Milestone 2 – Data Selection & Project Proposal

Name: Abed Tabbalat

Team Status: Individual

- **What types of model or models do you plan to use and why?**

The plan is to run 3 different models and see which ones fits best for the dataset. The models I am choosing to run and find out which is the best model are:

1. Random Forest
2. Logistic Regression
3. Decision Tree

The reason I chose the models above is because the dataset shows variables for different people with an outcome variable of ones and zeros that determines if the person is diabetic or not. The additional variables that come with it are characteristics of the person that could be triggers on being diabetic. Once the model's accuracy is tested, a confusion matrix will be tested on the models for a better visual.

- **How do you plan to evaluate your results?**

Results will be evaluated by choosing the most accurate model and applying a confusion matrix to the model and then an ROC curve to determine how fit the model will be. Accuracy percentage is the key to have a successful model.

- **What do you hope to learn?**

I recently found out that a close family member has gotten diabetes. I have never paid attention to it as it is a highly common condition people go through. Once I learned that this could be a genetic passthrough, my curiosity has been risen to learn more about it. What environmental factors could cause it? What variables from the data collected highly impacts the outcome of diabetes.

In addition, learning which model best fits certain types of datasets based on what variables it contains is one thing I would want to be excellent with. The ability to look at a dataset and narrowing down which models to use. Of course, this comes with experience and repetition, and this is one of the technical goals that I have.

- **Assess any risks with your proposal.**

The main risk I currently see is the dataset is less than 1,000 rows which can impact the model results. The more data and information we have the better the outcome of the model will be. This risk may not exist depending on how the variables are correlated with each other. In addition, the models I have chosen could result in inaccurate results if overfitting happens which could result in choosing a different approach in predictions.

- **Identify a contingency plan if your original project plan does not work out.**

If the plan does not work out because the analysis shows that there isn't enough data to support the model results, then the plan will be to find a larger dataset within the same topic. That said, having the same topic and possible similar variables should not impact the choice of the models that will be run. If the results are yet to be inaccurate due to the data not having enough features to predict. Finding a different dataset at that time will be a must and starting over. Even though that will be a stretch, a backup plan must be in place.

- **Include anything else you believe is important.**

The main question regarding diabetes would be, what variables are mostly impacting the outcome of having a positive diabetes diagnosis. To add, this topic or healthcare in general has never been my cup of tea, as my interest has always been in the P&C insurance industry. Sometimes life gets in the way, and we forget the importance of health that keeps us alive and well to be able to continue doing what we do. I take that for granted and there is nothing better than learning about it through datasets and predictions. Choosing diabetes in specific, as mentioned above, having a close family member diagnosed with it, while never paid attention to it except for "can't have sugar" is not enough anymore. This will help me understand finding out what correlates with the outcome what I should monitor for myself and others to avoid getting it. I believe as common as it is, anyone would be interested in knowing what variables do correlate to be able to monitor themselves and live a better life.