# Alvin Smart Money Management Classification

**Prepared by Endework Abera**

## Objective

The objective of this project is to develop a machine learning algorithm that classifies purchases into 13 distinct categories. This project was undertaken as part of a challenge hosted by Zindi. The datasets for this project were obtained from here.

## Data Description

The data for this challenge was collected over an 11-month period during Alvin's Beta release. If a user is the first to make a purchase at a merchant, the app prompts the user to manually classify the merchant. The next user to purchase at that merchant can either confirm the suggestion or enter a new categorization.

Alvin currently registers user transaction data via MPESA SMS receipts, but some users classify their purchases manually. The training set contains approximately 400 purchases, and the test set contains about 600 purchases. These ~1000 transactions have been verified by Alvin as correctly classified. There are ~10,000 unverified transactions in the unverified file; these are purchases users have classified themselves, and Alvin has not checked.

The datasets include the following variables:

- MERCHANT_CATEGORIZED_AT: The time the merchant was categorized by the customer.
- MERCHANT_NAME: The name of the merchant.
- MERCHANT_CATEGORIZED_AS: The category the merchant was assigned by the customer.
- PURCHASE_VALUE: The value of the purchase made by the customer.
- PURCHASED_AT: The time the purchase was made.
- IS_PURCHASE_PAID_VIA_MPESA_SEND_MONEY: If true, indicates that the merchant is not a registered business name.
- USER_EMAIL: The email of the customer.
- USER_AGE: The age of the customer.
- USER_GENDER: The gender of the customer.
- USER_HOUSEHOLD: The number of family members.
- USER_INCOME: The monthly income of the customer.

## Data Preprocessing

After loading the datasets using pandas, missing values were handled. For instance, the USER_AGE and USER_GENDER fields showed missing values in both the test and train datasets. For the

USER_GENDER, the most frequent gender value was calculated and used to fill the missing values. In this case, it was ‘female’. For the USER_AGE, the skewness of the data was examined, and the median was used to fill the null values.

Next, categorical data was converted to numerical as the algorithm accepts numeric values. In our dataset, the USER_GENDER and IS_PURCHASE_PAID_VIA_MPESA_SEND_MONEY fields are categorical, so they were converted to numerical using one-hot encoding.

# Exploratory Data Analysis (EDA)

An analysis was conducted to understand the distribution of different classes in the target variable. It appears that the classes in the target variable are imbalanced. For instance, ‘Bills & Fees’ is the most common type of transaction, while ‘Education’ is the least common. Categories like ‘Groceries’ and ‘Data & WiFi’ have similar counts, suggesting moderate frequency.

Given this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was used to generate synthetic data for underrepresented classes. Balancing the classes can help improve the performance of the model, especially for those classes that have fewer instances, as it provides a more even training signal across all classes.

# Feature Engineering

Feature importance was used to determine which features are best for the model. Based on this, PURCHASE_VALUE, USER_INCOME, USER_HOUSEHOLD, USER_GENDER, and the converted value of MERCHANT_NAMES were selected.

# Model Building and Evaluation

A split of 0.8 for the training set and 0.2 for the test set was utilized. Various models were evaluated, including logistic regression, decision tree, XGBoost, and random forest, with accuracies of 17%, 81%, 78%, and 83% respectively. Among all the models, logistic regression had the lowest performance with an accuracy of 17%, while the decision tree achieved an accuracy of 78%. The random forest classifier achieved an accuracy of 83%.

# Model Tuning

At this stage, the random forest classifier was selected for further evaluation. Its performance was assessed by tuning some parameters using GridSearchCV.

Attempts were made to improve the performance of the higher-performing models through model imputation, but this resulted in a decrease in accuracy.

# Conclusion and Future Work

This project demonstrates the application of various machine learning models to classify purchases into different categories. The models achieved a maximum accuracy of 83%. However, there is still room for improvement.

Future work could involve exploring more sophisticated techniques for handling class imbalance, conducting more extensive feature engineering, or applying more advanced models. Additionally, more rigorous hyper parameter tuning could potentially improve the model’s performance. Overall, this project serves as a valuable step towards developing more accurate and robust classification models for purchase data.