International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 10 Issue 10 Series II || October 2021 || PP 01-06

Comparative study of Decision Trees for Weather Data

K. Geetha[1], S. Venkatramana Reddy[2], B. Sarojamma[3*]
1&3: Department of Statistics, Sri Venkateswara University, Tirupati – 517 502, A.P., India.
2: Department of Physics, Sri Venkateswara University, Tirupati – 517 502, A.P., India.
*Author for Correspondence e-mail: saroja14397@gmail.com

ABSTRACT
Weather prediction is one of the most crucial demand tasks for weather forecasters since from the past. Precision plays vital role for detecting and giving warnings as natural calamities concern. In this paper, an analysis had been made by involving weather parameters like Minimum Temperature, Maximum Temperature, Precipitation, Wind speed, Visibility and Time, which consists of data from 2014 to 2019 in our country. From the last few decades, it has been seen that determining techniques have achieved good performance with their accuracy by analysis. This paper aims to compare the performance by means of few metrics using different Decision Trees such as J48, Random Forest, Random tree, Rep tree and Hoeffding tree. The result shows that Random Forest had a good level of accuracy than other algorithms.

KEYWORDS: J48, Random forest, Random tree, Rep tree, Hoeffding tree, WEKA.

Date of Submission: 07-10-2021 Date of Acceptance: 21-10-2021

I. INTRODUCTION

Weather Forecasting is one of the greatest difficulty faced by Meterology department. There is relation between Temperature, Precipitation, Windspeed, Visibility and Time. Wind speed plays major role now a days, Wind Energy is Renewable, sustainable and free. Wind Energy is produced with a low cost by Wind Turbines.

Md.Naral Amin et al[1] published an article Comparison of different classification techniques using WEKA for Haematological dataset. They took data samples of 600 and analysed. Data set consists of 298 samples for a given CBC test. White blood cells, Red Blood cells, Hemoglobin Features using Haematorists Normal values of Male and Females separately. Mean Cellular Volume, Mean Cellular Hemoglobin, Mean cellular Hemoglobin concentration, Platelet count, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils. The classifiers used for study is Decision Tree(J48), Naïve Bayes and Multilayer Neural Network. The results says that J48 Decision tree was best among 3 models taken using Kappa Statistic.

The Relationship between Wind Speed and Precipitation in the Pacific ITCZ was given by CARISSA E BACK et al[2]. The paper contains data of 4 years passive microwave satellite retrievals from the SSMI and TMI used to look at the relationship between daily wind speed and Precipitation. Correlation between Wind speed and Precipitation is significant. The slope of relation between Wind speed and Precipitation was increased in moister conditions. The area averaged Precipitation estimates derived from a radar at Kwajaleivi Island compared with microwave Precipitation estimates 2.50 Vector Mean Winds computed from Quick SCAT with the SSM/I-and TMI –derived Wind speeds.

Vidyulatha Pellakure et al[3] has published a paper entitled "applying Regression Techniques Environmental Data by WEKA". In this paper, they discussed Correlation, Regression and Prediction using Data mining process by WEKA tools for air pollutant data. "Machine Learning strategies for Time Series Forecasting" by Gianluca Bontempi et al[4]. They discussed about One-step Forecasting problems as supervised Learning tasks and they discussed about Multiple Step Forecasting Methods. Begum cigsar and Deniz Unal[5] was given a Research article on "Comparison of Data Mining Classification Algorithms Determining the default Risk". In their paper, they discussed Naives Bayes, Bayesian Networks, J48, Random Forest, Multilayer Perceptron and Logistic Regression i.e. 6 models for dataset using WEKA 3.9 datamining software. Chinnayan Ponnuraja et al[6] published research paper on " Performance Accuracy between Classifiers in sustain of Disease conversion for clinical trials Tuberculosis Data, DataMining Approach". For large dataset of TB data they used J48 classifier,iterative Dichofomister-3, a Multilayer Perceptron and a Naïve Bayes classifier by WEKA software. Razeet Mohd et al[7] published a paper on " Comparative study of Rainfall prediction Modelling Techniques" (A case study on Srinagar, J&K,India). For Prediction of Rainfall they used DataMining Techniques J48,RandomForest, Naives Bayes, Bayes Net, Logistic Regression, IBK, PART and Bagging for 5 attributes.

II. METHODOLOGY:

By taking atmosphere variables Windspeed, Minimum Temperature, Maximum Temperature, Precipitation, Visibility and Time for 2014 to 2019 day wise data[8], We prefer Decision trees such as J48, Random forests, Random Tree, Reptree and Hoeffding tree for comparison. The best among J48, Random forests, Random Tree, Reptree and Hoeffding tree are measures using Accuracy like Kappa statistic, RMSE, MSE and ROC Area.

