# Assignment 1
# ECO481
# University of Toronto

Instructor: Marlène Koffi

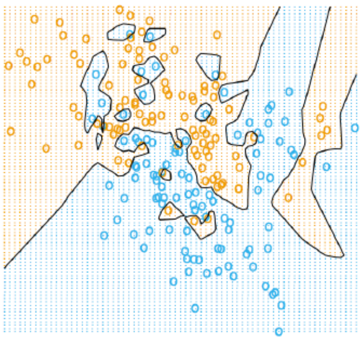
Fall 2022

## 1 Exercise 1 (25%)

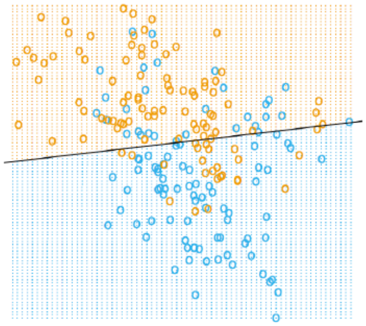
1. Your friend Albert builds a classification algorithm on the following data. It has 10,000 features and 100 observations. As the newest machine learning expert in your group of friends, he decided to ask you for help. (10 points)

   - Clearly explain to Albert what his model is likely to suffer from.
   - In fact, Albert has implemented the model and finds an accuracy rate of 99% on the training sample. Unfortunately, when it evaluates the model on a new dataset (the test sample), you have 50% accuracy. Explain to Albert what it means to have 50% accuracy.
   - Suggest a step you would take to fix the problem Albert is having.
   - After correcting the problem, you get two models. First, you use logistic regression and get an error rate of 10% on training data and 15% on test data. Then we use the 1-nearest neighbors (i.e. K=1) and get an average error rate (simple average over test and training datasets) of 9%. Based on these results, what method would you recommend to Albert for his classification exercise? Clearly explain why.

2. For each decision boundary, explain which classifier between the logistic regression and KNN is likely to have generated it. (5 points)



**Panel A**



**Panel B**

3. After training a logistic regression classifier, Anne and Bryan have one of the data point that is properly classifier and far away from the decision boundary. Bryan thinks that removing this point will not affect the decision boundary. Anne disagrees and think that it may affect the decision boundary. Who is right? Explain. (5 points)

4. You have a set of three data points with only one predictor and an output: $x_1 = 4$ and $y_1 = 1$, $x_2 = -2$ and $y_2 = 0$, $x_3 = 1$ and $y_3 = 1$. Suppose you use it as a training sample. In a logistic regression, what will be the value of the parameter $\beta$ associated with $x$? Give a short explanation for your choice. (5 points)

   - a) $\beta = 1$
   - b) $\beta = 0$
   - c) $\beta = \infty$

## 2 Exercise 2: (10%)

You work as a data analyst at Twitter. Because of all the scandals and growing fake news on social media, Twitter would like to improve its algorithm to detect fake news. As the junior analyst, you are in charge of analyzing the past “fake news” data.

You are given the following information:

- 10 out of 10000 tweets are classified as fake by the current twitter algorithm.
- With some human checks, we realize that 20% of the tweets classified as fake by the algorithm are not.
- 10% of tweets classified as ”non fake” by the algorithm are.

1. Your boss asks you to build the confusion matrix for 1,000,000 tweets (roughly the number of tweets per 5 minutes in 2021).(5 points)

2. You have a talk in front of your team. How would you summarize in a very meaningful way how well the current algorithm is performing? (show your calculation and explain your decision). (5 points)

## 3 Exercise 3: Application... (65%)

"Mobile money (m-money) refers to the use of mobile phones to perform financial and banking functions." (IFC Mobile Money Study 2011: Summary Report).

In low-income countries, mobile money is a substitute for banking access. In fact, individuals do not need a bank account to perform financial transactions (send and receive money) via their mobile service. One of its biggest advantages is that it can reach the most remote and vulnerable populations. Many observers agree that this new financial tool has an important role in widening financial inclusion in low-income countries (See Jack and Suri 2011 and Suri 2017 for a review).

