**REPORT: CAR EVALUATION**

**1.Introduction**

1. **Problem description:**

Using a private means for transportation is a norm in society today, and what better than a car could it be when it comes to one such means. Owning a car gives an individual a sense of status and the opportunity for being in personal control and autonomy. In thinly populated areas, owning a car is crucial, as it provides the only means for travelling long distances due to a lack of public transport. Even when public transport is available, people may express anxiety about the poor quality of public transport services. The lack of feasible transport alternatives to the private automobile make owning one undisputable. Keeping this in mind, it becomes even more important to evaluate a car which you want to own so that it’s a best fit in terms of price, maintenance, comfort, safety apart from fuel efficiency.

1. **Motivation:**

Buying a car is always exciting and it is also a composite process through which sometimes you end up overpaying an extra amount or may not be happy with a vehicle you purchased when you drive it. Researching and evaluating a car becomes essential before purchasing. Hence it is crucial to consider the factors like maintenance price, buying price, comfort, safety features, seating capacity, luggage boot space etc., before coming to a decision of owning a car. Thus, to get the best car that includes all these features along with being reasonably priced is our motivation behind choosing a Car evaluation dataset for the study.

1. **Brief description of your report organization**

The report begins with Introduction which consists of Problem description and Motivation. Then the Data Exploration part where we describe our dataset, followed by Methodology where we describe the data mining techniques we have used and its performance. Lastly but not the least we end the report by conclusion followed by Appendix and Reference.

**2. Data Exploration**

In our project, based on the dataset we think that there is no preprocessing required. We feel there is no need for discretization or aggregation since the attributes are character type. Every attribute is essential for classification; hence we are not performing feature selection.

**Type of dataset file used:** .arff file format

**Number of records:** The dataset consists of 1728 records.

**Number of attributes and description:** The dataset consists of six attributes and class values. All the six attributes are of type character. The model evaluates the car depending on six given attributes and classifies into unacceptable, acceptable, good and very good car class.

**Missing values and outliers:** None

**Classification model used:** We have considered all classification models taught in class and based on the obtained results; we have decided that Random Forest Algorithm is the best model to classify this dataset.

**3. Methodology**

1. **Data Preprocessing**

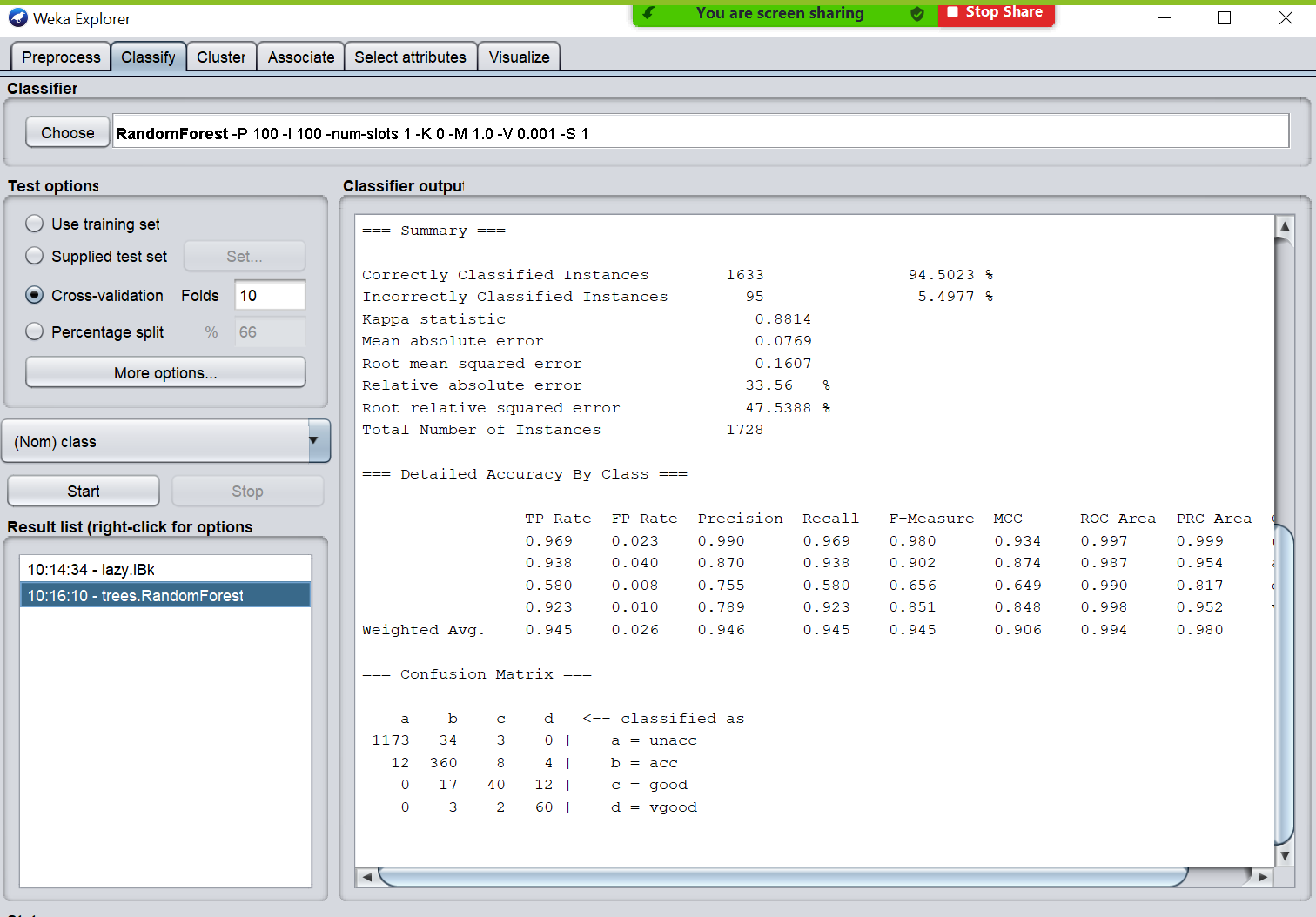
We did not perform Data Preprocessing techniques like discretization, aggregation, and feature selection since the dataset consisted only of character type and all attributes were equally important for model creation.

