CIND119 Project Milestone

# Members

Aida Judaki, [aida.judaki@torontomu.ca](mailto:aida.judaki@torontomu.ca)

Karem Allen, [karem.allen@torontomu.ca](mailto:karem.allen@torontomu.ca)   
Stephanie Boissonneault, [sboissonneault@torontomu.ca](mailto:sboissonneault@torontomu.ca)

**Dataset:** In this section, you need to mention which dataset you have selected for the project. And a brief explanation of how many rows, and features this dataset has:   
  
Our team has selected the Bank Marketing Dataset. This dataset has 4522 rows (including the header row) and 17 features of which 9 are qualitative data (job, marital, education, default, housing, loan, contact, month, Poutcome), 7 are numeric data (age, balance, day, duration, campaign, Pdays, previous) in addition to the Y class attribute which is a binary (yes or no) attribute showing whether the client has subscribed a term deposit or not.

| **Feature** | **Description** | **Data Type** |
| --- | --- | --- |
| Age | Age of the customer | numeric (discrete) |
| Job | Type of job | qualitative (nominal) |
| Marital | Marital status | qualitative (nominal) |
| Education | Education of the customer | qualitative (ordinal) |
| Default | Shows whether the customer has credit in default or not | qualitative (nominal, binary “yes” or “no”) |
| Balance | Average yearly balance in Euros | numeric (continuous) |
| Housing | Shows whether the customer has a housing loan or not | qualitative (nominal, binary “yes” or “no”) |
| Loan | Shows whether the customer has a personal loan or not | qualitative (nominal, binary “yes” or “no”) |
| Contact | Shows how the last contact for the marketing campaign has been made | qualitative (nominal) |
| Day | Shows on which day of the month the last time the customer was contacted | numeric (discrete) |
| Month | Shows on which month of the year last time the customer was contacted | qualitative (ordinal) |
| Duration | Shows the last contact duration in seconds | numeric (continuous) |
| Campaign | Number of contacts performed during the marketing campaign and for this  customer | numeric (discrete) |
| Pdays | Number of days that passed by after the client was last contacted by a  previous campaign | numeric (-1 means the client was not previously contacted) (discrete) |
| Previous | Number of contacts performed before this campaign and for this client | numeric (discrete) |
| Poutcome | The outcome of the previous marketing campaign | qualitative (nominal) |
| Y | Class attribute showing whether the client has subscribed to a term deposit or not | qualitative (nominal, binary (“yes”, “no”) |

**Project Approach:** In this section, you need to mention bullet points of your methodology.For example, in the approach, you can mention the steps of data preparation you plan to take, which tools you will use for modeling, and data split strategy (cross-validation or train-test set split).

For this project, our team must help a Portuguese bank develop an effective telemarketing strategy to sell long-term deposit accounts. We will do so by applying data analytic methods to help them determine which customers are most likely to subscribe to a long-term deposit account.

**1. Data preparation:**

**Use SAS to complete data preparation including**

* Assess and understand the data attributes and their type (nominal, ordinal, or quantitative)
  + Check whether there are missing or duplicate values, special characters, or errors in the data that need to be addressed during data cleaning
  + Handle missing values or transform values as needed
* Conduct exploratory analysis of data to investigate and understand what features are important during predictive modeling:
  + Conduct summary statistics: find max, min, mean, and standard deviation
  + Creating a boxplot of each attribute to help visualize data and note outliers for numeric values such as age, duration, pdays, and balance.
  + Plot a histogram of numeric attributes including age, duration, pdays, and balance against the class attribute (y) to analyze and identify relationships or certain patterns
  + Identify which attributes seem to be correlated
  + Based on the findings, will determine which attributes are of concern and which can be eliminated
* Investigate the proportion of records to determine whether the dataset has an imbalanced class distribution that needs to be addressed.

**2. Predictive Modeling (parts a and b)**

**Applying classification algorithms, including Decision Tree and Naive Bayes using python\_weka.ipynb file to help determine which customers are most likely to subscribe to a long-term deposit account.**

* Apply train-test split
  + To analyze the effectiveness of supervised learning and the generalization capability of our model
    - The data will be split into two non-overlapping subsets including a training set (80%) and a testing set(20%)
    - Our model will only use the training set to build the models
* Decision Tree:
  + Will use Python to apply a Classification Decision Tree classification algorithm
    - Create a baseline model using all attributes of the training set
    - Set a maximum depth to the model to stop splitting at 100.
    - Based on the algorithm output, the root node will be determined as the best predictor/classifier
    - Assess attributes used by the decision tree
* Naive Bayes:
  + Will use Python to apply the Naïve Bayes classification algorithm to predict the binary class attribute (Y).
  + Create a baseline model using all attributes of the training set
  + Features that were noted as important for consideration during Data preparation will be kept in mind.

**2. Predictive Modeling (parts c) and 3. Conclusion and Recommendations**

* Applying Evaluation metrics:
  + Once both models have been generated, we will use our test set (20% data unseen by the model) to generate a confusion matrix and evaluate our model’s performance based on true positives, false positives, precision, recall, and accuracy (to determine the proportion of correct predictions)
* Lastly, we will interpret the results and summarise our findings in the Project Submission Template Slides provided
  + Create a table to compare model accuracy, Precision, Recall, and F1 score
* Make recommendations based on data interpretation

# Workload Distribution

In this section, you need to mention who will complete what part of the project.

| Member Name | List of Tasks |
| --- | --- |
| Aida Judaki | 1. Data Preparation. |
| Karem Allen | 2. Predictive Modeling/Classification:  a. Classification using a Decision Tree.  b. Classification using Naive Bayes |
| Stephanie Boissonneault | c. Compare the results of the 2 techniques on original and filtered data (includes applying evaluation metrics)  3. Conclusions and Recommendations |