REPORT ON THE MACHINE LEARNING PROJECT TO PREDICT IF A PERSON IN CERTAIN AFRICAN COUNTRIES HAS A BANK ACCOUNT OR NOT

I started the project by carefully reading through the project presentation slide that was sent to the group chat. I also went to zindi to read about the project too.

I then developed a project plan. I’ll explain what I did for the project using the project plan that I created.

PROJECT PLAN:

1. PROJECT DEFINITON

At this step, I looked at the project in detail and explained to myself what this project is all about.

1. PROJECT PREPROCESSING/PREPARATION
2. I loaded the data my jupyter notebook
3. Then I checked out the all the columns of the dataset to see if they have missing values or other inconsistencies. But there was no dirt in the data. The dataset was very clean.
4. I checked for duplicates in the dataset, but there was none.
5. Then I removed the ID column of the dataset because, it is of no value to our project.
6. EXPLORATORY DATA ANALYSIS

At this step I analyzed the dataset in 3 ways:

1. Distribution Analysis using histograms and bar charts: During this analysis I noticed that the data was imbalanced. Most of the respondents were people without bank account. I also listed out lots of other things that I noticed
2. Variable Relationship Analysis using heatmaps: Beacause my dataset was mostly categorical, I could not calculate the correlation between the variables straight on. I had to change the two numerical columns to categorical, then use a contingency table and cramer’s v to measure the correlation between the variables before visualizing it in a heat map. Cramer’s V to put it simple is a statistical measure of how strong the correlation between two categorical variables are.(It measures this by checking the correlation between each two pairs of categories for each variable and then using the result to calculate a single number that represents the correlation between the two datasets.) It ranges form 0-1 instead of -1 to 1. From this heatmap I noticed that the correlation between variables of the dataset was weak and i also noticed a few other things that I listed in the notebook.
3. I analyzed the relationship with the target variable using grouped bar charts: I also listed the things i noticed in the notebook.
4. MODEL DEVELOPMENT/ EVALUATION

After fitting the dataset to the two models I choose, I finally selected the Random Forest classifier because it has the highest precision for the Yes category of the target variable. It also has a high accuracy for the No category. It has a high accuracy score of 0.8842 e.t.c.

1. PROJECT DEPLOYMENT

For project deployment I programmed a small tkinter window that contains drop down options and a predict button. The user can just select from the options below and click the predict button to see if a person has a bank account or not. (NOTE: The cell for this code is stand alone. I mean that you can just run the cell without running any other cell in the notebook and it will work)