**Diabetes Miletus Prediction using various Machine Learning Models**

By T. Sachin Soundar

1. **Introduction**
   1. **Overview**

Pima Indian Diabetes dataset is to predict the onset of diabetes based on diagnostic measures. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict whether a patient is suffering from diabetes or not based on diagnostic measurements.

* 1. **Purpose**

The diabetes dataset is a binary classification problem where it needs to be analysed whether a patient is suffering from the disease or not on the basis of many available features in the dataset. Different methods and procedures of cleaning the data, feature extraction, feature engineering and algorithms to predict the onset of diabetes are used based for diagnostic measure on Pima Indians Diabetes Dataset. Multiple Prediction models are built on the same Dataset and the accuracy of prediction of the various model involved is compared adjacently to detect and employ the better model in terms of sheer recall and precision for better prediction of diabetes.

1. **Literature Survey**

**•** In 2017, National Diabetes Statistic Report for Centre Disease Control and Prevention (CDC), gives the facts give an account of the United States that 30.3 million individuals have diabetes, among that 23.1 are analysed and 7.2 million are undiscovered individuals.

• In 2018, the American Diabetes Association models of therapeutic care in diabetes discharges a report about “Order and finding of diabetes” which incorporates the arrangement of diabetes, diabetes care, treatment objectives, criteria for conclusion test ranges and dangers esteems, chance engaged with diabetes.

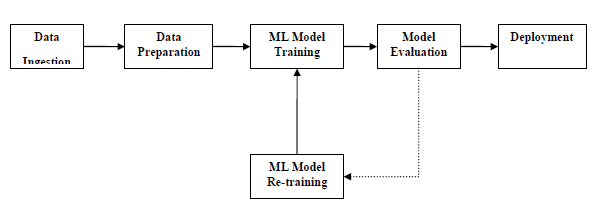
• In 2017, Global provides details regarding Diabetes by world wellbeing association, it expresses the weight of diabetes, hazard components and inconveniences of diabetes. Likewise, gives the data about counteracting diabetes in individuals with high hazard and overseeing diabetes at beginning times with fundamental solutions to be taken.

1. **Theoretical Analysis**
   1. **Analysing the Dataset**

# Pima Indian Diabetes dataset has 9 attributes in total. All the person in records are females and the **number of pregnancies** they have had has been recorded as the first attribute of the dataset. Second is the value of **Plasma glucose concentration** a 2 hours in an oral glucose tolerance test and then is the **Diastolic blood pressure** (mm Hg), fourth in line is the **Triceps skin fold thickness** (mm), then is the **2-Hour serum insulin** (mu U/ml), sixth is **Body mass index** (weight in kg/ (height in m) ^2) and then seventh is the **Diabetes pedigree function** and the second last value is the that of the **Age** (years). The ninth column is that of the Class variable (**Outcome - 0 or 1**), 0 for no diabetes and 1 for the presence.

# To start with we first take a description of the dataset. We infer not much from this except the facts like we have a data **dataset of 768 lines** and the maximum values of the Age and Pregnancies. Nothing more is of much use for the prediction. We also calculated the number of datasets that were positive to the test of diabetes and those who were negative and the value came out to be **268 and 500** respectively.

* 1. **Block Diagram**



The blocks of ML pipelines are as follows

**Data ingestion** − As the name suggests, it is the process of importing the data for use in ML project. The data can be extracted in real time or batches from single or multiple systems. It is one of the most challenging steps because the quality of data can affect the whole ML model.

**Data Preparation** − After importing the data, we need to prepare data to be used for our ML model. Data pre-processing is one of the most important technique of data preparation.

**ML Model Training** − Next step is to train our ML model. We have various ML algorithms like supervised, unsupervised, reinforcement to extract the features from data, and make predictions.

**Model Evaluation** − Next, we need to evaluate the ML model. In case of AutoML pipeline, ML model can be evaluated with the help of various statistical methods and business rules.

**ML Model retraining** − In case of AutoML pipeline, it is not necessary that the first model is best one. The first model is considered as a baseline model and we can train it repeatably to increase model’s accuracy.

**Deployment** − At last, we need to deploy the model. This step involves applying and migrating the model to business operations for their use.

1. **Experimental Investigations**
   1. **Data Analysis**

# We decided to take the mean value of BMI and found that the average value of a person suffering from the disease has mean BMI value as 35.14 which means that they are not healthy and obese. It is also interesting to note that the mean BMI value for the people who are not suffering from the disease is 30 which is the threshold value of people becoming obese.

The mean value of the second parameter Glucose (Plasma glucose concentration) was done we found that those who suffered from the disease had mean value as 141.25 which indicates pre-diabetic state of hyperglycaemia that is associated with insulin resistance and increased risk of cardiovascular pathology.

