

|  |  |
| --- | --- |
| **ASSIGNMENT** | |
| **Subject Code** | CSE402A |
| **Subject Name** | Data Mining |
| **Programme/Course** | B. Tech |
| **Department** | CSE |
| **Faculty** | FET |

#### 

|  |  |
| --- | --- |
| **Name of the Student** | Ashish Kumar |
| **Reg. No** | 16ETCS002144 |
| **Semester/Year** | 7th sem/ 4th year |
| **Subject Leader/s** | Prof. N. D. Gangadhar |

**Ramaiah University of Applied Sciences**

**University House, Gnanagangothri Campus, New BEL Road,**

**M S R Nagar, Bangalore, Karnataka, INDIA - 560 054.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Declaration Sheet** | | | | | | | | |
| Student Name | Ashish Kumar | | | | | | | |
| Reg. No | 16ETCS002144 | | | | | | | |
| Programme/Course | B.Tech | | | | | Semester/Year | 7th sem/ 4th year | |
| Subject Code | CSE402A | | | | | | | |
| Subject Title | Data Mining | | | | | | | |
| Subject Date |  | | to | |  | | | |
| Subject Leader | Prof. N. D. Gangadhar | | | | | | | |
| **Declaration**  The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly. | | | | | | | | |
| Signature of the Student | |  | | | | | Date |  |
| Submission date stamp  (by Examination & Assessment Section) | |  | | | | | | |
| Signature of the Subject Leader and date | | | | Signature of the Reviewer and date | | | | |
|  | | | |  | | | | |