Therefore, it is crucial to understand the mobile money users and non-users characteristics. In this application, we want to analyze the determinants of mobile money adoption using (real) data from a survey of 2,282 households in Kenya on M-PESA ("M" for Mobile and "Pesa" for Money in Swahili), one of the most successful mobile money applications.

1. First, you have access to the following table:

| Personal ID | Large household size | Have a cell phone | Have a mattress at home | M-pesa user |
|---|---|---|---|---|
| 1 | False | True | False | True |
| 2 | False | False | False | False |
| 3 | False | True | False | True |
| 4 | True | False | False | False |
| 5 | True | False | True | True |
| 6 | True | False | False | False |

- Using the table and an entropy-based information gain, construct a decision tree (by hand, i.e make the calculus and find the relevant splits) that would predict the use of M-pesa for an individual. (NB: the logarithm to use in the entropy measurement is the logarithm to the base 2.) (15 points)
- What will be the prediction generated by the tree for: "Large Household"= false, "Have a cell phone"= False and "Have a mattress at home"=true.(1.5 points)
- What will be the prediction generated by the tree for: "Large Household"= True, "Have a cell phone"= True and "Have a mattress at home"=true.(1.5 points)

2. Now, you have access to a more complete database. Read the file "mobile_money.csv" in Python. (2 points)

3. Present descriptive statistics on the outcome variable mpesa_user. Comment.(3 points)

4. Present descriptive statistics on the following variables depending on the mpesa_user status. (10 points: 5 points for each label)

   - Own Cell Phone
   - Per Capita Consumption
   - Per Capita Food Consumption
   - Total Wealth
   - Household Size
   - Education of Head (Years)
   - Positive Shock
   - Negative Shock
   - Weather/Agricultural shock
   - Illness Shock
   - Send Remittances
   - Receive Remittances
   - Bank account
   - Mattress
   - Savings & Credit Cooperative (SACCO)
   - Merry Go Round/ ROSCA
   - Farmer
   - Public Service
   - Professional Occupation
   - Househelp
   - Run a Business
   - Sales
   - In Industry
   - Other Occupation

- Unemployed

5. Comment on the descriptive statistics' main takeaways in question 3 (no more than 3 lines).(2 points)
6. Construct the following classifiers using the outcome variable mpesa_user (11 points):

   - Logistic Classifier
   - Decision Tree Classifier
   - Random Forest classifier

   NB: Consider a train-test split of 80-20. Consider also standardizing the data before.
7. Comparing the accuracy rate and the area under the curve (AUC) criteria, find the best classifier among those in question 5.(6 points: 2-2-2)
8. What are the top 3 predictors based on the best classifier found in question 6? (3 points)
9. Consider now a KNN classifier. Using a loop "for", consider a value of K from 1 to 10 by step of 1. In the ML jargon, we are doing a grid search. It aims to tune (find) the value of the hyperparameter "K"). Using cross-validation methods on the training data set, for each value of K, find the optimal value of neighbours K. (5 points)
10. Is the optimal KNN classifier, as found in question 8, outperforming the one found in 6?(2 points)
11. Based on what you have found, what is the key recommendation that you can make to a government that would like to foster the use of M-PESA among the population? (no more than 3 lines).(3 points)