* 1. **Statistical Analysis of Dataset**

# **DataFrame.describe()** method generates descriptive statistics that summarize the central tendency, dispersion and shape of a dataset’s distribution, excluding NaN values. This method tells us a lot of things about a dataset. One important thing is that the describe() method deals only with numeric values. It doesn't work with any categorical values. So if there are any categorical values in a column the describe() method will ignore it and display summary for the other columns unless parameter include="all" is passed.

# Now, let's understand the statistics that are generated by the describe() method:

# **count** tells us the number of NoN-empty rows in a feature.

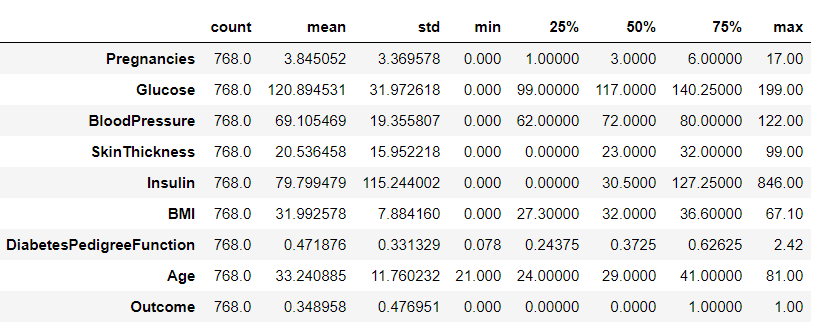
# **mean** tells us the mean value of that feature.

# b tells us the Standard Deviation Value of that feature.

# b tells us the minimum value of that feature.

# **25%, 50%, and 75%** are the percentile/quartile of each features. This quartile information helps us to detect Outliers.

# **max** tells us the maximum value of that feature.



# On these columns, a value of zero does not make sense and thus indicates missing value. Following columns or variables have an invalid zero value:

# Glucose

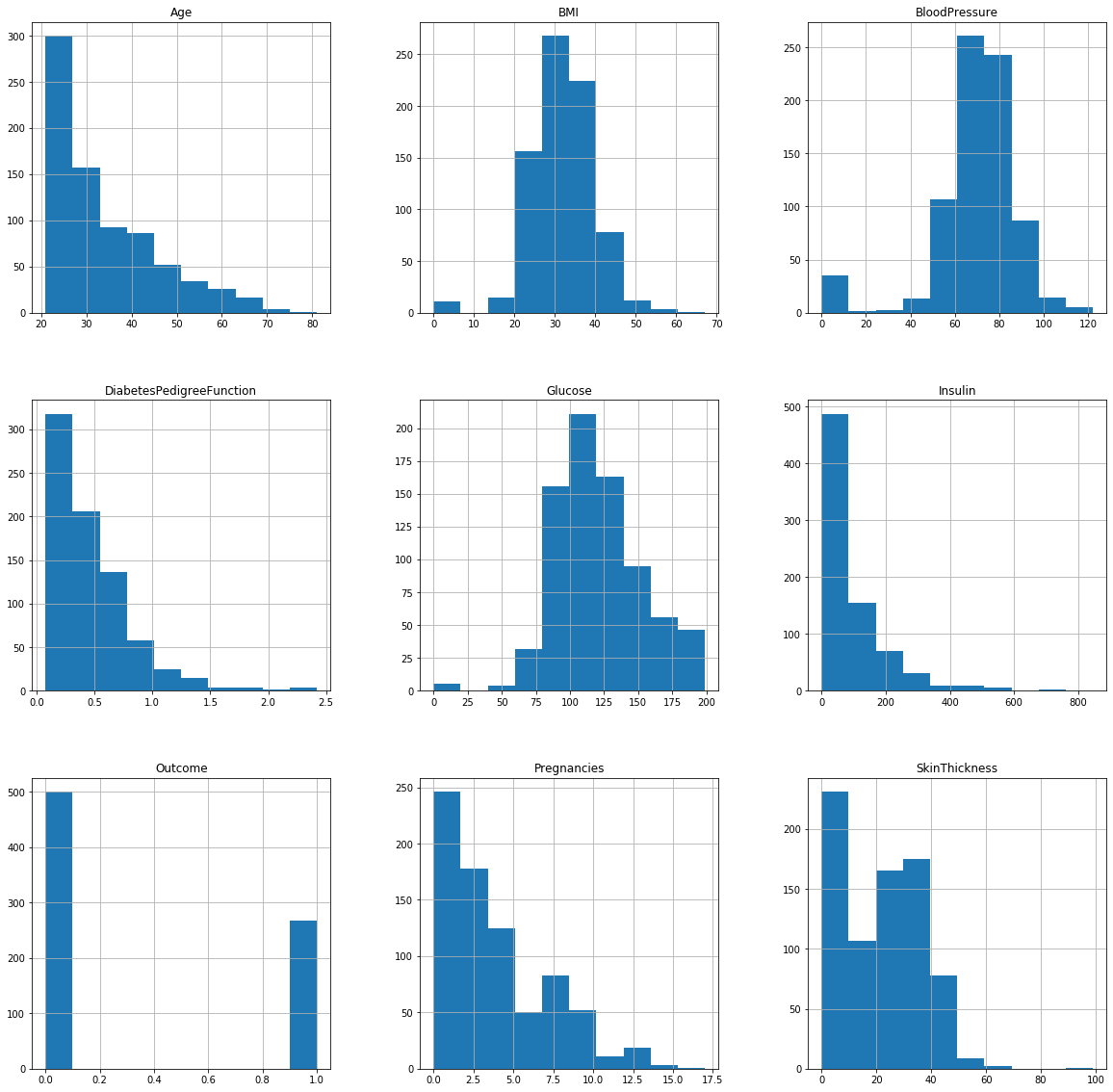
# BloodPressure

# SkinThickness

# Insulin

# BMI

#### To fill these NaN values the data distribution needs to be understood. The following graphs indicate the spread of the data values in the dataset pertaining to the range of the respective parameters.

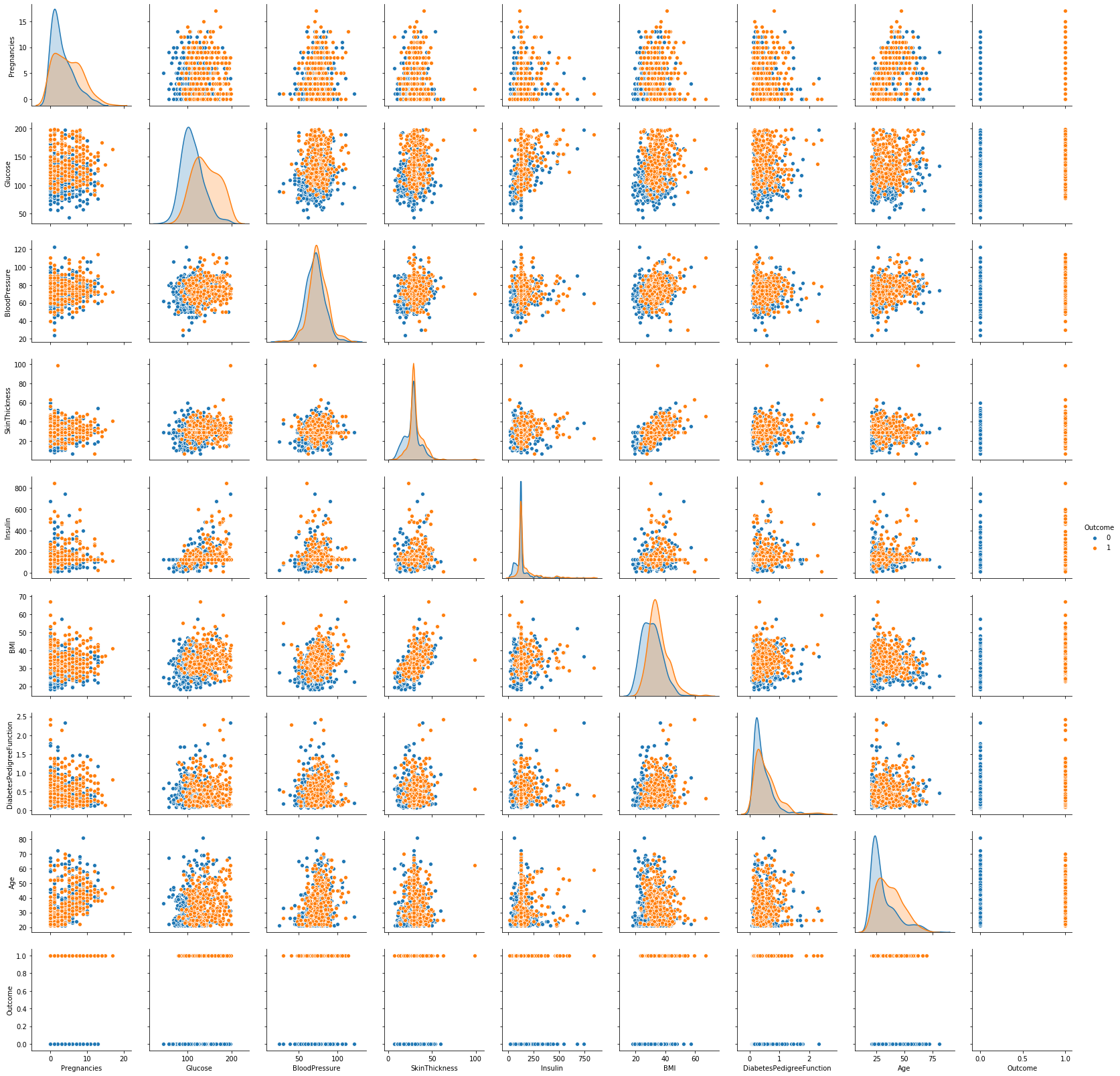


From the above depiction, the dataset would be fairly accuracy and meaningful if the missing data values are being substituted by the median of the particular variables, as it wouldn’t affect the overall spread of the histograms being presented above.

