Week 1

1. I have uploaded the data into the Python Shell.
2. I then had a look at the first 4 columns.
3. I realized that I had zeros that did not mean anything to the data columns, “Glucose”, “Insulin”, “Skin Thickness”, “Blood Pressure” and “BMI”, so I then replaced all those zeros with missing values “Nan”, then replaced the “Nan” values with the median of each column respectively.
4. Finally, I create a count (frequency) plot describing the data types and the count of variables.

Week 2

1. I checked the balance of the data by plotting the count of outcomes by their value and I realized that 34.9% of the people were diabetic. My plan is to find out which factors contribute to people being diabetic and to build a model that will better predict the likelihood of someone being Diabetic.
2. I created scatter charts and realized that Age and Pregnancy are strongly positively correlated. I also observed that Diabetes Pedigree Function and blood pressure are strongly negatively correlated.
3. I then performed correlation analysis and I visually explored it using a heat map.

Week 3

1. I decided to use two models, namely a Logistic Regression model, as it will help me predicts a dependent data variable by analyzing the relationship between one or more existing independent variables, and a Decision Tree model so I can predict a set of binary rules to predict a target value.
2. I have noticed that the Logistic regression models gives me a better prediction accuracy of 77%.

Week 4

**Precision**

Precision is the ability of a classifier not to label an instance positive that is actually negative.

The model’s prediction is accuracy 78% for “None Diabetic” and 69% for “Diabetic”.

**Recall**

Recall is the ability of a classifier to find all positive instances.

The model correctly classified “None Diabetic” at 86% accuracy, and “Diabetic” at 55%.

**F1 Score**

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

The precision of the model is very strong as both values are closer to 1.0

**Conclusion**

The model is very precise and strong. I have also noticed that there is a huge correlation between Glucose and High blood pressure in ages 20-25 and 25-30, which may be a result of unhealthy lifestyle habits which may result in them being Diabetic. They may try and change the food they consume, reduce alcohol consumption or join a gym.