# **Question No. 1**

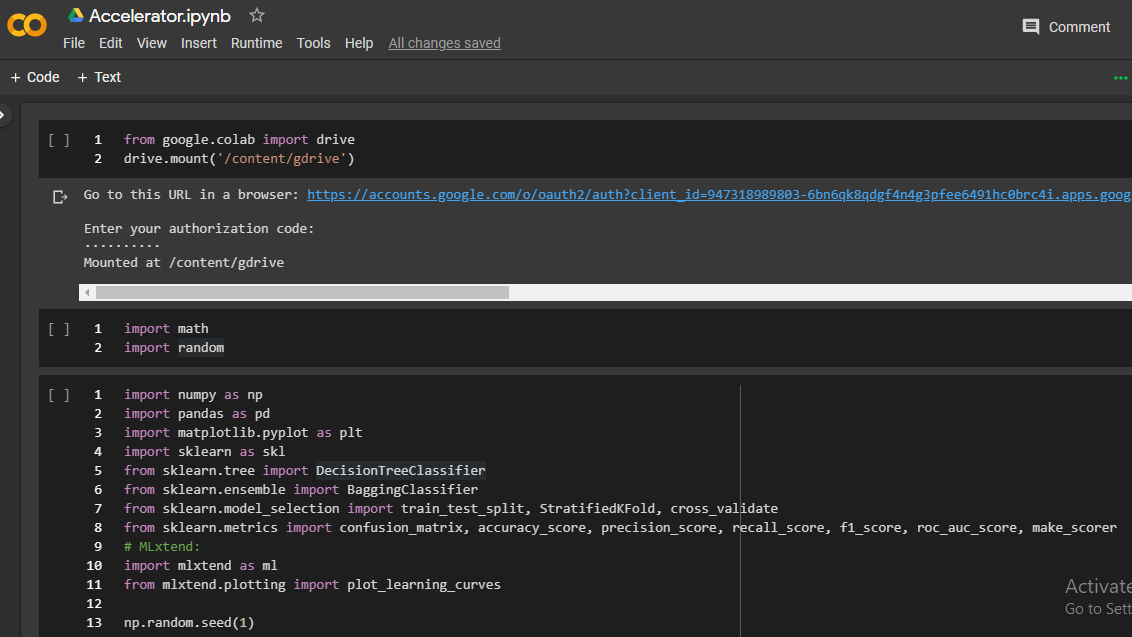
**Solution to Question No. 1:**

## Introduction:

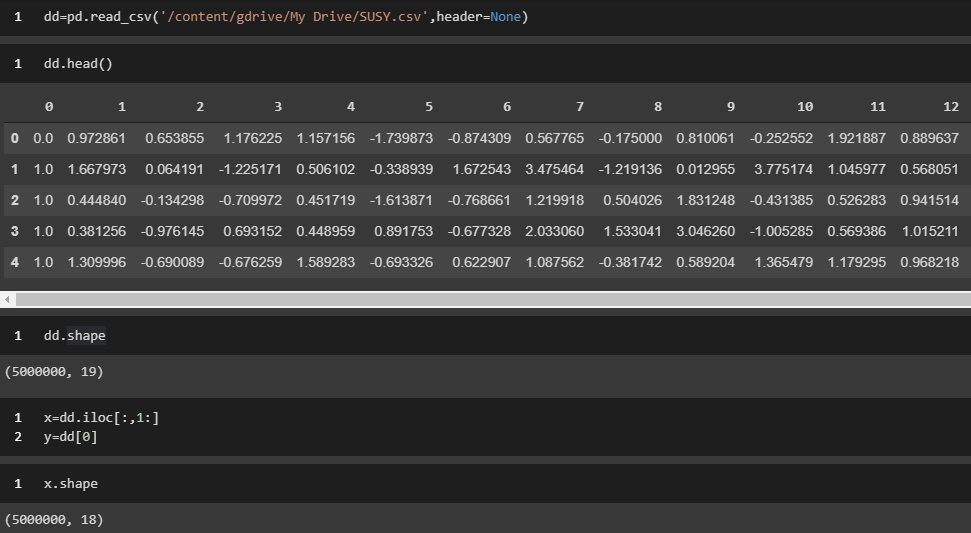
## Classification is a method of identifying an object’s class by learning from a set of known data and their classes, and use that information to classify the class of object whose class is unknown. In a particle accelerator we try to detect presence of fundamental particles like Higgs. To detect these particle, we can use classifier with two labels, particle is present and particle is not present. The presence of a particle depends on the traces left by the experiment. The attributes for classification consists of eight kinematic values measured by the particle detectors and ten derived values from these eight values. We can use different classifier algorithms to classify the above data like SVM, Neural Network and Decision tree etc. In this assignment we will use decision tree as base classifier and random ensemble decision tree because SVM is very complex classifier and neural network has very high computation expense so for such huge data set like in this problem neural network is also not helpful. Decision tree are the best choice as they are very good for tabular data and are simpler to implement and computationally simple too hence we use decision tree classifier. Ensemble of decision tree is used to improve the performance of the algorithm by selecting multiple decision tree to work at a time hence the confidence of the ensemble with be higher compared to single decision tree and will help take more accurate decisions.

## 1.2 Decision Tree Classifier:

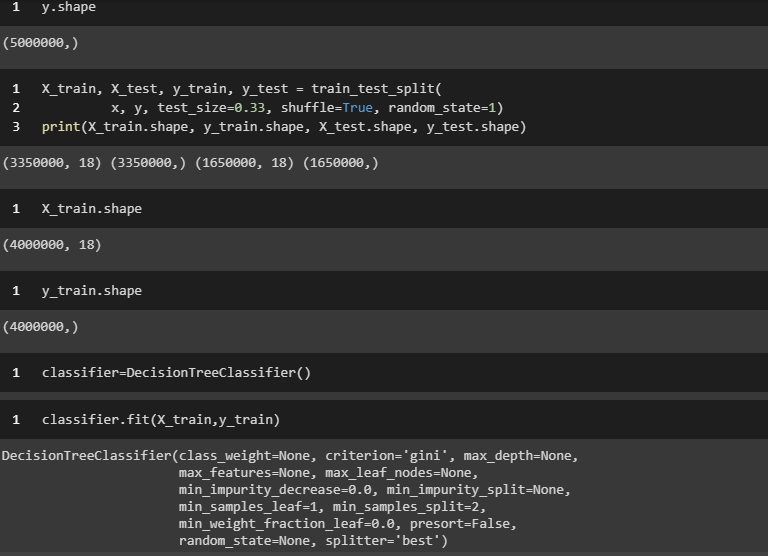
The decision tree classifier implementation:



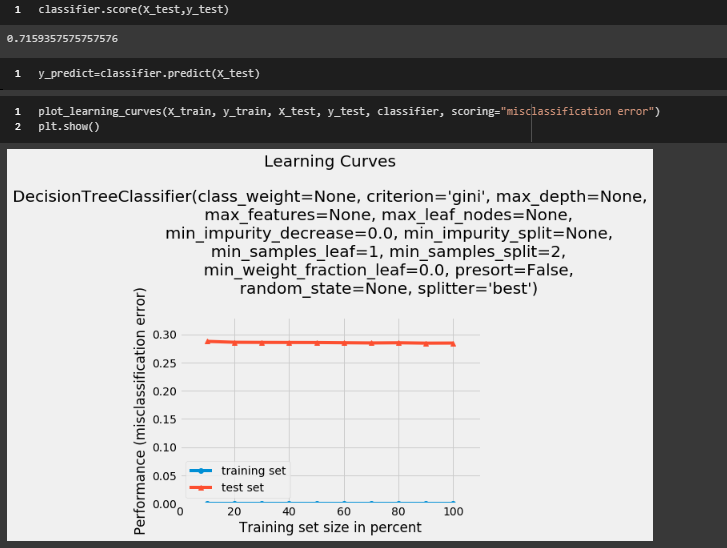
The above figure shows the mounting of the google drive to the google colaboratory notebook and the necessary libraries imported.



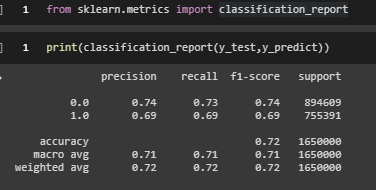
The above figure shows the reading of dataset from drive to colab and extracting the data and labels into x and y variables respectively.



The above figure shows the split of random shuffled data into train and test set in a ratio of 2:1 and the decision tree classifier fitting to the data.