* 1. **Exploratory Data Analysis**

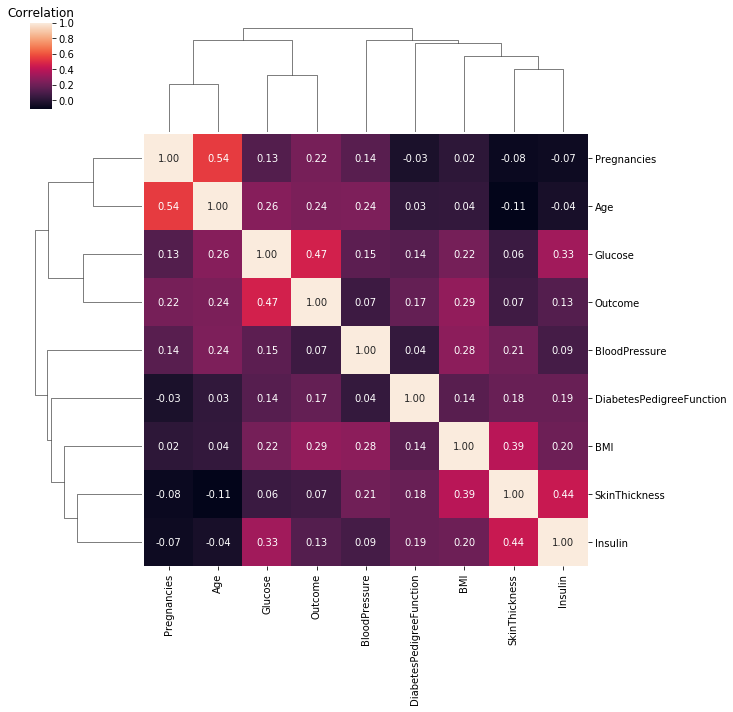
# To understand the data well we can do two things either we can just simply look at the data and check the values of each feature or we can do exploratory data analysis. If we choose the first option we would become bias to the data and our mind inclines to the best know algorithm for the data but that would be good only for training. Hence we perform EDA to ensure the data stay unbiased and relevant.

# The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables.



One thing that we were able to deduce from this image was that all the parameters overlap for the Outcome value, i.e., no matter if you are suffering from the disease of not, you can have the same parameters.

# **Pearson's Correlation Coefficient**: helps you find out the relationship between two quantities. It gives you the measure of the strength of association between two variables. The value of Pearson's Correlation Coefficient can be between -1 to +1. 1 means that they are highly correlated and 0 means no correlation. A heat map is a two-dimensional representation of information with the help of colors. Heat maps can help the user visualize simple or complex information.



This heat map was of much use as we got to know that the following pairs had a positive correlation coefficient between them as compared to the other parameters:

1) Pregnancies and Age

2) Insulin and Skin thickness

3) BMI and Skin thickness

4) Insulin and Glucose

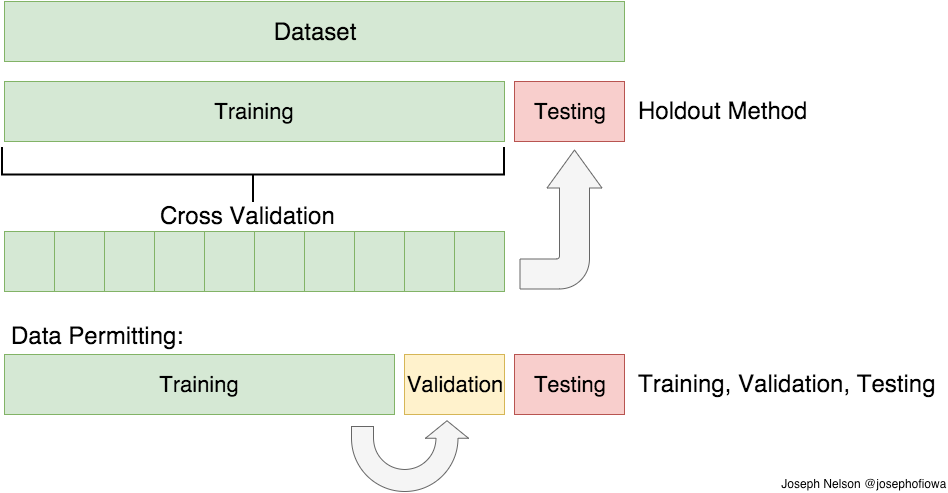
And with the Outcome value, Glucose and BMI values related the most.

