CI512 – Intelligent Systems

Pima Indians Diabetes Database

Vickshan Vicknakumaran

## **Introduction**

The Pima Indians Diabetes Database is based on the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of the dataset is to diagnostically predict if a patient has diabetes or not. Due to the number of patients with diabetes, the data were restricted to only include females over 21 years old who had Pima Indians Heritage. The datasets are divided into three categories: prediction, outcome, and BMI. The dataset includes the number of pregnancies and the patient's insulin level. Also included are the number of times a person has given birth and the height of their body mass index. They also include the plasma glucose concentration and the Triceps skin fold thickness. Using machine learning, I was able to predict whether a patient had diabetes.

## **Algorithms**

### K-nearest neighbours (KNN)

The supervised learning technique K-nearest neighbours (KNN) is used for both regression and classification. By calculating the distance between the test data and all the training points, KNN attempts to predict the correct class for the test data. Then chooses the K number of points that are closest to the test data. The reason for my use of the KNN algorithm as it can provide highly accurate predictions which can determine if a patient has diabetes or not.

### Decision Tree Classifier

Through a Decision Tree, the model can learn and predict the class or value of a target variable by learning the simple decision rules given from the data. The decision tree starts from the root and compares the values to the attributes in the tree. The classifier is used to predict the response for given data. I had used this algorithm as it is easy to interpret and understand.

### Random Forest Classifier

Random forest is a machine learning algorithm that learns and builds decision trees on different samples and takes their majority vote for classification problems. The reason behind using this classifier as it compatible with big dataset. Furthermore, while more trees go through the algorithm, the random forest adds more randomness to the model. When splitting a node, it looks for the best feature from a random subset of features rather than the most essential feature. As a result, there is a lot of variety, which leads to a better model.

### Multilayer Perception

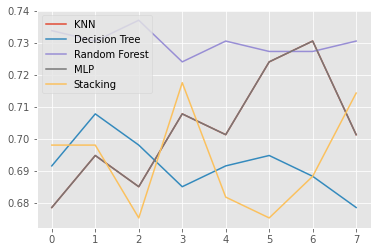
The multilayer perceptron is a system that consists of three layers: the output layer, the hidden layer, and the input layer. Hidden layers are the real engines of the system. Classification will be executed by the output layer. The data will flow through the layers towards both the input and output layers. The reason behind the use of this algorithm as it can predict the frequency of certain events and solve problems that are not linearly separable.

### Stacking Classifier

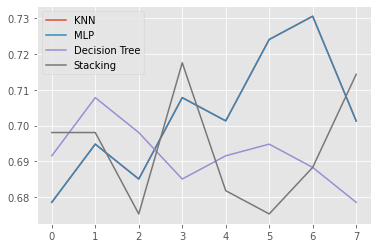
A stack is a technique where multiple classifiers are used to train a meta-classifier. Stacked learning involves stacking the output of several estimators and calculating the result with a classifier. Stacking allows you to take use of each individual estimate's strengths by sending their output into a final estimator.

## **Results**

The results are represented as a line graph for easy interpretation. In the dataset there were few outliers and null results. To exclude these, we replaced every result that was 0 to the categories mean value. This is so that the data can make accurate scores during each model. I used 10-fold validation using 70% for training and 30% for testing for all algorithms. This is so that the prediction is accurate.



I created a line graph which shows relationship of each model used. Based on this, I can understand that random forest outlies the data compare to the rest of model. Therefore, I excluded random forest algorithm.



Based on the second graph excluding the Random Forest, I was able to identify that MLP prediction score was progressively higher than the other models. Therefore, the most successful model for this dataset would be MLP.