****

**MOBILE MONEY AND FINANCIAL INCLUSION IN TANZANIA.**

**TABLE OF CONTENTS.**

[**LIST OF FIGURES.** 3](#_Toc112954479)

[**1.** **INTRODUCTION** 4](#_Toc112954480)

[**1.1.** **BACKGROUND.** 4](#_Toc112954481)

[**1.2.** **PROBLEM** 4](#_Toc112954482)

[**1.3.** **OBJECTIVE** 4](#_Toc112954483)

[**1.3.1.** **Main Objective** 4](#_Toc112954484)

[**1.3.2.** **Specific Objectives** 4](#_Toc112954485)

[**1.4.** **TARGET** 5](#_Toc112954486)

[**1.5.** **PROJECT IMPACT** 5](#_Toc112954487)

[**1.6.** **TEAM MEMBERS** 5](#_Toc112954488)

[**2.** **DATA** 6](#_Toc112954489)

[**2.1.** **DATA SOURCES** 6](#_Toc112954490)

[**2.2.** **DATA CLEANING** 6](#_Toc112954491)

[**2.2.1.** **Missing values.** 6](#_Toc112954492)

[**2.2.2.** **Outliers** 7](#_Toc112954493)

[**2.2.3.** **Mis-Value type.** 7](#_Toc112954494)

[**3.** **METHODOLOGY** 8](#_Toc112954495)

[**3.1 EXPLORATORY DATA ANALYSIS (EDA)** 8](#_Toc112954496)

[**3.1.1.** **Hypothesis generation** 8](#_Toc112954497)

[**3.1.2.** **Univariate analysis** 8](#_Toc112954498)

[**3.1.3.** **Bivariate analysis** 10](#_Toc112954499)

[**3.2.** **FEATURE ENGINEERING** 11](#_Toc112954500)

[**3.2.1.** **Dropping columns** 11](#_Toc112954501)

[**3.2.2.** **Dataset split** 11](#_Toc112954502)

[**3.2.3.** **Label encoding of binary categorical features** 12](#_Toc112954503)

[**3.2.4.** **One hot encoding** 12](#_Toc112954504)

[**3.2.5.** **Continuous features; min-max normalization** 13](#_Toc112954505)

[**3.2.6.** **Feature selection** 14](#_Toc112954506)

[**3.3.** **MODEL SELECTION AND EVALUATION** 15](#_Toc112954507)

[**4.** **MODEL DEPLOYMENT** 21](#_Toc112954508)

[**4.1.** **CREATING FRONTEND** 22](#_Toc112954509)

[**4.2.** **Running Web App** 23](#_Toc112954510)

[**4.3.** **Pushing deployment folder on github** 25](#_Toc112954511)

[**5.** **CONCLUSION** 26](#_Toc112954512)

[**6.** **REFERENCES** 26](#_Toc112954513)

# **LIST OF FIGURES.**

[Figure 1 showing data shape 6](#_Toc112954514)

[Figure 2 checking for the presence of missing values in our dataset 6](#_Toc112954515)

[Figure 3 Showing outliers in our dataset 7](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954516)

[Figure 4 Showing all columns with outliers being dropped 7](#_Toc112954517)

[Figure 5 showing target distribution exploration 9](#_Toc112954518)

[Figure 6 showing hypothesis 1 testing by bivariate analysis 10](#_Toc112954519)

[Figure 7 showing dropping of unnecessary columns 11](#_Toc112954520)

[Figure 8 splitting data 12](#_Toc112954521)

[Figure 9 replacing 2 with 0 12](#_Toc112954522)

[Figure 10 one hot encoding 13](#_Toc112954523)

[Figure 11 min max scaler 13](#_Toc112954524)

[Figure 12 best feature selection 14](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954525)

[Figure 13 model selection 16](#_Toc112954526)

[Figure 14 classifiers accuracy 17](#_Toc112954527)

[Figure 15 voting classifier 17](#_Toc112954528)

[Figure 16 model evaluation 18](#_Toc112954529)