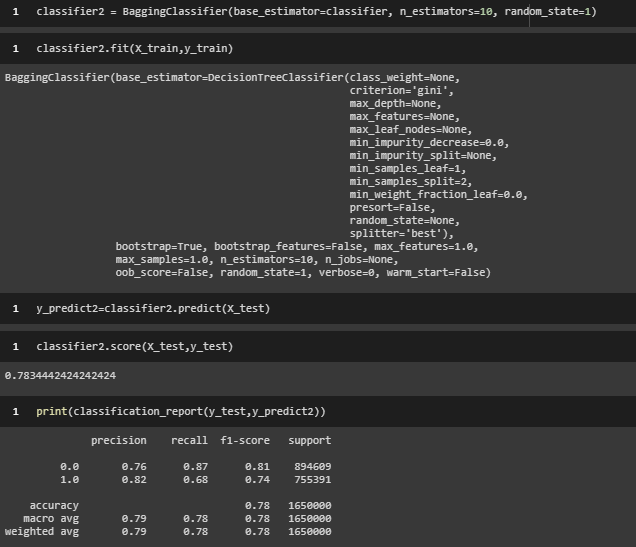
The above figure shows the accuracy score of the classifier (0.7159) and the learning curve. The learning curve can be seen decreasing slowly as the training data is feed to the algorithm, the error being close to less than 0.28.



The above figure shows the report of the classifier, the precision, recall, f1-score, support and accuracy. The accuracy of the algorithm can be seen to be 0.72 (OR 72%).

## 1.3 Ensemble of Decision Tree Classifiers

Ensemble of Decision Tree classifier:



The above figure shows the ensemble of decision tree using bagging classifier with base classifier being the decision tree classifier developed above and no of estimator (no of decision trees) used is 10. We can also see the fit of the bagging algorithm and the accuracy score bring 0.78344 which is higher as compared to the 0.7159 that was achieved with a single decision tree. The classification report shows the precision, recall, f1-score, support and accuracy of the classifier.



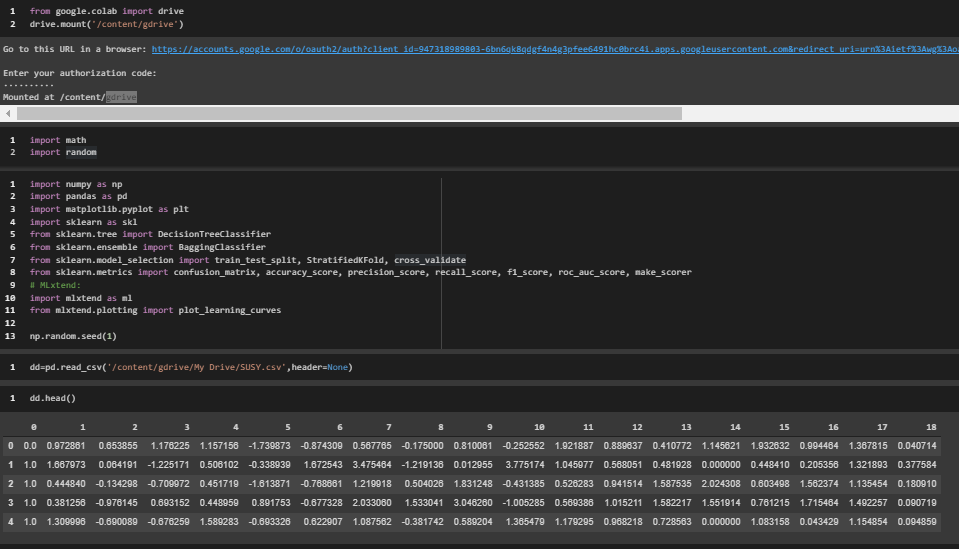
The above figure shows the learning curve of the bagging classifier where we can see the gradual decrease in the error as the no of data feed to the algorithm increases, the error being close to 0.22 which is less compared to 0.28 obtained for the single decision tree.

## 1.4 Decision Tree Classifier with Selected Data Attributes

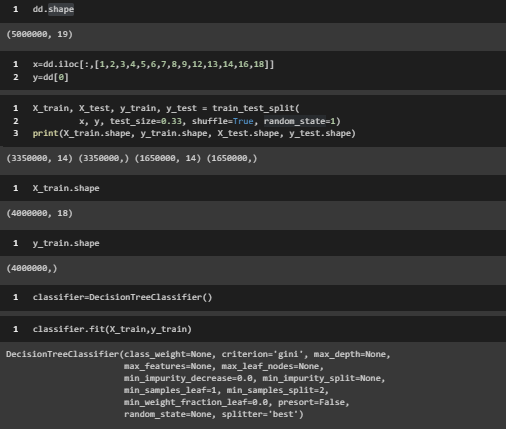
## Decision tree classifier with selected data attributes