1. **Mining the data**

**I. Data mining techniques** - Random Forest and KNN Data mining techniques

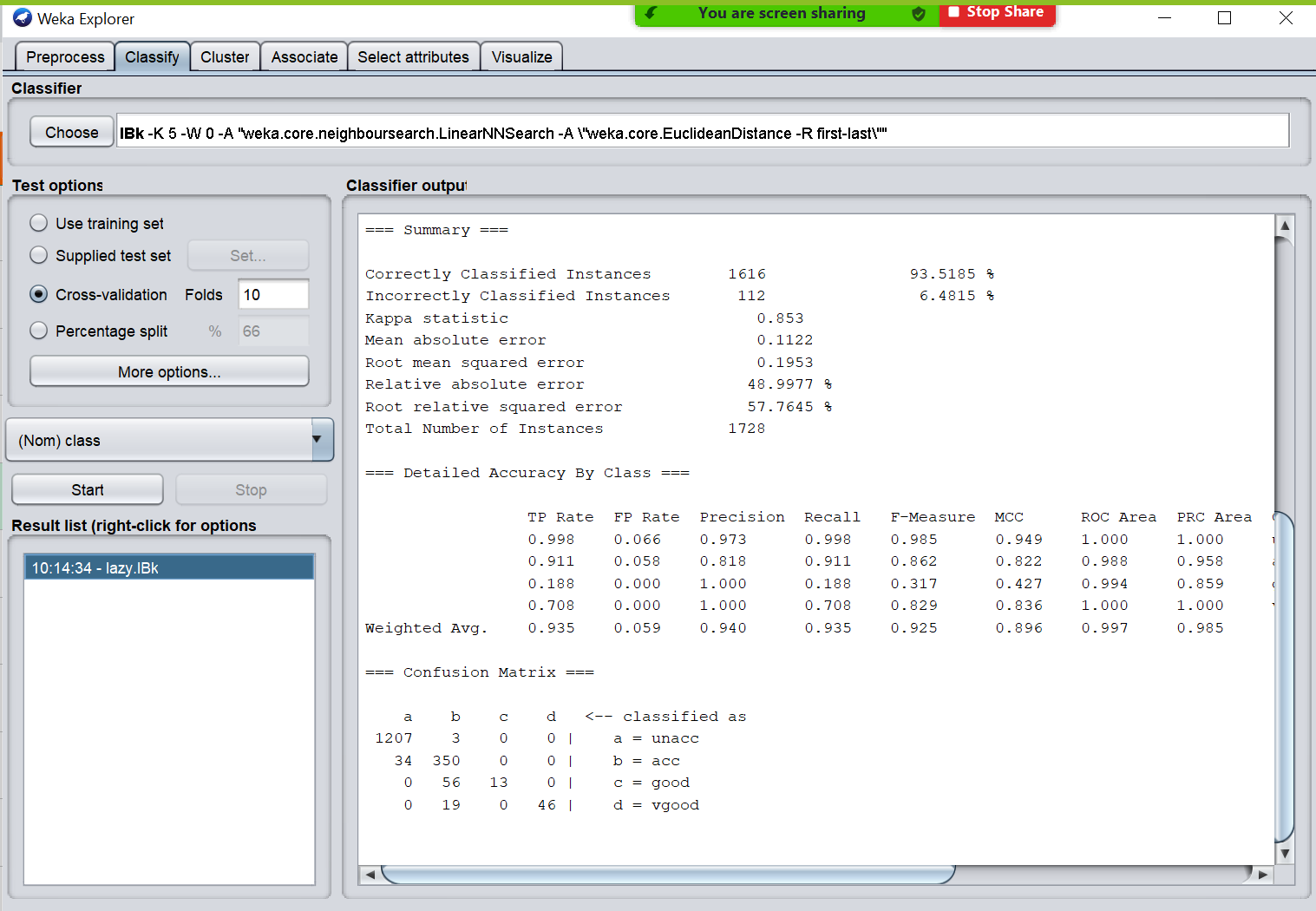
* **Random Forest**

Random Forest is an ensemble classification Data mining technique which builds multiple Decision trees and takes the Majority predicted class votes from all the Decision trees to predict the class. We tuned parameters like batchsize, seeds and a few others, but at the end, the result like correctly classified instances, ROC Area and confusion matrix did not change. Random Forest has the highest accuracy among many other tested algorithms, and accuracy alone isn’t enough to say that it suits our dataset. Based on the below figure, all the measures like TP Rate, FP rate, Precision, F-Measure and a few others prove that Random Forest is a good fit for our dataset. TP rate is the True Positive rate, which is the ratio of number of True Positives to sum of number of True Positive and False Negative. FP rate is the False Positive rate, which is the ratio of number of False Positives to sum of number of True Negative and False Positive. Precision is the proportion of positive identifications which was actually correct. Recall is the proportion of actual positives that was correctly identified.



* **K-Nearest Neighbors**

K-Nearest Neighbors is a classification Algorithm, as the name suggests the class predicted based on the given number[K] of Nearest Neighbors. We tuned parameters like K-value, nearestNeighbourSearchAlgorithm, and a few others, but the results did not change unless we give extreme K-values, which is not suggested for K-Nearest Neighbors Algorithm. Has the second highest accuracy after Random Forest among other tested algorithms. We also got a precision of 1 for classes good and very good, also the ROC Area is 1 for classes unacceptable and very good. We believe that IBk model is also a good fit for our dataset as it illustrates satisfactory model performance values.



**II. Models Performance**

We used Correctly classified instances (accuracy), ROC Area and Confusion Matrix to evaluate the Model Performance. Correctly classified instances (accuracy) is the measure of the number of correctly classified instances comparing the model’s predicted class to the actual class after the model is trained. Receiver operator Characteristics curve (ROC) is a graph showing performance of a classification model which is obtained by plotting false positives rate against true positives rate for different thresholds. Greater the area under the curve (AUC), better is the efficiency. Confusion Matrix is the table that also describes the Models performance, comparing all the predicted classes with the actual class, each row in the confusion matrix represents the actual class, and each column represents the predicted class. The diagonal elements of the matrix are the correctly classified instances for the given class, the rest are misclassified instances.