[Figure 17 confusion matrix 19](#_Toc112954530)

[Figure 18 saving model by joblib 21](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954531)

[Figure 19 loading model 21](#_Toc112954532)

[Figure 20 prediction and results 22](#_Toc112954533)

[Figure 21 number\_input and selectbox 22](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954534)

[Figure 22 adding banner image 22](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954535)

[Figure 23 run web app 23](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954536)

[Figure 24 streamlit app 24](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954537)

[Figure 25 streamlit app cont.. 25](file:///C:\Users\Candy\Desktop\MOBILE%20MONEY%20AND%20FINANCIAL%20INCLUSION%20IN%20TANZANIA%20finaalll.docx#_Toc112954538)

# **INTRODUCTION**

## **BACKGROUND.**

In early 2008, the FinTech company E-Fulusi, launched the first mobile wallet in Tanzania called MobiPawa, shortly followed by the launch of M-Pesa by Vodacom in April 2008, then followed by other financial services like Tigo Pesa, Airtel Money and Z-Pesa (Zantel).

Tanzania financial inclusion was spearheaded by the enactment of the banking and financial institutions Act in 1991, which allowed the entrance of private sector (both local and foreign) in the Tanzanian financial sector. The new Act also paved way for the introduction of financial markets in the economy.

## **PROBLEM**

A small percent of people in Tanzania have bank accounts and a lot more of Tanzanians who don’t have bank accounts do other types of formal financial services, primarily, mobile money.

Mobile money and other financial service providers face difficulties in identifying which people are more likely to use their services. This hinders the proper optimization of their resources and services to their clients.

By predicting which individuals are most likely to use mobile money, a precise target area can also be attained.

This problem is a classification problem whereby, we have to predict which individuals are most likely to use mobile money.

## **OBJECTIVE**

### **Main Objective**

The main objective of this project is to create a solution that will help mobile money service providers predict which individuals are most likely to use mobile money and other financial services (savings, credit and insurance).

### **Specific Objectives**

The sum of these processes meet the general objective as a result, and these processes are;

* To obtain a ‘mobile money and financial inclusion in Tanzania’ dataset.
* To visualize and perform Exploratory Data Analysis (EDA) of the dataset.
* To perform featureengineeringand feature selection.
* To create and evaluate a machine learning model that will be used to reach the main objective.

## **TARGET**

This project aims at mobile money service providersand other financial service providers.

## **PROJECT IMPACT**

* This solution can help mobile money providers target new clients and markets across Tanzania more effectively and hence increase profits among their business organizations.
* It can also help other financial service providers cross-sell other financial services (saving, credit and insurance) to the existing mobile money customer base so that they can maximize their profit by providing maximum services accordingly.
* Also, it will provide better financial conveniencefor the said markets and clients.

## **TEAM MEMBERS**

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **FULL NAME** | **COLLEGE NAME** | **YEAR OF STUDY** |
| 1 | Candy Jeremiah | UDSM-CoICT | 3 |
| 2 | Pawa .N. Makeresia | UDSM-CoNAS | 2 |
| 3 | Hussein Mzenzi | KAMPALA INTERNATIONAL UNIVERSITY | 3 |

# **DATA**

## **DATA SOURCES**

The dataset was obtained from **Zindi**.

Zindi is a platform that hosts the largest community of African data scientists, working to solve the world’s most pressing challenges using machine learning and AI.

Our dataset is available via: <https://zindi.africa/competitions/mobile-money-and-financial-inclusion-in-tanzania-challenge/data>.

It contains 7094 rows and 37 columns as can be seen in the figure below;

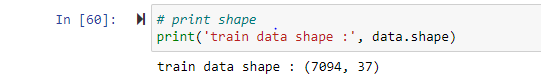


Figure showing data shape

## **DATA CLEANING**

Data cleaning refers to the removing of outliers, missing values, mis-value types and/or other things that reduce the quality of the data.