## Selected attributes are attribute no: 1,2,3,4,5,6,7,8,9,12,13,14,16,18

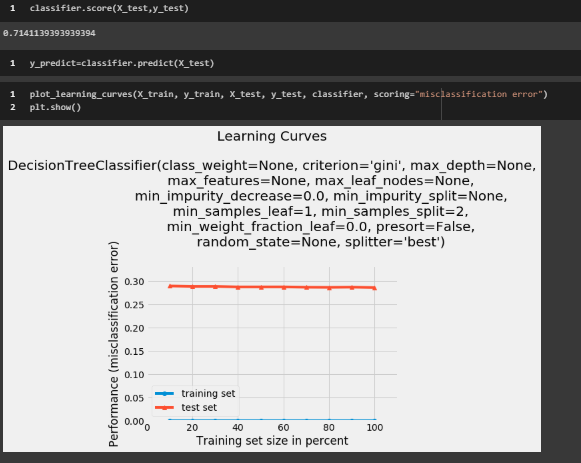
## Decision tree for the selected 14 attributes:



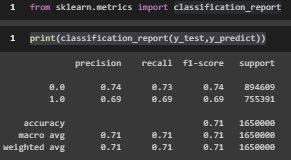
The above figure shows the google drive mounted to the google colab and the necessary libraries imported and the data read from the drive.



The above figure shows the selected attributes selected for training in x variable and the split of random shuffled dataset into train and test set in ratio 2:1. Then the decision tree classifier is fit to the selected data.



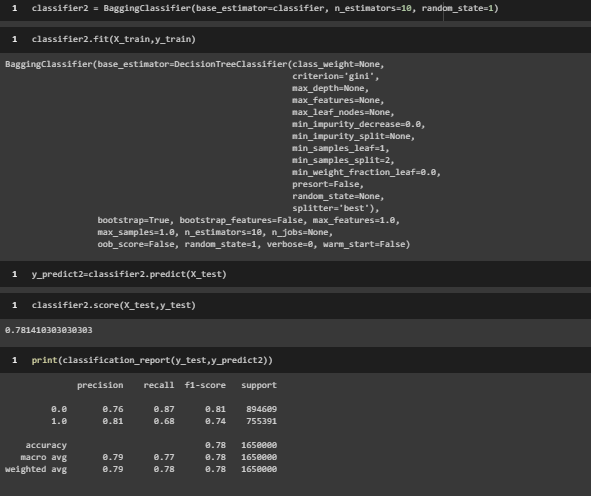
The above figure shows the accuracy score of the classifier 0.71411 which is less than the accuracy score of the decision tree classifier with all the attributes due to loss of information by not selecting some attributes. The figure also shows the learning curve of the classifier which is decreasing as more training data is feed to the algorithm, the error is closed to less than 0.28.



The above figure shows the classification report of the classifier, the precision, recall, f1-score, support and accuracy.

## 1.5 Ensemble of Decision Tree Classifiers with Selected Attributes:

Ensemble of decision tree classifier with selected attributes:



The above figure shows the fit of the bagging classifier with selected attributes and the accuracy score 0.7814 which is more than the accuracy of decision tree with selected attributes but less than the bagging classifier with all the attributes due to loss of information with loss of attributes. The classifier report shows the precision, recall, f1-score, support and accuracy of the algorithm.



The above figure shows the learning curve of the ensemble of decision tree with selected attributes. It can be seen that the error gradually decreases as more data is feed to the algorithm, the error being close to less than 0.22 which is less than decision tree with selected attributes and almost equal to the error of ensemble decision tree classifier with all the attributes.

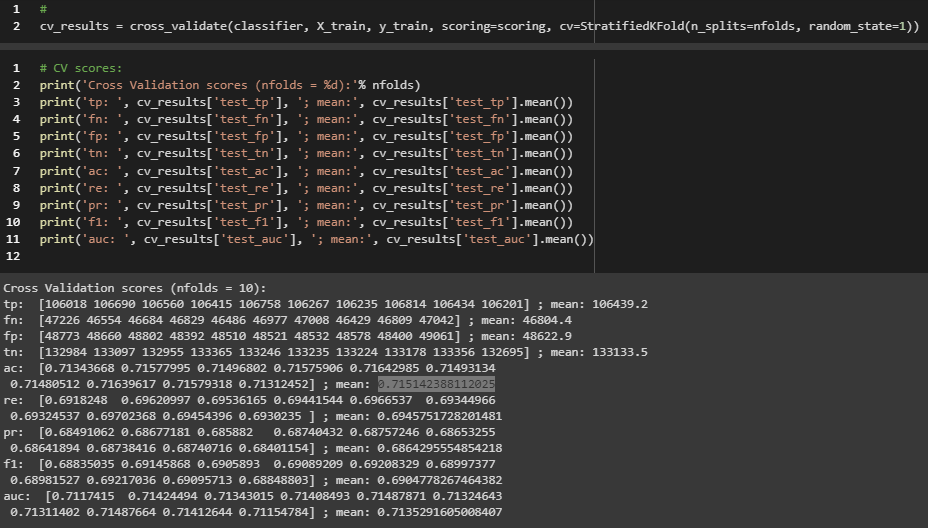
## 1.6 Comparative evaluation of the Classifiers