**III. Motivation behind choosing the Models Performance**

We believe that the measures which have been chosen depicts the true performance of the Models. ROC Area gives a better idea about accuracy. Confusion Matrix gives a clear picture of the Models Performance, as it reveals how many instances are incorrectly classified and which class they belong to.

**IV. Performance measures estimation**

We chose Cross validation and gave 10 as the input for the number of folds. Cross Validation will ensure that all instances in the dataset are being used as Test and Training data, by this way the entire dataset will be used for both training and testing purposes.

**V. Performance of Different modes**

| **Test Options** | **Training set** | **Percentage Split - 66%** | **10-fold cross validation** |
| --- | --- | --- | --- |
| IBk | 100% | 92.675% | 93.457% |
| Random Forest | 100% | 92.517% | 94.502% |

The training set results are always expected to be higher and most of the time the expected accuracy is 100% since the same set is used for both training and testing purposes, which evidently result in higher accuracy and misleading. The Percentage Split is done at 66%, which means that 66% of the dataset is used for training purposes and 34% of the dataset is used for testing purposes. Although Percentage split may seem a better mode to evaluate the performance, we may include a few outliers in the training set to make overfit and model or underfit the model be excluding some crucial instances while training the model. Cross validation is by far the best way to evaluate the model as it uses the entire dataset for training and testing purposes.

**4. CONCLUSIONS:**

1. **Result Comparison**

**I. Patterns confirmed by different models:**

We studied the dataset using different models like SMO, IBk, JRip, Naïve Bayes, J48, Random Forest and Bagging. Each model was tested using 10-fold cross validation and Percentage split using Weka tool. The following patterns were confirmed by different models:

For SMO, highest accuracy with cross validation was 92.935%.

For IBk, highest accuracy with cross validation was 93.457%.

For JRip, highest accuracy with cross validation was 86.458%.

For Naïve Bayes, highest accuracy with cross validation was 85.532% .

For J48, highest accuracy with cross validation was 92.361%.

For Random Forest, highest accuracy with cross validation was 94.502%.

For Bagging, highest accuracy with cross validation was 91.666%.

Thus, it is seen that in most of the models 10 fold cross validation gives better accuracy as it takes the average of the squared errors in 10 iterations as compared to percentage split.

From the Weka tool, it is clearly understood that Random forest is an aggregation of multiple J48’s; Hence the accuracy of both Random forest as well as J48 is high. Since Naïve Bayes uses the probability outcomes, the accuracy is low.