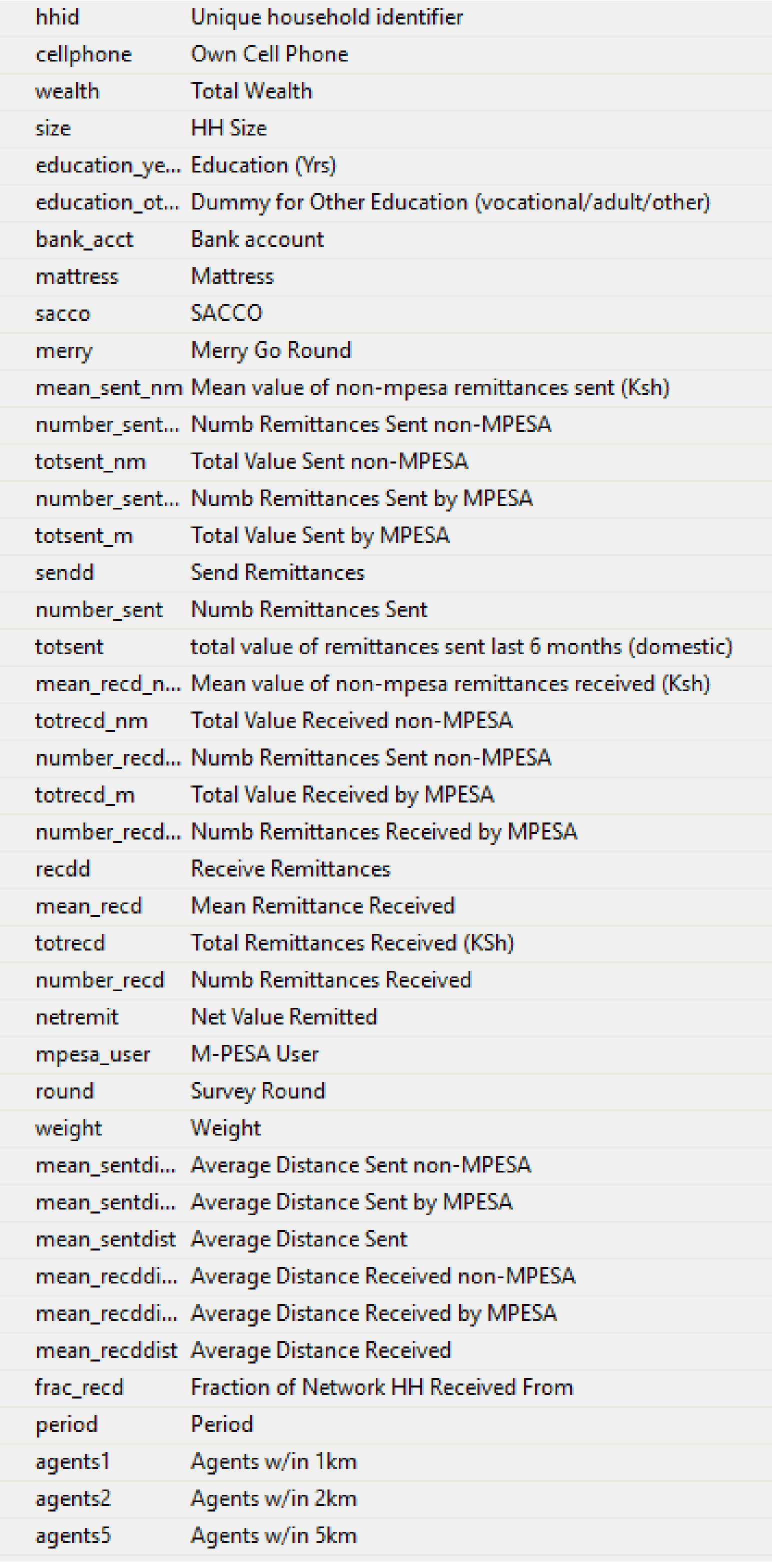
# References:

Jack, W., & Suri, T. (2011). Mobile money: The economics of M-PESA (No. w16721). National Bureau of Economic Research.