## Comparative evaluation of the classifiers:

**Cross validation for decision tree with all attributes**

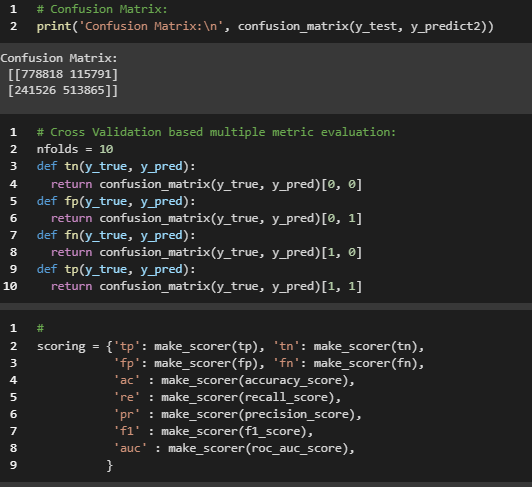


The above figure shows the confusion matrix of the decision tree algorithm.

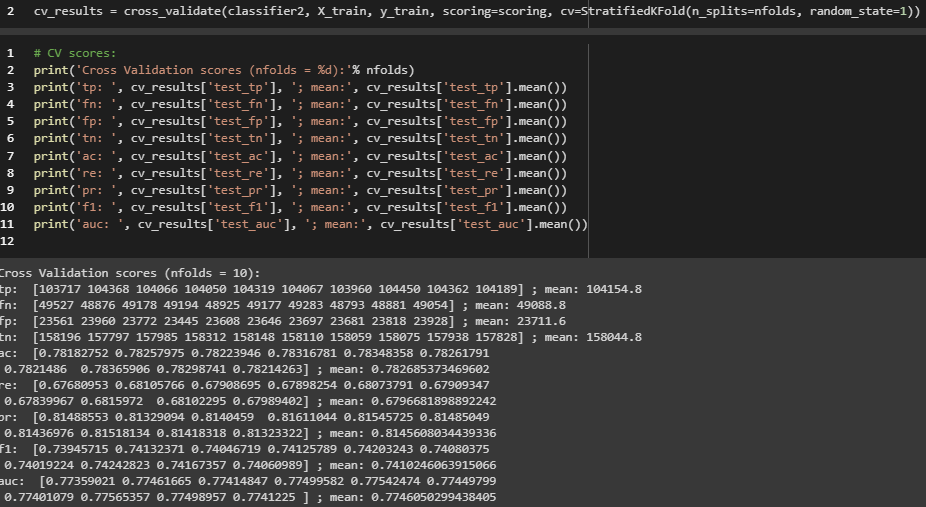


The above figure shows the cross validation for 10 fold for the decision tree algorithm. We can see the values of 10 folds true positive, false positive, true negative, false negative, accuracy, recall, precision, f1-score and AUC (area under the curve). The accuracy and auc is above 0.5 hence the algorithm is working correctly.

**Cross validation for bagging algorithm with all attributes:**

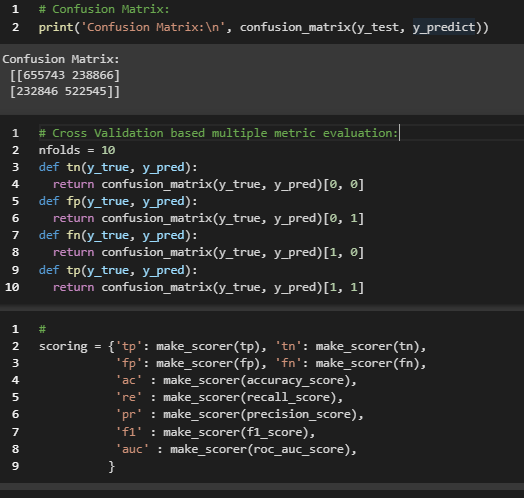


The above figure shows the confusion matrix of bagging algorithm.

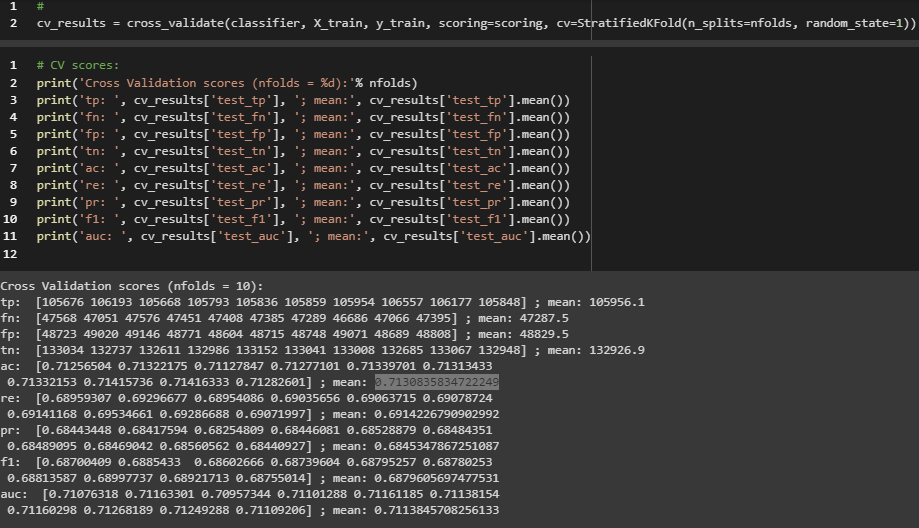


The above figure shows the cross validation for 10 fold for the bagging algorithm. We can see the values of 10 folds true positive, false positive, true negative, false negative, accuracy, recall, precision, f1-score and AUC (area under the curve). The accuracy and auc is above 0.5 hence the algorithm is working correctly.