This helped us to know that Glucose and BMI are the parameters we need to take special care of.

* 1. **Train Test Split and Cross Validation**

***Train Test Split*** : To have unknown datapoints to test the data rather than testing with the same points with which the model was trained. This helps capture the model performance much better.

***Cross Validation***: When model is split into training and testing it can be possible that specific type of data point may go entirely into either training or testing portion. This would lead the model to perform poorly. Hence over-fitting and underfitting problems can be well avoided with cross validation techniques.



Before going for any algorithm, we divided our dataset into two parts training and testing in the 70:30 ratios.

* 1. **Decision Trees**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

Decision tree algorithm follows:

• The attribute/feature best for set is taken as root

• Distribute the set into different sets having same attribute values for particular value.

• Repeat the above steps till we get to the leaf nodes of the tree where no further division can take place.

* 1. **Random Forest**

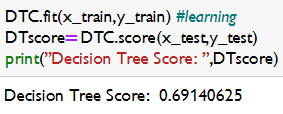
### Random Forest is the one of the Classifier which is used for Classifications problems. Random Forest is ensemble classifier made using many decision trees where ensemble means that uses multiple machine learning algorithm to obtain the predictive performance It is better than other for the prediction of diabetes mellitus.

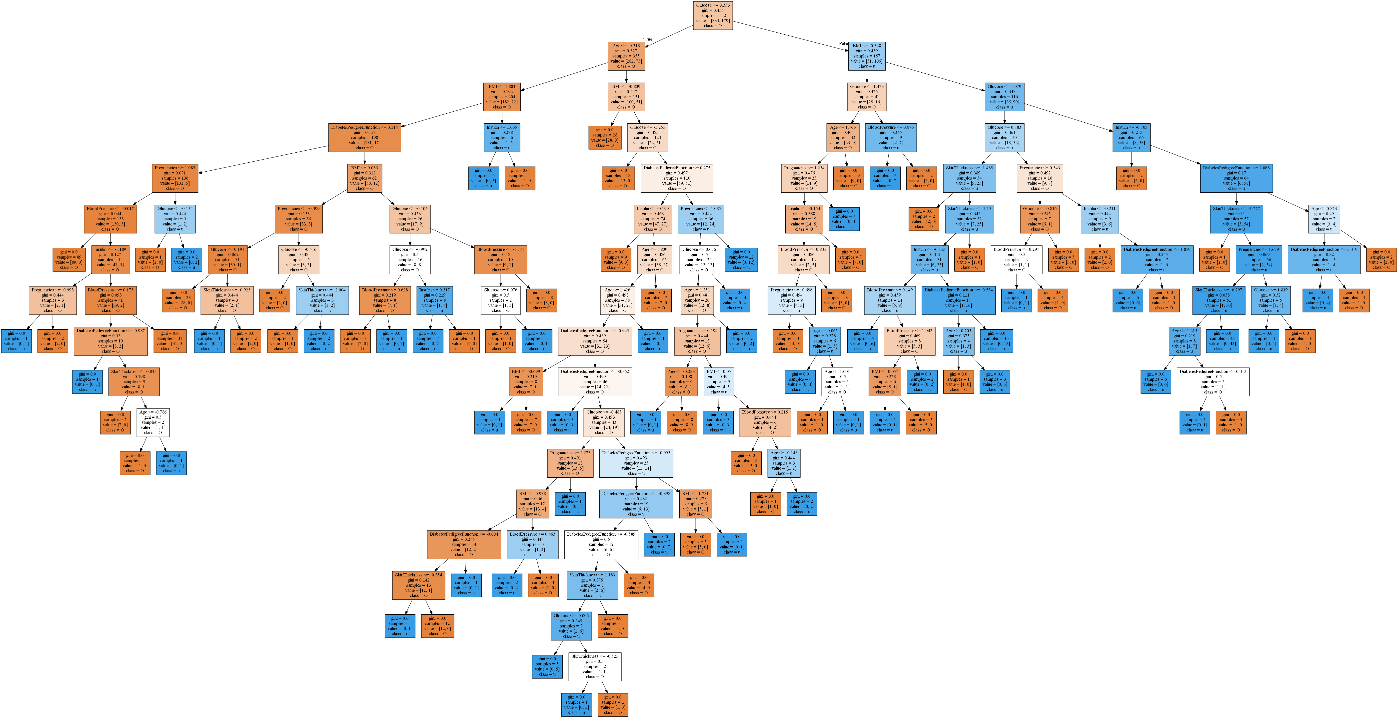
* 1. **Artificial Neural Network**

A Sequential Model is a class of feedforward artificial neural network. This Sequential Model consists of seven layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. Sequential Model utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish Sequential Model from a linear perceptron. It can distinguish data that is not linearly separable.