**II. Particular patterns discovered**:

We used the Weka tool to study the car evaluation dataset using classification models. We found patterns in the model output. Training set always gives highest efficiency for any model. But we cannot rely on it as Training data refers to the data used to "build the model". Percentage split produces reliably accurate results for different models. But, the split is about partitioning the dataset you provide into a training and a test set. The classifier is trained on the train set and subsequently evaluated on the test set. The larger the training set, the more accurate the classifier. But the less informative is the score you get on the test set, since a small test set does not give conclusive results. 10 fold cross validation is found to be best technique used to evaluate different models for accuracy as the dataset is divided into 9training sets and one test set. Thus 10 iterations are carried out every time with different training and test sets and finally average of the errors is taken. Hence we can rely on the accuracy obtained. While testing under IBk model, when k>12, the accuracy decreases greatly. Thus, we understood that having a large value for k will include incorrect nearest neighbors resulting in wrong prediction. With this in picture, we understood that it is a good idea to have odd values for k in the IBk algorithm. Also, better to have a k < 10.

**III. Models that work better:**

Among the different classification model used in the study, we found that Random Forest with an accuracy of 94.502% under 10-fold cross validation is the best model to be used for car evaluation dataset. Apart from accuracy, even the confusion matrix with least false positives and the highest area under receiver operator characteristics curve along with TP Rate, FP Rate, F-measure support our observation. It is followed by IBk as the second best model with an accuracy of 93.457%

1. **Summarization of problem and conclusions drawn from the built models:**

It is crucial to consider the factors like maintenance price, buying price, comfort, safety features, seating capacity, luggage boot space etc., before coming to a decision of owning a car. So, we had to research how to decide the best values to be considered for each of these factors so that we can evaluate the car into the appropriate class in an efficient manner. It was also challenging for us to consider various stakeholders’ point of view as well as the owner’s point of view while evaluating a car to get a wholesome perspective from both the parties involved.

When it comes to classifying the dataset using different models, we had to make sure that we considered all the different types of classification models available in the Weka tool. Thus, ensured that we did a comparative study and were able to find the best classification model. We also had to test the different models using all the available test options like training set, percentage split and 10-fold cross validation so that better accuracy could be attained. When it came to choosing the best model for car evaluation we also looked at correctly classified instances, confusion matrix and AUC-ROC. With all these factors in picture, we were able to find out that 10 fold cross validation is the suitable test option to be considered to get reliable accuracy. We should look for ROC to have a greater area. The confusion matrix should have the maximum diagonal values which belong to the correct class and least number of false positives which belong to the incorrect class. Thus, with all these considerations in picture, we found that the Random Forest algorithm proves to be a best fit classification model to evaluate the Car evaluation dataset.

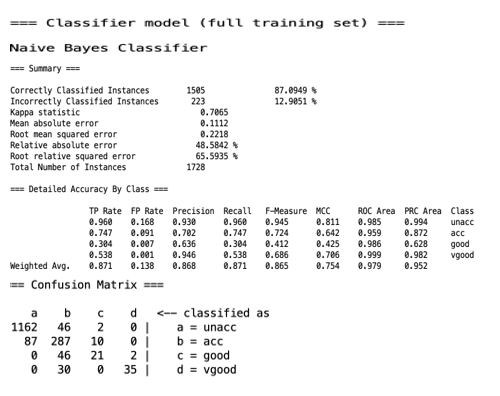
1. **What next can be done if the project was continued:**

We can consider additional attributes that can be included in the dataset to study car evaluation. Currently, all our attribute values are of character type. So, we can try modifying the type to integer, decimal or any other type as suitable to check if the obtained results change with change in data type. We do not have outliers or missing attributes in the dataset. So, we can introduce some missing values and outliers to test how they influence the classification results/accuracy. Study techniques that can be used to reduce false positives and obtain better accuracy. Extend our study to a few more models that can be applied during classification and make a comprehensive comparative study. Develop an open source platform having a front end and back end application to help both car buyers and sellers.