### **Missing values.**

There were no missing values in our dataset, as can be seen below;

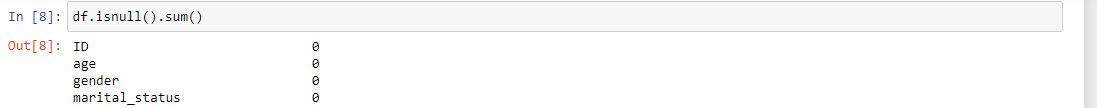


Figure checking for the presence of missing values in our dataset

### **Outliers**

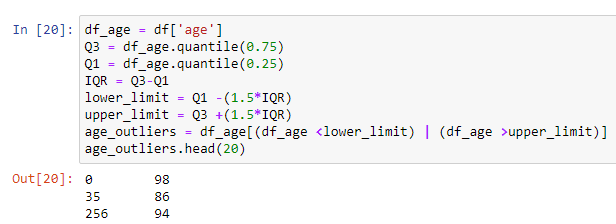
Outliers are values that are much smaller or larger than most of the other values in a dataset. Due to this, they tend to affect the mean of the data in a negative way. In our dataset, the age column contained outliers, as can be seen below;

Figure Showing outliers in our dataset

Following this, all the columns containing outliers were dropped;

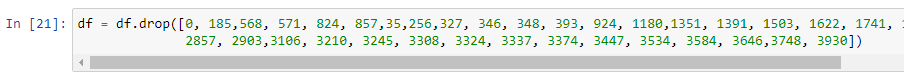


Figure Showing all columns with outliers being dropped

### **Mis-Value type.**

Refers to the situation in which the value of a data does not match with its type. For example, the value of ‘age’, which is numerical, being a float or string type instead of an integer type.

There were no mis-value types in our dataset.

# **METHODOLOGY**

## **3.1 EXPLORATORY DATA ANALYSIS (EDA)**

This refers to an approach of analyzing a dataset to summarize their main characteristics, often using statistical graphics and other data visualization methods.

### **Hypothesis generation**

Hypothesis generation is a very important stage in any data science/machine learning pipeline. It involves understanding the problem in detail by brainstorming as many factors as possible which can impact the outcome. It is done by understanding the problem statement thoroughly and, before looking at the data.

The following were the hypotheses generated;

* People with age above 60 will not have both mobile money and other financial service.
* People with lower education level are less likely to possess mobile money.
* People with mobile phones are most likely to use mobile money.
* Almost everyone who gets money through salaries/wages is less likely to use mobile money only.
* Almost everyone who works for government has insurance.

Also, EDA has a lot of categories. The two most common/main are;

* Univariate Analysis
* Bivariate Analysis

### **Univariate analysis**

This is the simplest form of analyzing data where only one variable is involved. Its main purpose is to describe.

For example, below is a figure showing a univariate analysis for our **target**, mobile\_money\_classification;

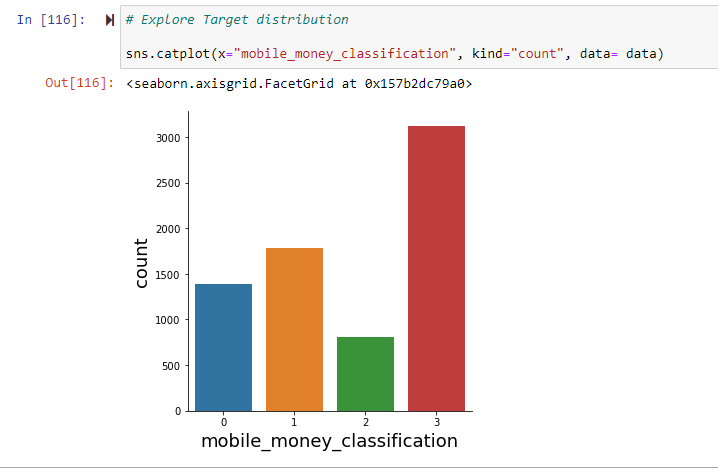


Figure showing target distribution exploration

##### Where;

0 - no mobile money and no other financial service (saving, borrowing, insurance)

1 - no mobile money, but at least one other financial service

2 - mobile money only

3 - mobile money and at least one other financial service

The analysis shows that we have a large number 3 class and the smallest number of 2 class

### **Bivariate analysis**

This involves the relationship between two variables for the purpose of determining the empirical relationship between them. It is helpful in testing hypotheses.