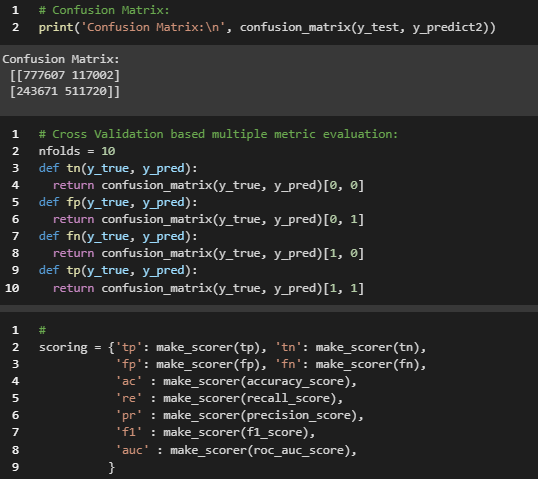
**Cross validation for decision tree with selected attributes:**



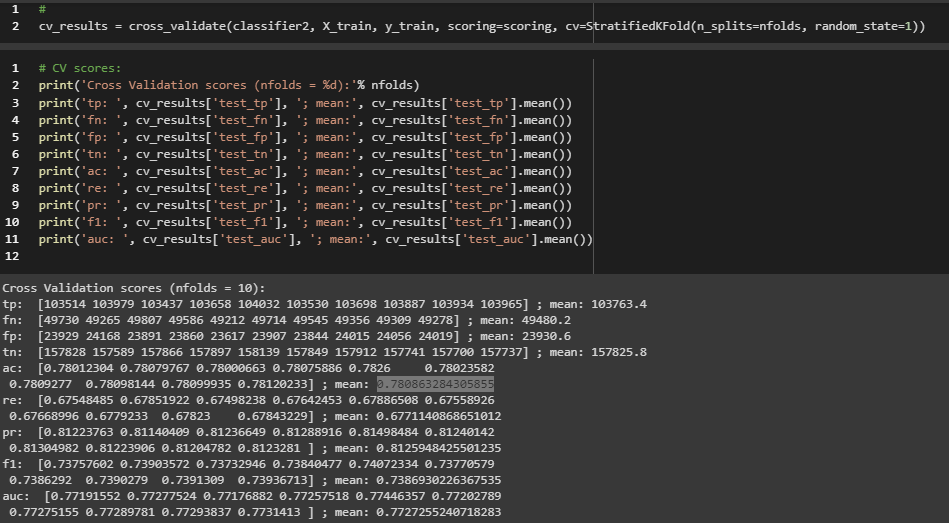
The above figure shows the confusion matrix of the decision tree algorithm with selected attributes.



The above figure shows the cross validation for 10 fold for the decision tree algorithm. We can see the values of 10 folds true positive, false positive, true negative, false negative, accuracy, recall, precision, f1-score and AUC (area under the curve). The accuracy and auc is above 0.5 hence the algorithm is working correctly.

**Cross validation for bagging algorithm with selected attributes:**  


The above figure shows the confusion matrix for bagging algorithm with selected attributes.



The above figure shows the cross validation for 10 fold for the decision tree algorithm. We can see the values of 10 folds true positive, false positive, true negative, false negative, accuracy, recall, precision, f1-score and AUC (area under the curve). The accuracy and auc is above 0.5 hence the algorithm is working correctly.

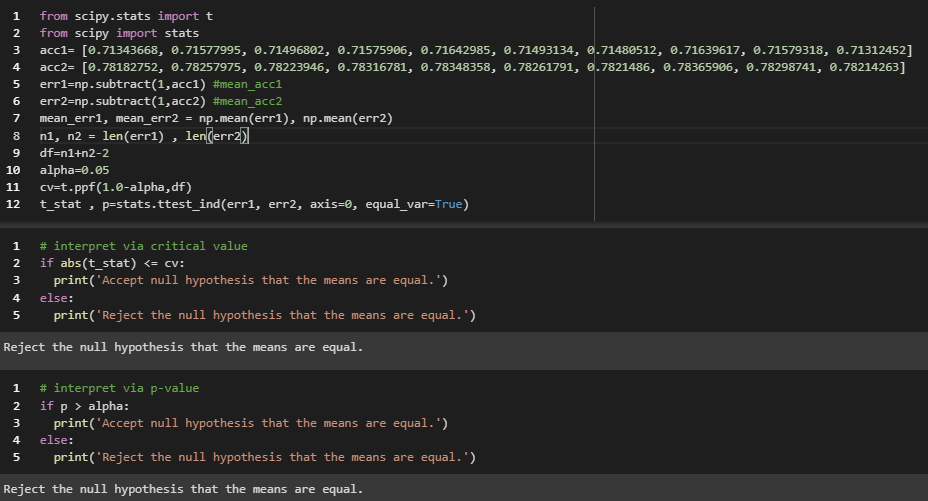
**Comparison between decision tree algorithm with all attributes and bagging algorithm with all attributes:**