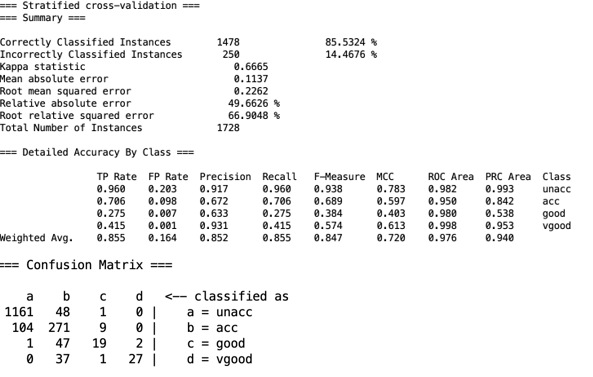
**APPENDIX**

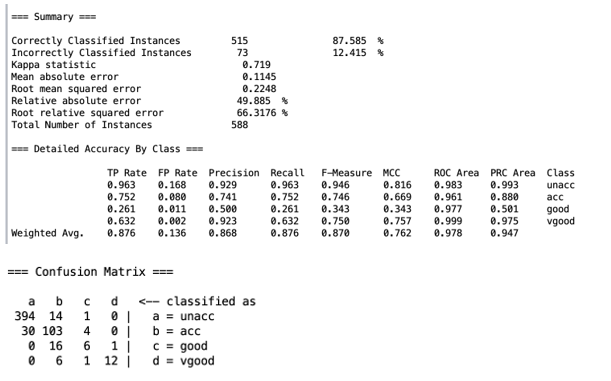
**Naïve Bayes Classifier:**

**Using training set**



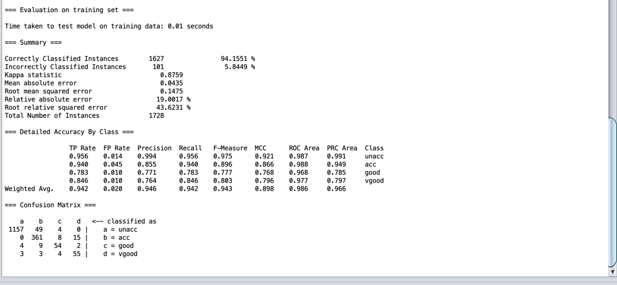
**Using cross validation**



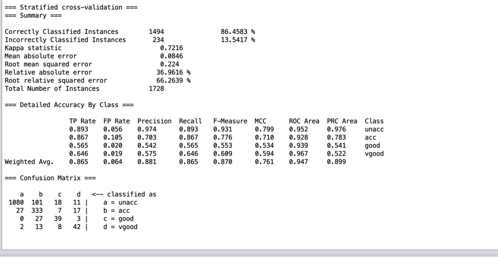
**Using Percentage Split**

**JRip:**

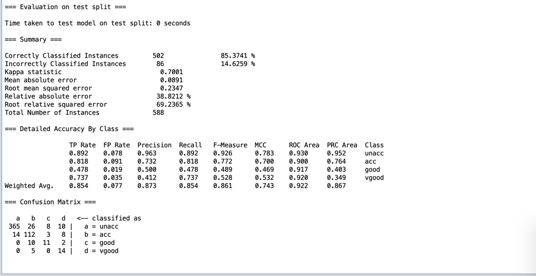
**Using training set**



**Using cross validation**

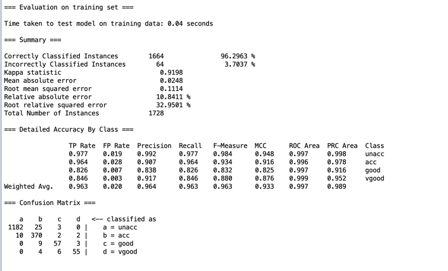


**Using Percentage split**

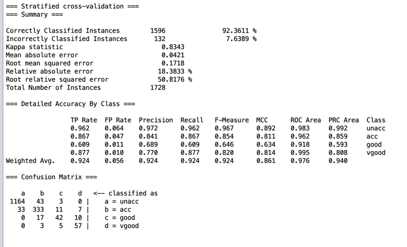


**J48:**

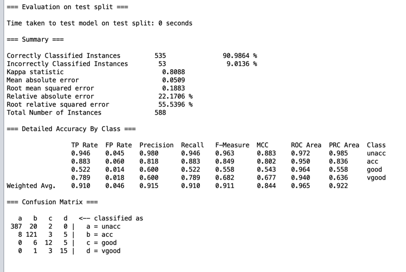
