

|  |  |
| --- | --- |
|  | **UNIVERSITI TUNKU ABDUL RAHMAN** |
|  | **Lab 1** |
| Course Code: | UEMH3163/UECS2053/UECS2153 |
| Course Name: | Artificial Intelligence |
| Lecturer: | Dr. Ng Oon-Ee |
| Academic Session: | 2019/05 |
| Title: | Supervised Learning |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Name | I.D. No | Course | Practical Group |
| Tan Ying Yao | 1703648 | SE | P1 |
|  |  |  | Mark: /100 |

INTRODUCTION

The dataset selected was Heart disease dataset. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The aim of the analysis is to detect the presence of heart disease in a patient. Supervised learning is used to attempt to distinguish presence (values 1,2,3,4) from absence (value 0). A wide variety of supervised learning tools will be used to measure accuracy of the detection

DATASET COLUMNS FEATURE

Age (age in years)

Sex (1 = male; 0 = female)

CP (chest pain type)

TRESTBPS (resting blood pressure (in mm Hg on admission to the hospital))

CHOL (serum cholesterol in mg/dl)

FPS (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

RESTECH (resting electrocardiographic results)

THALACH (maximum heart rate achieved)

EXANG (exercise induced angina (1 = yes; 0 = no))

OLDPEAK (ST depression induced by exercise relative to rest)

SLOPE (the slope of the peak exercise ST segment)

CA (number of major vessels (0-3)

THAL (3 = normal; 6 = fixed defect; 7 = reversable defect)

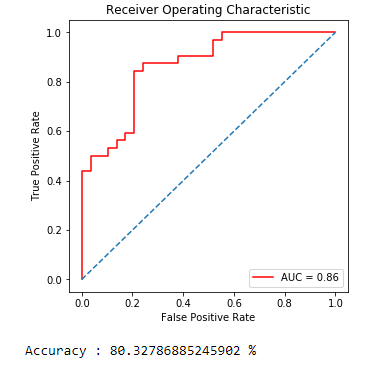
TARGET (1 or 0)

A variety of supervised learning techniques are used to evaluate performance. These are logistic regression, support vector machine (SVM), K nearest neighborhood (kNN), GradientBoostingClassifier and RandomForestClassifier algorithms.

Analysis of Result:

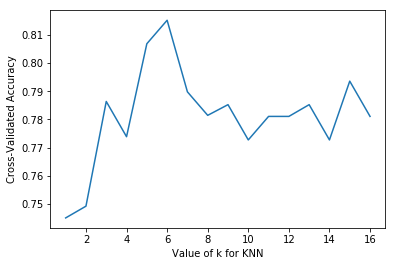
1. Logistic Regression

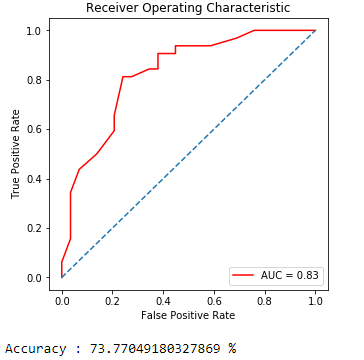
To implement this algorithm model, we need to separate dependent and independent variables within our data sets. In addition, we created a combination of features between different features to make different experiments. While creating these parameters, the process of finding the best results was made by giving hyper parameter values.



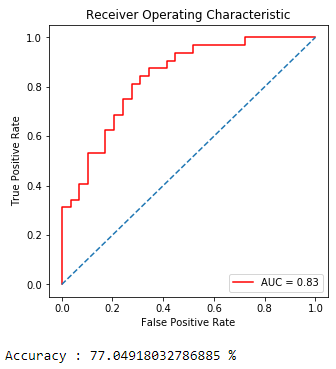
1. K-Nearest Neighbours

Accuracy gained vs the number of neighbors used:Cross validation accuracy vs value of k for KNN. It reaches 81.52% when using knn = 6

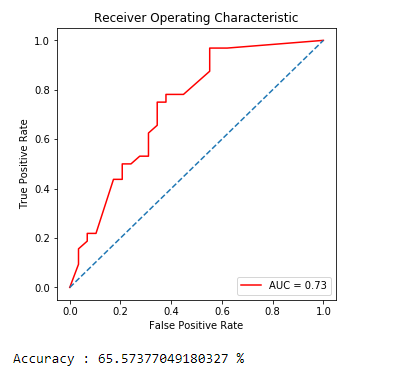




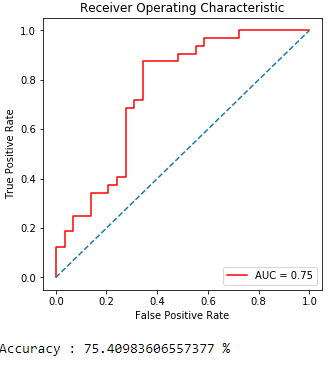
1. Naive Bayes Algorithm



1. Gradient Boosting



1. Random Forest Classifier



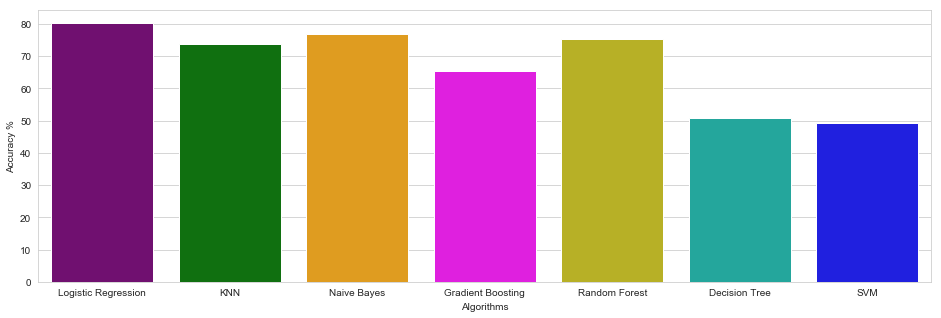
1. Decision Tree

Decision Tree Test Accuracy 50.82%

1. Support Vector Machine

In a data set, the data that are distant from each other are made to scale between each other by making a specific scaling. As a result of this operation, the data takes a value of 0.1. This may change in some scaling operations. Standard and Normalization scale will be used for our operation. There is a big change between the data obtained. Therefore, we need to use this method for SVM algorithm. Accuracy : 49.18 %

Comparison of Model:



From the graph above it is obvious that Logistic Regression has the highest accuracy (80.32%) in comparison to other models. Naives Bayes ranked second (77.04%) while Random Forest takes the third place (75.41%). Decision Tree has a low accuracy of 50.82%). KNN has a high accuracy of 73.77%. The lowest accuracy recorded is by SVM ( 49.18%). Gradient Boosting has an average accuracy of 65.57%.