For example, from the figure below;

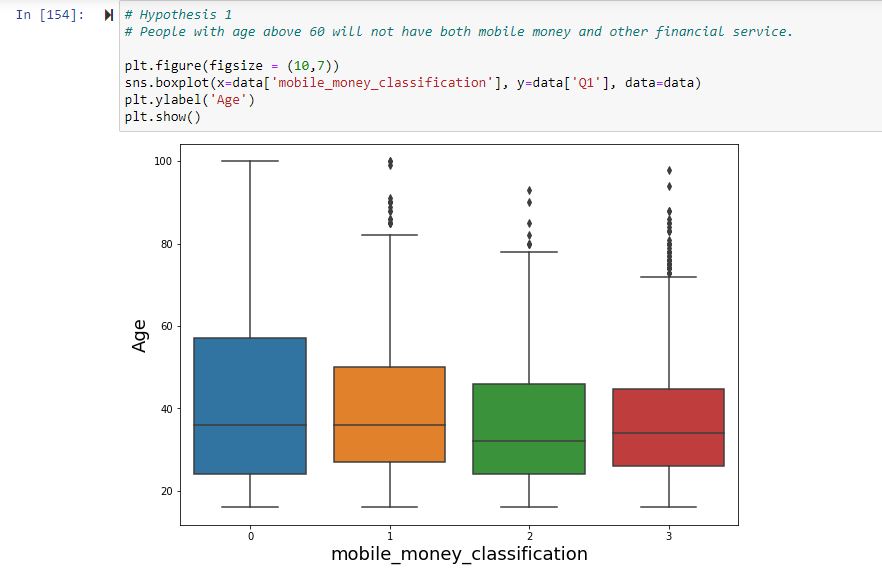


Figure showing hypothesis 1 testing by bivariate analysis

##### Where;

0 - no mobile money and no other financial service (saving, borrowing, insurance)

1 - no mobile money, but at least one other financial service

2 - mobile money only

3 - mobile money and at least one other financial service

Hence result; Hypothesis 1 = TRUE.

## **FEATURE ENGINEERING**

Feature engineering refers to a process of selecting and transforming variables/features when creating a predictive modelusing machine learning.

Feature engineering has two goals:

* Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
* Improving the performance of machine learning models.

### **Dropping columns**

In the dataset of Mobile money and financial inclusion in Tanzania, there were columns which were removed from the dataset.

ID column was removed because it had no any significant impact to the target column. Also, Latitude and Longitude columns were removed since they were for determining the location of the respondent which had no significant impact.

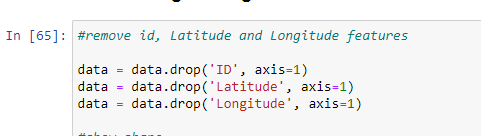


Figure showing dropping of unnecessary columns

### **Dataset split**

After dropping the irrelevant columns, the dataset was split into features and target. The target column being mobile\_money\_classification column, which classifies an individual into any one of four mutually exclusive groups,

* whether an individual use mobile money only,
* no mobile money but uses other financial services,
* uses mobile money and at least one other financial service, or
* no mobile money and no any other financial service.

The features in the dataset were 33 columns after removing the target column.

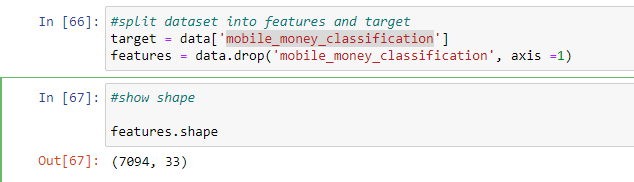


Figure splitting data

### **Label encoding of binary categorical features**

Some of the binary categorical features had options encoded 1 and 2 instead of 0 and 1 as intended to be used by the model. So, the values 2 were changed to 0 as per label encoding rules.