**Using training set**



**Using cross validation**

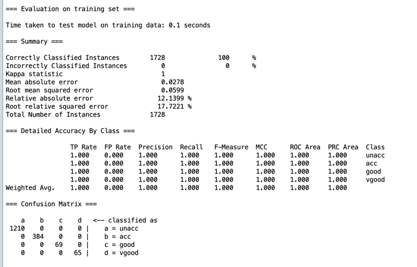


**Using percentage split**

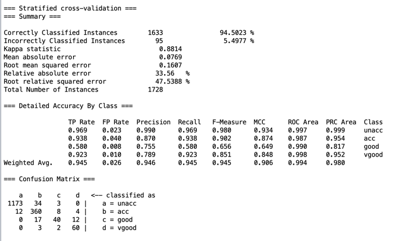


**Random Forest:**

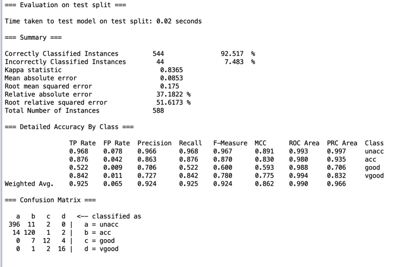
**Using training set**



**Using cross validation**

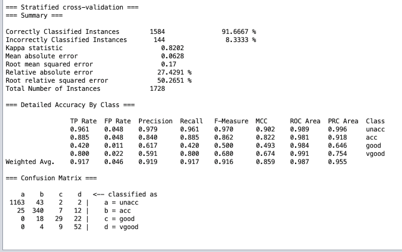


**Using percentage split**

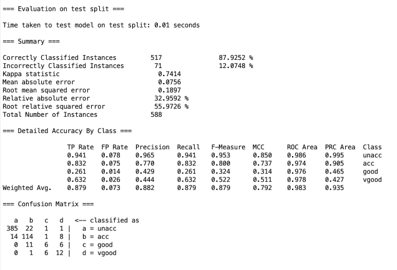


**Bagging**

**Using cross validation**

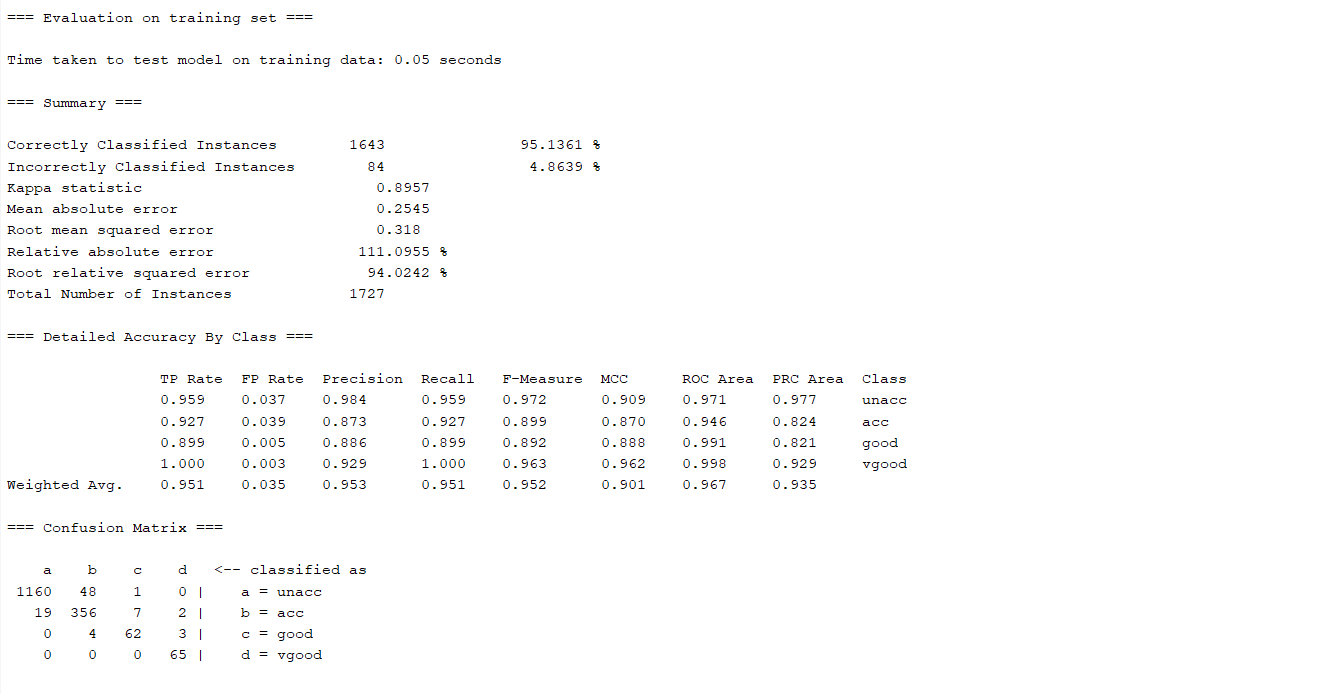


**Using percentage split**

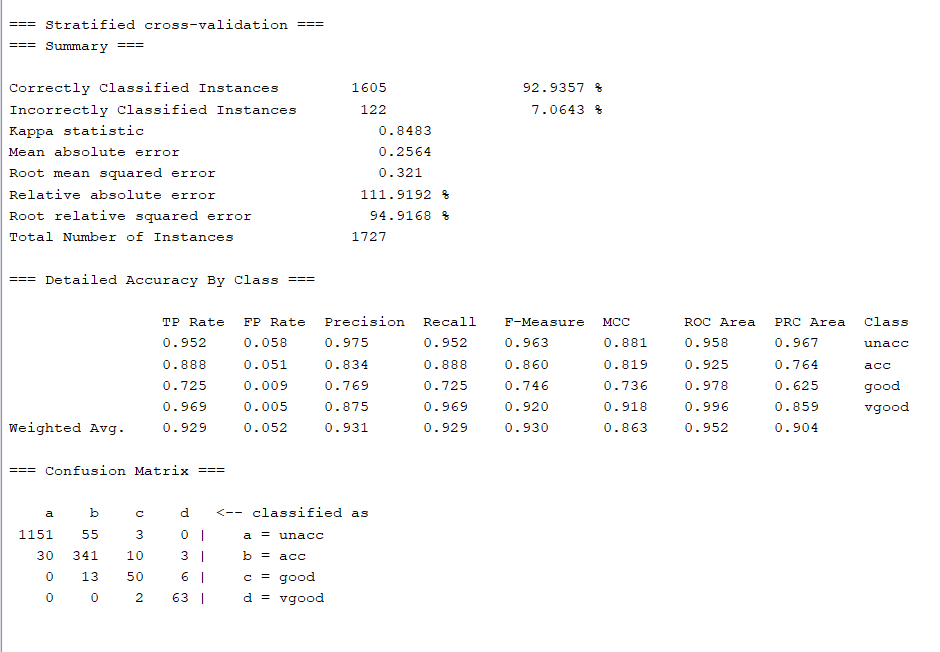