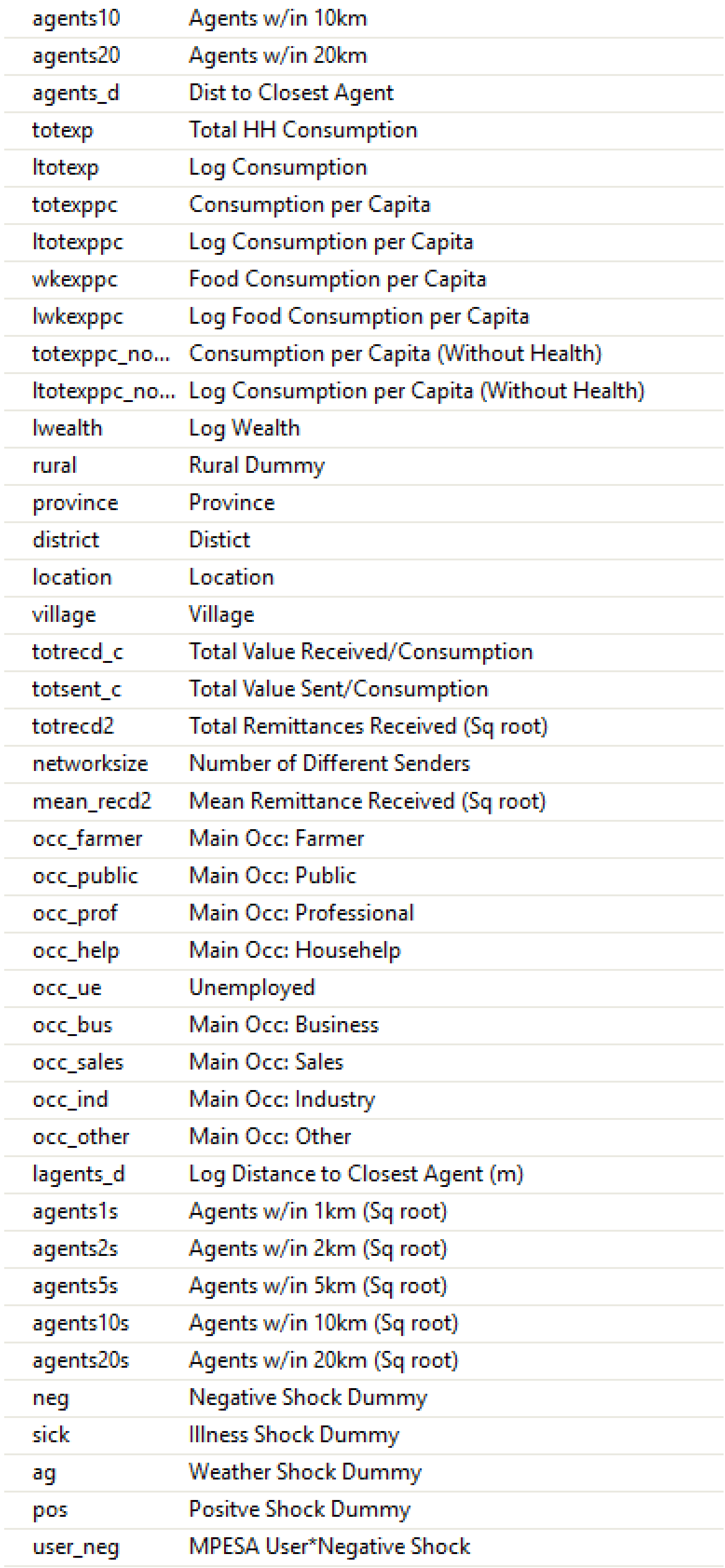
Jack, W., & Suri, T. (2014). Risk sharing and transactions costs: Evidence from Kenya's mobile money revolution. American Economic Review, 104(1), 183-223.

Suri, T. (2017). Mobile money. Annual Review of Economics, 9, 497-520.