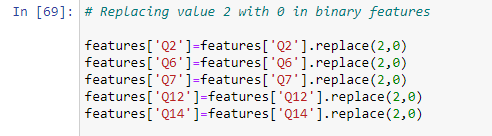


Figure replacing 2 with 0

### **One hot encoding**

Other features had more than two options, so to change them into an appropriate format for the model, a one-hot-encoding preprocessing method was employed as shown in the figure below;

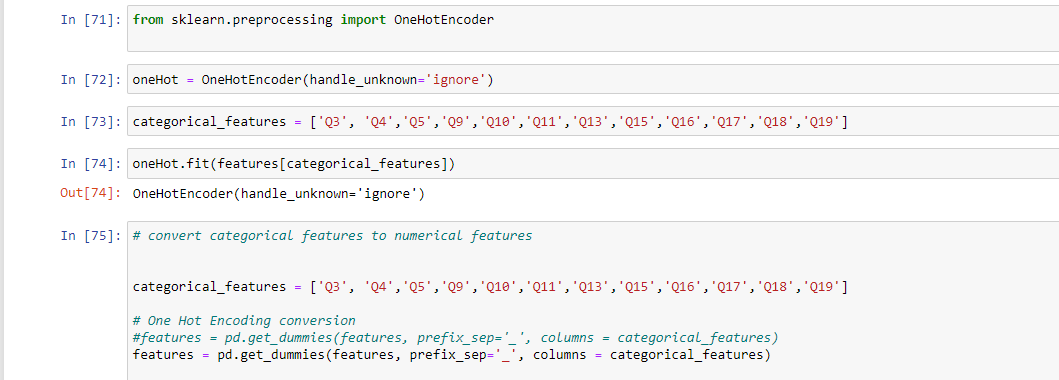


Figure one hot encoding

### **Continuous features; min-max normalization**

Continuous features such as age and prices have different range of value hence undergo a standardization process so that the model may perform well.

To do so, Min-Max Scaler method / Standardization method is used.

In our case, Min-Max scaler was employed for column ‘Q1’ (Age), as shown in the figure below;

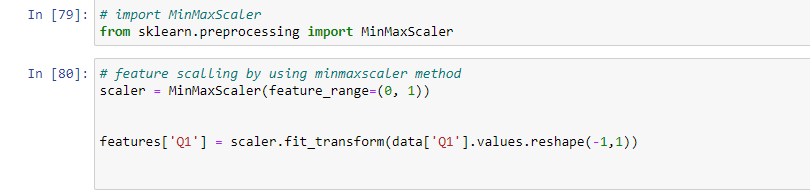


Figure min max scaler

### **Feature selection**

Feature Selection is the process where you automatically or manually select those features which have the most contribution to the prediction variable or output you are interested in.

Having irrelevant features in your data can *decrease* the accuracy of the models and make your model learn based on irrelevant features.

Top reasons to use feature selection are;

1. It enables the machine learning algorithm to train faster.
2. It reduces the complexity of a model and makes it easier to interpret.
3. It improves the accuracy of a model if the right subset is chosen.
4. It reduces overfitting.

In selecting the best features, SelectKBest class was applied to extract top 40 best features from the list of all features by using chi square score, as shown in the diagram below;



Figure best feature selection

## **MODEL SELECTION AND EVALUATION**

**Model selection** is the process of selecting final machine learning model from among a collection of candidate machine learning models for a training dataset.

For example, we have a dataset for which we are interested in developing a classification model of mobile money and financial inclusion challenge in Tanzania. We do not know beforehand as to which model will perform best on this problem, as it is unknowable. Therefore, we fit and evaluate a suite of different models on the problem.