We can observe from the value of accuracy and AUC of both the algorithms below:

|  |  |  |
| --- | --- | --- |
|  | Decision tree algorithm | Bagging of decision tree algorithm |
| Accuracy | 0.715142388112025 | 0.782685373469602 |
| AUC | 0.7135291605008407 | 0.7746050299438405 |

The value of accuracy and AUC of both the algorithm is more than 0.5 hence the algorithm is able to predict correctly for more than 50% of times. The accuracy of decision tree is 71.51% whereas accuracy of bagging algorithm is 78.26% which is higher than the accuracy of decision tree. This is due to ensemble of the decision tree in bagging which improves the performance of the algorithm by using different decision tree for different cases, in other word the ensemble contains different decision trees which has better performance for different types of input hence a single decision tree may miss classify some of the inputs but ensemble is more likely to correctly classify the inputs.

T-test for verifying the skill scores:



The above figure shows the t-test for verifying the skill scores and since the hypothesis is rejected hence we can conclude that the skill scores is statistically significant and the above conclusion holds correctly.

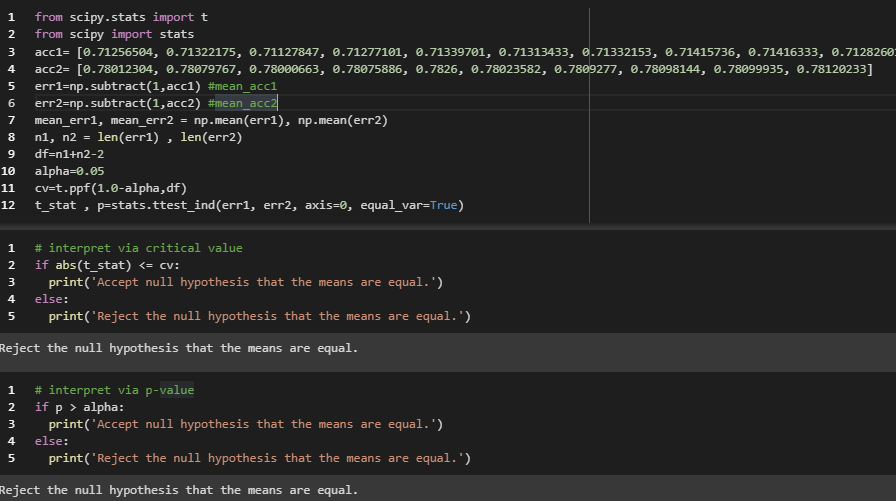
**Comparison between decision tree algorithm with selected attributes and bagging algorithm with selected attributes:**

We can observe from the value of accuracy and AUC of both the algorithms below:

|  |  |  |
| --- | --- | --- |
|  | Decision tree algorithm | Bagging of decision tree algorithm |
| Accuracy | 0.7130835834722249 | 0.780863284305855 |
| AUC | 0.7113845708256133 | 0.7727255240718283 |

The value of accuracy and AUC of both the algorithm is more than 0.5 hence the algorithm is able to predict correctly for more than 50% of times. The accuracy of decision tree is 71.30% whereas accuracy of bagging algorithm is 78.08% which is higher than the accuracy of decision tree. This is due to ensemble of the decision tree in bagging which improves the performance of the algorithm by using different decision tree for different cases, in other word the ensemble contains different decision trees which has better performance for different types of input hence a single decision tree may miss classify some of the inputs but ensemble is more likely to correctly classify the inputs.

T-test for verifying the skill scores:



The above figure shows the t-test for verifying the skill scores and since the hypothesis is rejected hence we can conclude that the skill scores is statistically significant and the above conclusion holds correctly.

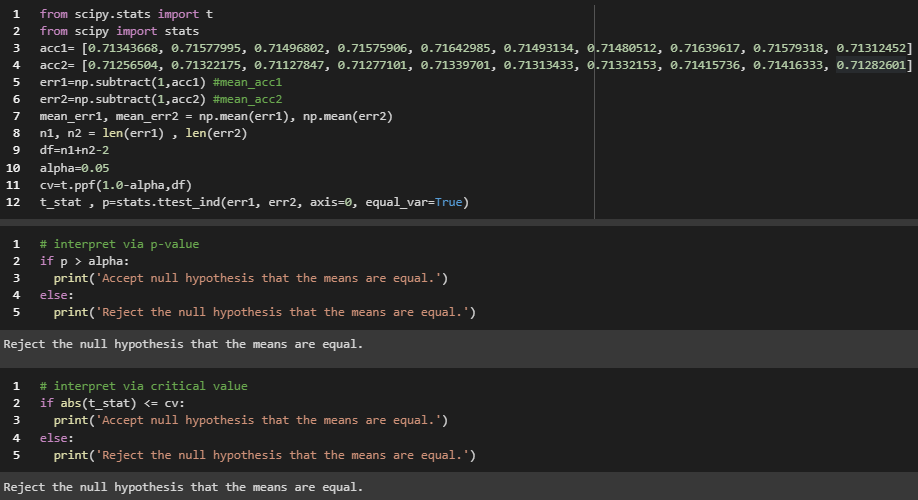
**Comparison between decision tree algorithm with all attributes and decision tree with selected attributes:**