1. **Results**
   1. **Decision Trees**

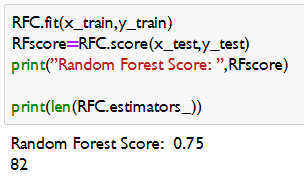
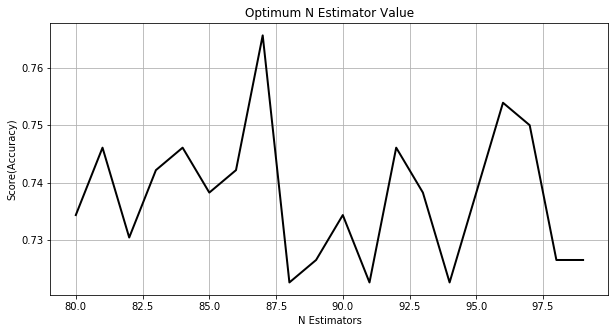
To have a confirmation on the best features we found the feature importance using the Decision Tree. We also performed the prediction using the same algorithm which gave us an accuracy on training and testing data as 74.0 and 69.0 percentages respectively.



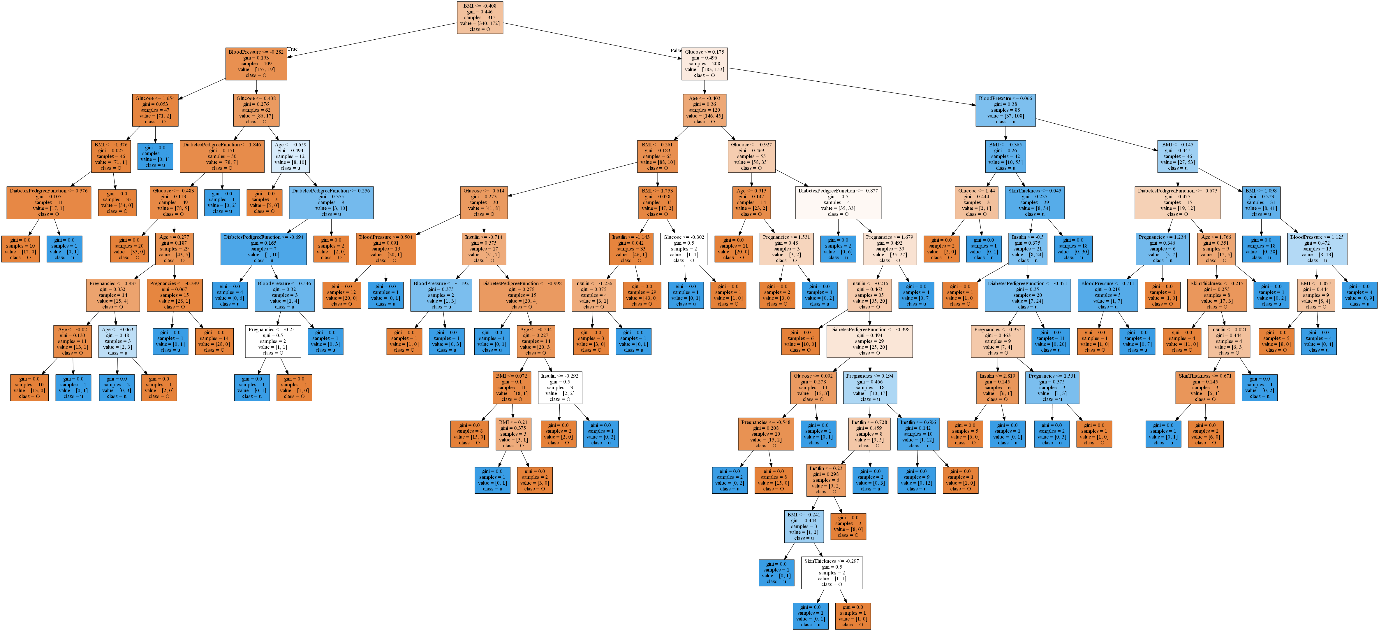
****

* 1. **Random Forest**

The score of random forest which performed bagging for around 82 estimated tress and found the best result of recall and precision with accuracy about 0.75.