**What do we care about when choosing a final model?**

* A model that meets the requirements and constraints of project stakeholders.
* A model that is sufficiently skillful given the time and resources available.
* A model that is skillful relative to other tested models

By using the **Voting Classifier** technique, where by instead of creating dedicated separate models and finding the accuracy for each of them, we created a single model which trains by these models and predicts output based on their combined majority of voting for each output class.



Figure model selection

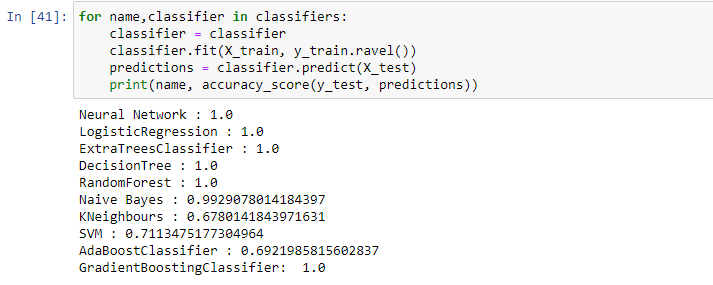


Figure classifier's accuracy

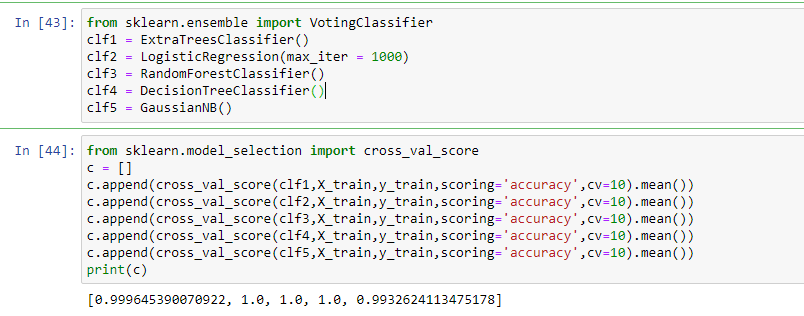


Figure voting classifier

So, the best models were those with higher accuracy scores like Logistic Regression and Random Forest.

We used RandomForest classifier.

**Model evaluation** is the process of using different evaluation metrics to understand a machine learning model’s performance, as well as its strengths and weaknesses.

Model evaluation is important to assess the efficiency of a model during initial phases, and it also plays a role in model monitoring. To understand if your model(s) is working well with new data, you can leverage a number of evaluation metrics.

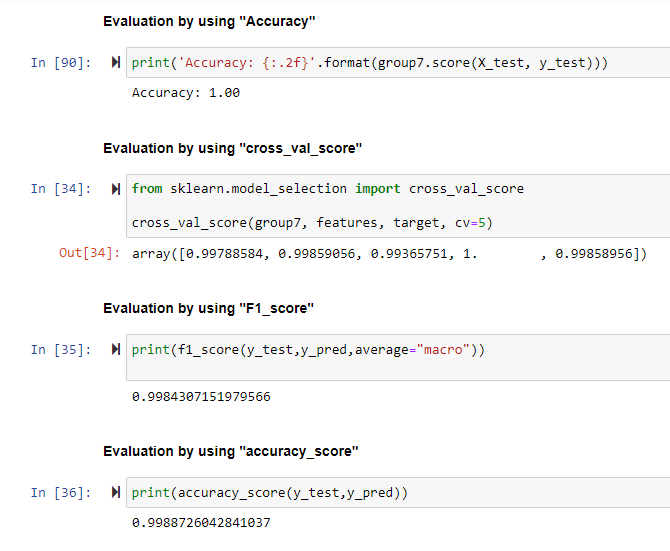


Figure model evaluation

**Accuracy metric;**

It measures how often the classifier makes the correct predictions, as it is the ratio between the number of correct predictions and the total number of predictions.