We can observe from the value of accuracy and AUC of both the algorithms below:

|  |  |  |
| --- | --- | --- |
|  | Decision tree with selected attributes | Decision tree with all attributes |
| Accuracy | 0.7130835834722249 | 0.715142388112025 |
| AUC | 0.7113845708256133 | 0.7135291605008407 |

The value of accuracy and AUC of both the algorithm is more than 0.5 hence the algorithm is able to predict correctly for more than 50% of times. The accuracy of decision tree with selected attributes is 71.30% whereas accuracy of decision tree with all attributes is 71.51% which is higher than the accuracy of decision tree. This is due to loss of information in the decision tree with selected attributes, more data means more information which can help build better classifier with better performance of the algorithm. By using all attributes the classifier uses all the information available hence makes better prediction hence decision tree with all attributes has higher accuracy than decision tree with selected attributes.

T-test for verifying the skill scores:



The above figure shows the t-test for verifying the skill scores and since the hypothesis is rejected hence we can conclude that the skill scores is statistically significant and the above conclusion holds correctly.

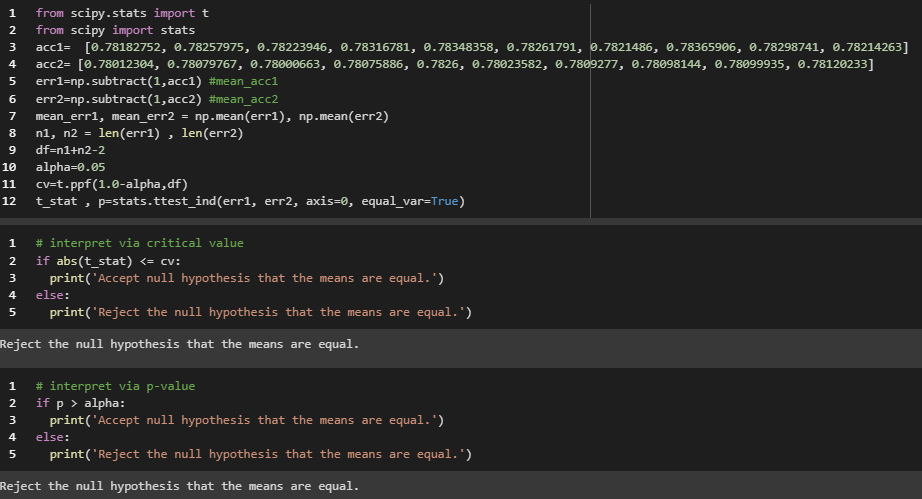
**Comparison between ensemble of decision tree algorithm with all attributes and ensemble of decision tree with selected attributes:**

We can observe from the value of accuracy and AUC of both the algorithms below:

|  |  |  |
| --- | --- | --- |
|  | Ensemble with selected attributes | Ensemble with all attributes |
| Accuracy | 0.780863284305855 | 0.782685373469602 |
| AUC | 0.7727255240718283 | 0.7746050299438405 |

The value of accuracy and AUC of both the algorithm is more than 0.5 hence the algorithm is able to predict correctly for more than 50% of times. The accuracy of ensemble of decision tree with selected attributes is 78.08% whereas accuracy of ensemble of decision tree with all attributes is 78.26% which is higher than the accuracy of decision tree. This is due to loss of information in the ensemble of decision tree with selected attributes, more data means more information which can help build better classifier with better performance of the algorithm. By using all attributes the classifier uses all the information available hence makes better prediction hence ensemble of decision tree with all attributes has higher accuracy than ensemble of decision tree with selected attributes.

T-test for verifying the skill scores:



The above figure shows the t-test for verifying the skill scores and since the hypothesis is rejected hence we can conclude that the skill scores is statistically significant and the above conclusion holds correctly.

## 1.7 Conclusions

In this assignment 4 classifiers were developed, decision tree algorithm, ensemble of decision tree, decision tree with selected attributes and ensemble of decision tree with selected attributes. All the different algorithm had different performance for the same input and test data. The ensemble of decision tree performed better than the single decision tree for both the cases due to ensemble which helped the algorithm perform better by using multiple decision trees. For ensemble 10 estimators are used with base classifier being the decision tree itself. The performance skill score verified by the T-test which rejects the hypothesis hence concluding that the skill scores are statistically significant. It is also seen that the algorithms with selected attributes has less accuracy than the algorithms that use all the attributes which is due to loss of information in selecting attributes and leaving some attributes. Therefore, ensemble of decision tree with attributes has best performance among all the 4 algorithms and the decision tree with selected attributes has least performance. Hence it can be concluded that best algorithm to use for this problem is ensemble algorithm with all attributes unless there is a problem of overfitting in which case few attributes can be dropped.