**SMO:**

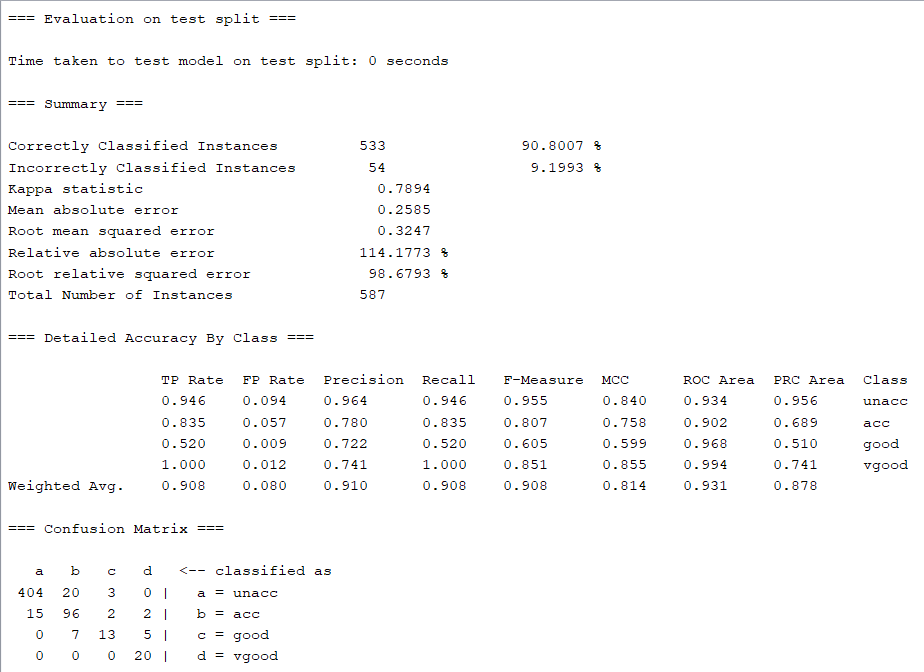
**Using training set**



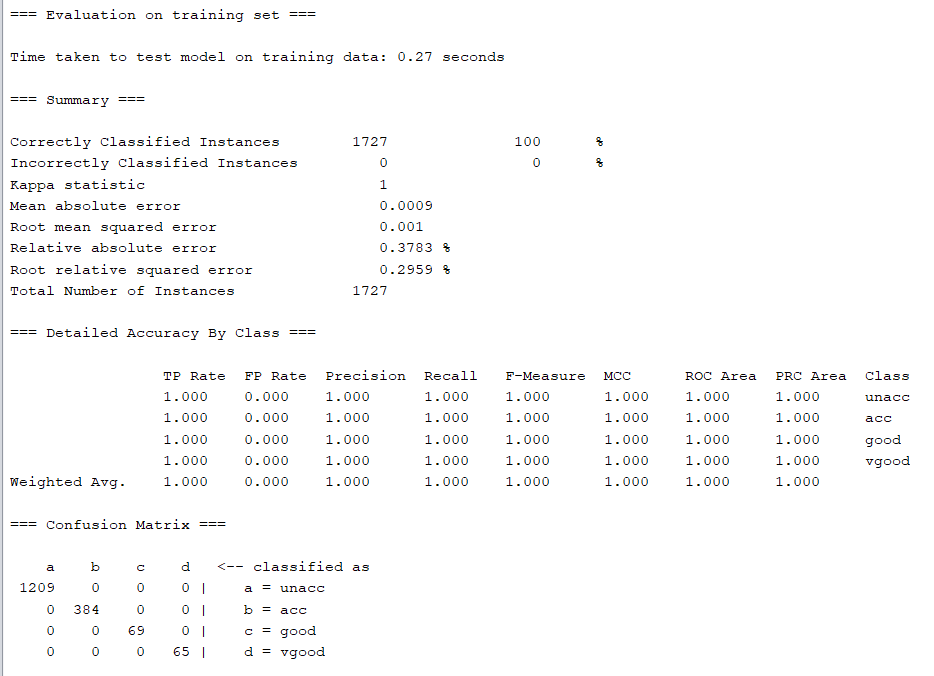
**Using Cross validation:**



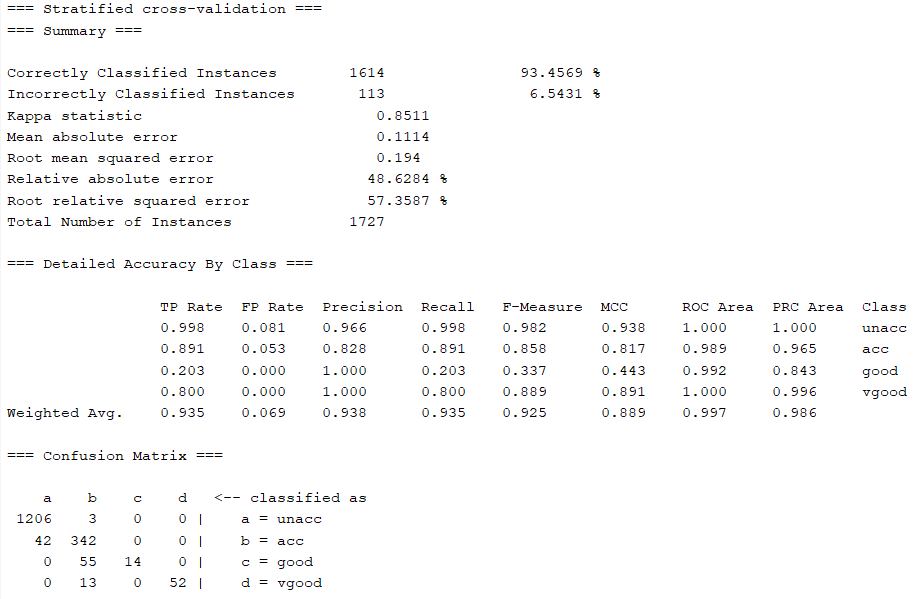
**Using Percentage split:**



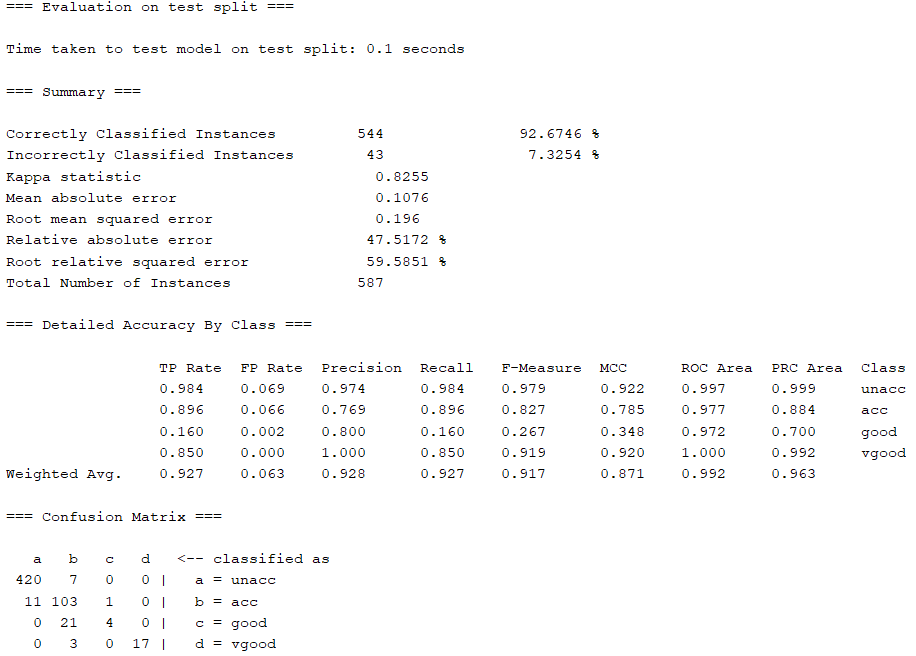
**IBK:**

**Using training set**

**Using cross validation**



**Using percentage split**



REFERENCE:

[1]P. Srikanth and D. Deverapalli, "A Critical Study of Classification Algorithms Using Diabetes Diagnosis," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, 2016, pp. 245-249, doi: 10.1109/IACC.2016.54.

[2]S. Thaiparnit, N. Chumuang and M. Ketcham, "A Comparitive Study of Clasification Liver Dysfunction with Machine Learning," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-4, doi: 10.1109/iSAI-NLP.2018.8692808.

[3]“Introduction to Data Mining” 2ndEd. by P. Tan, M. Steinbach, A. Karpatne& V. Kumar.

[4]“Data Mining: Practical Machine Learning Tools & Techniques” by I. Witten, E. Frank & M. Hall. Morgan Kaufmann (2011,2017)

[5]http://archive.ics.uci.edu/ml/datasets/Car+Evaluation