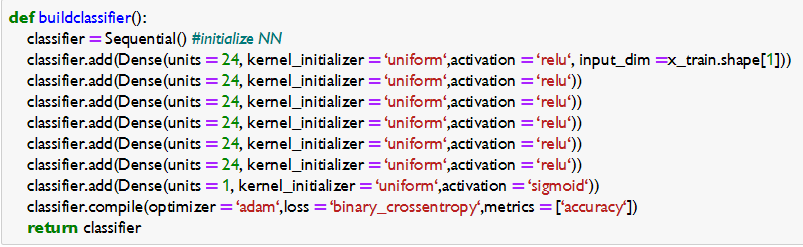
 ****

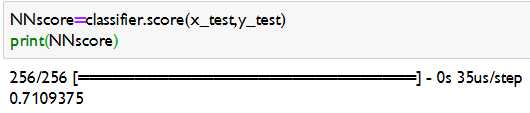
This Tree representation is just one of the possible tree which would be taken into account by the Random Forest Classifier for improving accuracy.



* 1. **Artificial Neural Networks**

The Neural Network is a Sequential Neural model built using 7 dense layers.





1. **Comparison of Performance**
   1. **Confusion Matrix Report**

The confusion matrix is a technique used for summarizing the performance of a classification algorithm i.e. it has binary outputs.

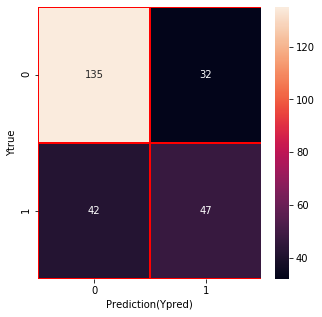
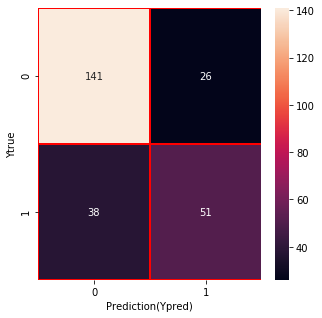
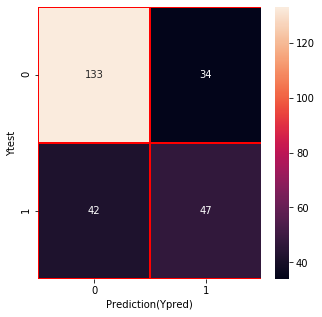
**Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

*Precision = TP/TP+FP*

**Recall (Sensitivity)** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? A recall greater than 0.5 is good.

*Recall = TP/TP+FN*

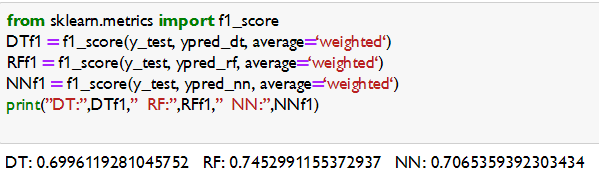
**Decision Tree Random Forest Neural Network**