**Confusion matrix;**

The **confusion matrix** (or confusion table) shows a more detailed breakdown of correct and incorrect classifications for each class. Using a confusion matrix is useful when you want to understand the distinction between classes, particularly when the cost of misclassification might differ for the two classes, or you have a lot more test data on one class than the other.

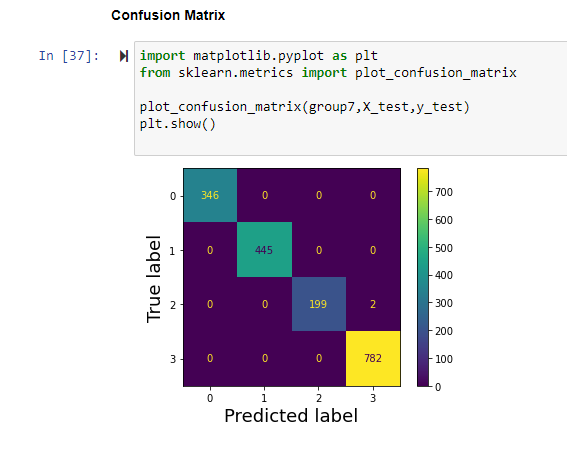


Figure confusion matrix

**MODEL SELECTION AND EVALUATION TABLE**

|  |  |  |  |
| --- | --- | --- | --- |
| No | Features | Algorithm | Perfomance(Accuracy) |
| 1. | 33 | Neural Network | 1.0 |
| 2. | 33 | Logistic Regression | 1.0 |
| 3 | 33 | Extra Trees Classifier | 1.0 |
| 4 | 33 | Decision Tree | 1.0 |
| 5 | 33 | Random Forest | 1.0 |
| 6 | 33 | Naive Bayes | 0.99 |
| 7 | 33 | K Neighbours | 0.678 |
| 8 | 33 | SVM | 0.71 |
| 9 | 33 | Ada Boost Classifier | 0.69 |
| 10 | 33 | Gradient Boosting Classifier | 1.0 |
| 11 | 33 | Light GBM | 1.0 |
| 12 | 33 | XG Boost | 1.0 |
| 13 | 33 | Cat Boost | 1.0 |
| 14 | 33 | Histogram-based gradient boosting classification tree | 1.0 |

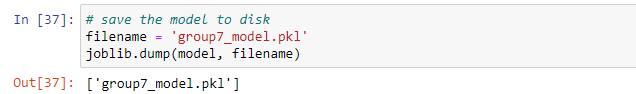
# **MODEL DEPLOYMENT**

Model deployment is the method where you integrate a machine learning  model into an existing production environment to make practical business decisions based on data.

In order to get the most out of machine learning models, it is important to seamlessly deploy them into production so a business can start using them to make practical decisions.

Once we were done training, we saved the model by using the joblib python package as shown in the figure below;

Figure saving model by joblib



After that, we deployed the models in the form of web applications to increase their accessibility.

There are several options of deploying a project on the web like; Flash, Django and Streamlit, but we preferred streamlit due to the following reasons;

* Streamlit lets you create apps for your machine learning project using simple code
* No need to write a backend and no need to define different routes or handle HTTP requests like in Django and Flask.

So, we created a streamlit file then loaded the built model for web application by using the joblib package on the python.

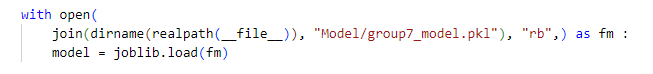


Figure loading model

Then, we performed prediction by using predict function and displayed the results;

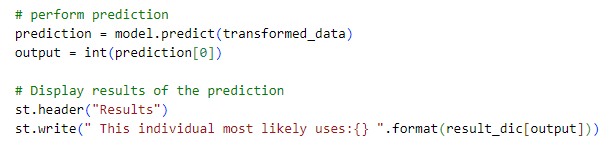


Figure prediction and results

## **CREATING FRONTEND**

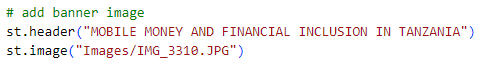
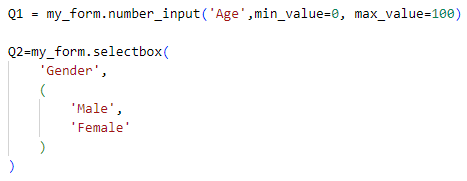
Here is where the user interface creation on the web is done. We used the streamlit scripts form and its elements like selectbox, number\_input, submit and others like setting image scripts for creating the interface on the web application as shown on figures below;