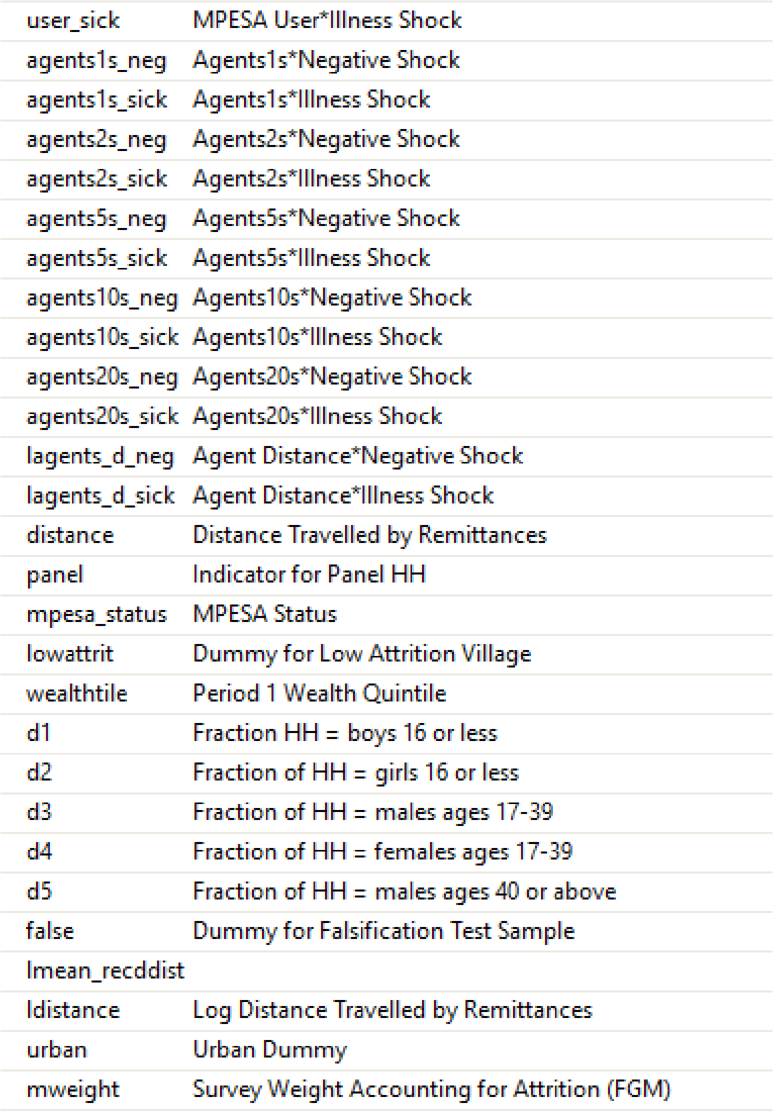
# Variables Labels



| | |
|---|---|
| hhid | Unique household identifier |
| cellphone | Own Cell Phone |
| wealth | Total Wealth |
| size | HH Size |
| education_ye... | Education (Yrs) |
| education_ot... | Dummy for Other Education (vocational/adult/other) |
| bank_acct | Bank account |
| mattress | Mattress |
| sacco | SACCO |
| merry | Merry Go Round |
| mean_sent_nm | Mean value of non-mpesa remittances sent (Ksh) |
| number_sent... | Numb Remittances Sent non-MPESA |
| totsent_nm | Total Value Sent non-MPESA |
| number_sent... | Numb Remittances Sent by MPESA |
| totsent_m | Total Value Sent by MPESA |
| sendd | Send Remittances |
| number_sent | Numb Remittances Sent |
| totsent | total value of remittances sent last 6 months (domestic) |
| mean_recd_n... | Mean value of non-mpesa remittances received (Ksh) |
| totrecd_nm | Total Value Received non-MPESA |
| number_recd... | Numb Remittances Sent non-MPESA |
| totrecd_m | Total Value Received by MPESA |
| number_recd... | Numb Remittances Received by MPESA |
| recdd | Receive Remittances |
| mean_recd | Mean Remittance Received |
| totrecd | Total Remittances Received (KSh) |
| number_recd | Numb Remittances Received |
| netremit | Net Value Remitted |
| mpesa_user | M-PESA User |
| round | Survey Round |
| weight | Weight |
| mean_sentdi... | Average Distance Sent non-MPESA |
| mean_sentdi... | Average Distance Sent by MPESA |
| mean_sentdist | Average Distance Sent |
| mean_recddi... | Average Distance Received non-MPESA |
| mean_recddi... | Average Distance Received by MPESA |
| mean_recddist | Average Distance Received |
| frac_recd | Fraction of Network HH Received From |
| period | Period |
| agents1 | Agents w/in 1km |
| agents2 | Agents w/in 2km |
| agents5 | Agents w/in 5km |



