**Group 2 Project 4**

Our team comprises Manal Bayoumi, Rocio Cantu, Johnny Troung, and Jennifer Melendez. Our project shows the data that banks collect, how they view people's loans and whether they default or do not default on the loans. Our project will discuss the factors that make people default on loans. Also, it will look at the customer’s standard of living, whether they are male or female, income level, education level, and amount of loan they default on.

**There are several questions to ask using that dataset:**

· Why do people default on their loans?

· Which profession defaults the most on their loans?

· Do educated or uneducated default the most on their loans?

· Which income level defaults the most on their loans?

We will add more questions as we proceed with our project. Our dataset consists of about 25 columns and 30,000 rows. Our data is large enough to analyze and provide information about the loans and how people fall into having loans and defaulting on them. We will use Python or SQL Database to upload our data, that might change according to our data and how the project will proceed. We will work hard to meet on zoom and during class hours. Also, we will create visualizations for different aspects of the project, such as whether loans default or not default.

Concerning the rough breakdown of tasks, tasks will be divided equally within that project. Manal will be responsible for cleaning, training, and testing the data. Rocio and Jennifer will prepare the presentation, and Johnny will write the analysis. Jennifer will do a tableau story and different dashboards for our group. We will all input our thoughts in the analysis. **Our group tasks are subject to change according to the project's needs.** We hope that our project will get sufficient funding from the shareholders of our company.

   Within that project, we will use supervised Machine Learning Process that has the following features: the size of the loan, sex, the income level of the borrower, the number of opened accounts, marital status, education level, and the total debt that the borrower has. Also, it has a label showing whether the loan defaults or not. Using that data, we can do a first model and oversampling data. We will have two models to calculate precision, accuracy, and recall level. We are hoping that will get higher accuracy, precision and recall.

For the models, we will use the following:

* Prepared Data: Import data from csv file, change the data type, and normalize the data, resampling the data if required(most critical part).
* Separate the data into features and labels and use train\_test\_split to separate them into training and testing data sets.
* Initialize the model.
* Fitting the model using the training data(X\_train, y\_train).
* Use this model to make predictions(use X\_test to predict y\_predictions).
* Evaluate the predictions(compare y\_predictions and y\_test).

**Link to our data:**

<https://www.kaggle.com/datasets/samanemami/credit-approval-loan?select=Loan_Defaulter.csv>