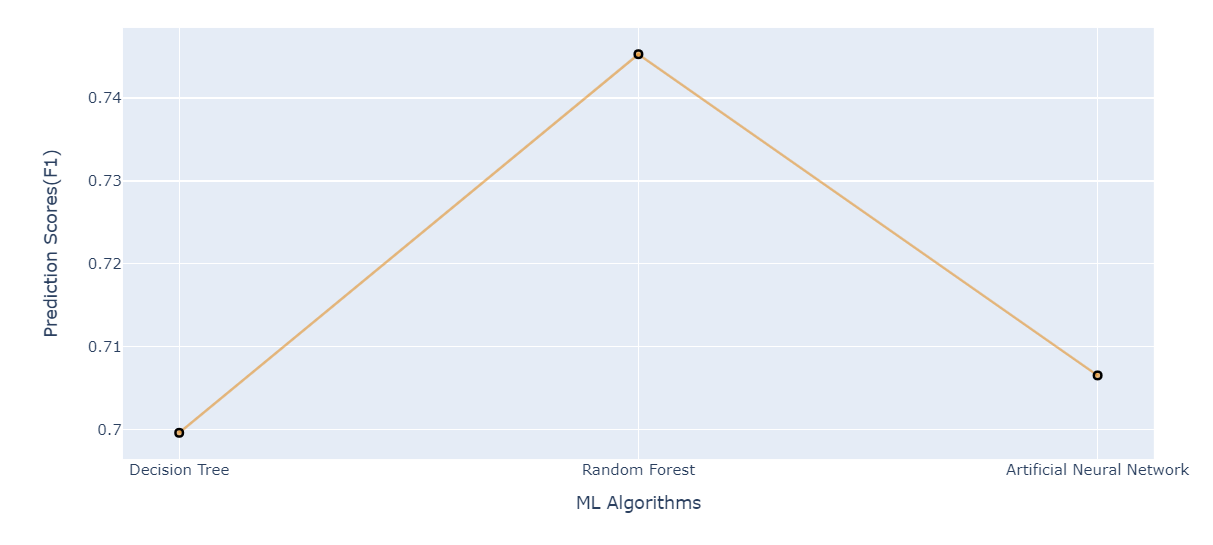
****

* 1. **F1 Scores**

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

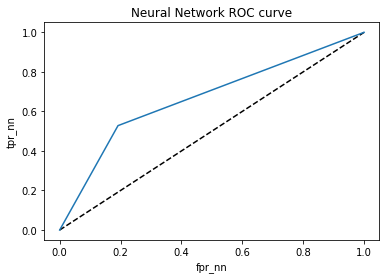
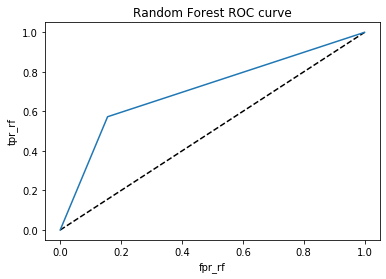
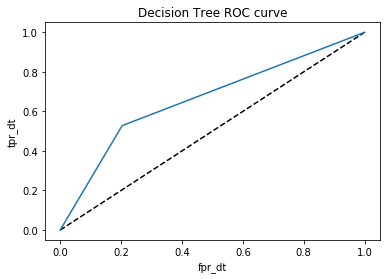
*F1 Score = 2(Recall Precision) / (Recall + Precision)*

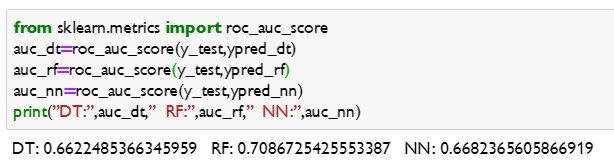




* 1. **ROC – AUC**

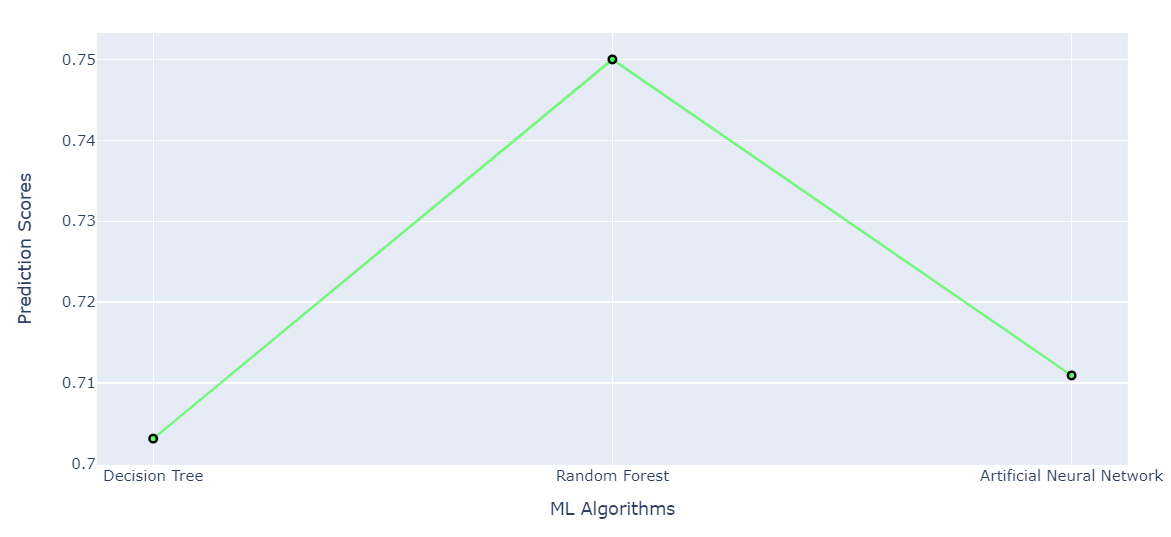
ROC (Receiver Operating Characteristic) Curve tells us about how good the model can distinguish between two things (e.g If a patient has a disease or no). Better models can accurately distinguish between the two. Whereas, a poor model will have difficulties in distinguishing between the two.

****



The Area under Curve of random forest is around 0.7 which makes Random forest better at predicting the diabetes outcome as it has the best distinguishing potential out of the three chosen models.

1. **Conclusion**



By comparing the accuracy of the three models, we can clearly say that random forest has given better outcome and a greater prediction potential compared to the other models. This fact is backed-up by the higher f1 score of random forest model and a greater ROC – AUC score. The Neural Network which is a much more powerful model, needs to be trained and tuned much more accurately to provide better accuracy scores. A different Sequential net could have provided a better score than the current NN Score, Nevertheless, Random Forest outperformed the other model in these situations.

1. **Bibliography**

<https://keras.io/guides/sequential_model/#when-to-use-a-sequential-model>

<https://www.kaggle.com/shrutimechlearn/step-by-step-diabetes-classification-knn-detailed>

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>

<https://medium.com/@djocz/confusion-matrix-aint-that-confusing-d29e18403327>