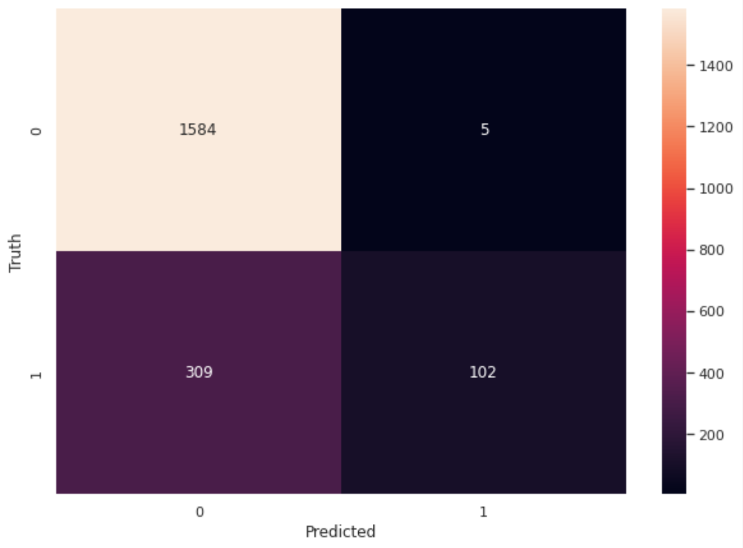
- LR - Linear Regression
- KNN - K-Nearest Neighbors
- CART - Classification and Regression Trees
- NB - Naive Bayes
- SVM - Support Vector Machine
- RF - Random Forest



The most efficient models were Support Vector Machine (SVM) with 85.37% accuracy followed by Random Forest (RF) with 84.81% accuracy.
After trying different settings to optimize both and try to improve their metrics, we got the following result:

- Support Vector Machine (SVM) with 85.45% accuracy
- Random Forest (RF) with 85.10% accuracy

## DEEP LEARNING APROACH



**Deep Learning (DL) showed impressive results with 85.80% accuracy.**

After fine-tuning, the Support Vector Machine (SVM) showed 85.45% accuracy.
Both models are timing-consuming.

**For long-term use, probably the Deep Learning approach would be more accurate and time-efficient.**