| | |
|---|---|
| agents10 | Agents w/in 10km |
| agents20 | Agents w/in 20km |
| agents_d | Dist to Closest Agent |
| totexp | Total HH Consumption |
| ltotexp | Log Consumption |
| totexppc | Consumption per Capita |
| ltotexppc | Log Consumption per Capita |
| wkexppc | Food Consumption per Capita |
| lwkexppc | Log Food Consumption per Capita |
| totexppc_no... | Consumption per Capita (Without Health) |
| ltotexppc_no... | Log Consumption per Capita (Without Health) |
| lwealth | Log Wealth |
| rural | Rural Dummy |
| province | Province |
| district | Distict |
| location | Location |
| village | Village |
| totrecd_c | Total Value Received/Consumption |
| totsent_c | Total Value Sent/Consumption |
| totrecd2 | Total Remittances Received (Sq root) |
| networksize | Number of Different Senders |
| mean_recd2 | Mean Remittance Received (Sq root) |
| occ_farmer | Main Occ: Farmer |
| occ_public | Main Occ: Public |
| occ_prof | Main Occ: Professional |
| occ_help | Main Occ: Househelp |
| occ_ue | Unemployed |
| occ_bus | Main Occ: Business |
| occ_sales | Main Occ: Sales |
| occ_ind | Main Occ: Industry |
| occ_other | Main Occ: Other |
| lagents_d | Log Distance to Closest Agent (m) |
| agents1s | Agents w/in 1km (Sq root) |
| agents2s | Agents w/in 2km (Sq root) |
| agents5s | Agents w/in 5km (Sq root) |
| agents10s | Agents w/in 10km (Sq root) |
| agents20s | Agents w/in 20km (Sq root) |
| neg | Negative Shock Dummy |
| sick | Illness Shock Dummy |
| ag | Weather Shock Dummy |
| pos | Positve Shock Dummy |
| user_neg | MPESA User*Negative Shock |



| | |
|---|---|
| user_sick | MPESA User*Illness Shock |
| agents1s_neg | Agents1s*Negative Shock |
| agents1s_sick | Agents1s*Illness Shock |
| agents2s_neg | Agents2s*Negative Shock |
| agents2s_sick | Agents2s*Illness Shock |
| agents5s_neg | Agents5s*Negative Shock |
| agents5s_sick | Agents5s*Illness Shock |
| agents10s_neg | Agents10s*Negative Shock |
| agents10s_sick | Agents10s*Illness Shock |
| agents20s_neg | Agents20s*Negative Shock |
| agents20s_sick | Agents20s*Illness Shock |
| lagents_d_neg | Agent Distance*Negative Shock |
| lagents_d_sick | Agent Distance*Illness Shock |
| distance | Distance Travelled by Remittances |
| panel | Indicator for Panel HH |
| mpesa_status | MPESA Status |
| lowattrit | Dummy for Low Attrition Village |
| wealthtile | Period 1 Wealth Quintile |
| d1 | Fraction HH = boys 16 or less |
| d2 | Fraction of HH = girls 16 or less |
| d3 | Fraction of HH = males ages 17-39 |
| d4 | Fraction of HH = females ages 17-39 |
| d5 | Fraction of HH = males ages 40 or above |
| false | Dummy for Falsification Test Sample |
| lmean_recddist | |
| ldistance | Log Distance Travelled by Remittances |
| urban | Urban Dummy |
| mweight | Survey Weight Accounting for Attrition (FGM) |