Figure number\_input and selectbox

Figure adding banner image



## **Running Web App**

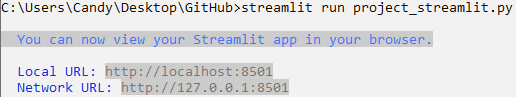
After that we ran the web app on our browser by running the command:-

Figure run web app

Then, the above Local URL was copied then pasted on browser.

Finally, we viewed our streamlit app which appeared just as shown in figure below;

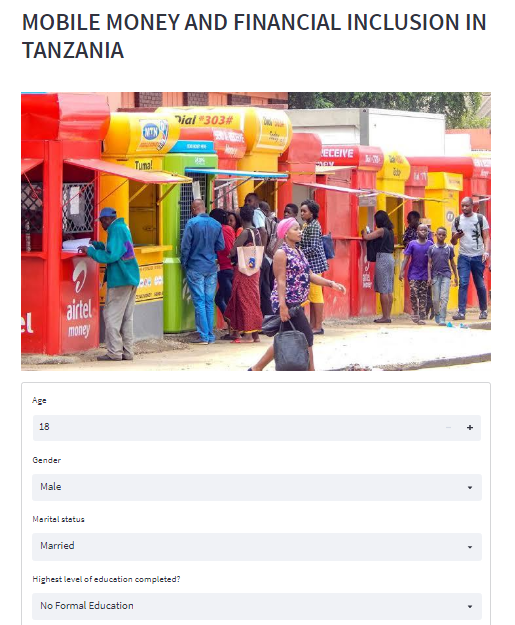


Figure streamlit app

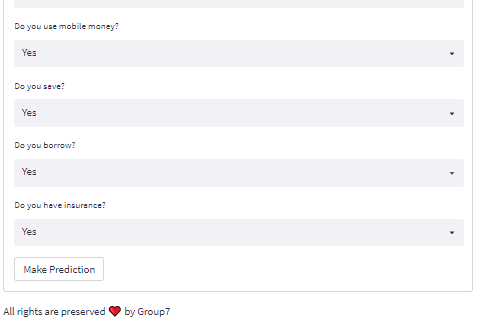


Figure streamlit app cont..

The app can then be used by mobile money and financial service providers in Tanzania.

## **Pushing deployment folder on github**

We created an account on github for pushing the deployment folder(local repo) to the Github(remote repo).

Github link: <https://github.com/CandyTheophil/Mobile-money-and-financial-inclusion-Tanzania>

# **CONCLUSION**

This solution will help mobile money providers target new clients and markets across Tanzania more effectively hence increase **profits** among their business organizations and also provide **better financial convenience** for the said markets and clients.

But, it is important to identify the various challenges which were encountered while doing this project such as:-

* Problem in understanding the dataset
* Presence of difficult features such as Latitude and Longitude and,
* A big number of categories in categorical features

Overall, all limitations end on a positive note seeing as how we worked together to provide a solution that indicates Tanzania’s hopes of achieving the goal of building a modern ICT-based economy are ripe.

# **REFERENCES**

Ephraim, I, Mhina, D, (n.d). FSDT CASE STUDY The story of digital finance in Tanzania. Retrieved from <https://www.fsdt.or.tz/wp-content/uploads/2021/05/FSDT_DFS_CaseStudy.pdf>

Zindi, (2019). Mobile Money and Financial Inclusion Challenge. Retrieved from <https://zindi.africa/competitions/mobile-money-and-financial-inclusion-in-tanzania-challenge/data>