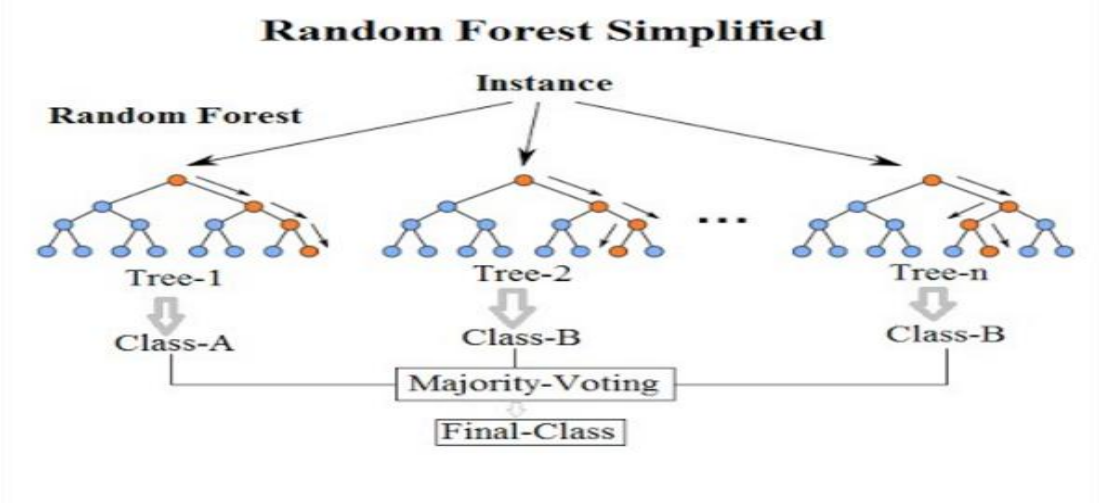
2.1 Dataset and Preprocessing: Classification methods:

J48 Algorithm:

J48 algorithm is to create a trimmed C4.5 decision tree. In this algorithm, information is split into minor susbets to base on a decision. J48 gets the results by split the information choosing an attribute. In the split strategies, stop is a subset and has place with a similar class in all the instances. Expected estimations of the classifiers utilizing decision node was developed by J48.

Random Forest:

Random Forests or Random Decision Forests are learning methods for Classification and Regression operates by constructing a multitude of Decision trees of Training time and mode or mean of Regression of individuals trees was developed as output.



Random Tree

It is similar to Random Forest. A Random Tree is built on an entire dataset using all the features or variables of interest where as Random Forest randomly selects observations or rows and specific features to build multiple Decision trees from and then averages the results. Random trees are powerful because it limits over fitting without substantially increasing error due to bias.

Reptree

Reptree Algorithm is a fast Decision tree Learner. It is also based on C4.5 Algorithm and can produce Classification or Regression trees. It builds a model using information and prunes it using reduced error pruning i.e it reduces the size of decision tree by removing decisions of the tree that do not have importance in classify.

Hoeffding Tree

Hoeffding Tree uses bound for construction and analysis of the decision tree. It uses to decide the number of instances to be run in order to achieve a certain level of confidence. It is capable of learning from bulk data stress.

Performance Measures

There are many performance measures for Classification Algorithms. In this work, we have discussed the following measures : Accuracy, Kappa Statistic, RMSE,MAE,ROC.

i) Accuracy :

It is the percentage of correctly classified modules. It is one of the most widely used Classification performance Metrics.

$$\text{Overall Accuracy} = \frac{TN+TP}{TP+FP+FN+TN}$$

where **a) True positive (TP) :** It is number of correctly classified fault prone modules. It is also called Sensitivity Measure.

$$\text{TP rate} = \frac{TP}{TP+FN}$$

b) False Positive (FP) : FP is number of non fault prone modules that is misclassified as fault prone class.

$$\text{FP rate} = \frac{FP}{FP+TN}$$

c) True Negative (TN): It is number of classified non-fault prone modules.

$$\text{TN rate} = \frac{TN}{TN+FP}$$

d) False Negative (FN): FN is number of fault prone modules that is misclassified as non fault prone class.

$$\text{FN rate} = \frac{FN}{FN+TP}$$

ii) F-measure:

It is harmonic mean of Precision and Recall.

$$\text{F-measure} = \frac{2Xprecision\ \ Xrecall}{precision\ +recall}$$

where **a) Precision :** It is number of classified fault prone modules that actually are fault prone modules.

$$\text{Precision} = \frac{TP}{TP+FP}$$

b) Recall : It is percentage of fault prone modules that are correctly classified.

$$\text{Recall} = \frac{TP}{TP+FN}$$

iii) ROC area : ROC(Receiver Operating Characteristic) is tool for comparing capabilities of Classification model. It plots true positive rates on y-axis and false positive rate on x-axis.

iv) MAE(Mean Absolute Error):

MAE is Average of difference between actual and predicted values in all test cases.

v) RMSE(Root Mean Square Error):

RMSE is the measure of difference between actually observed from the thing which is being modeled or estimates and values predicted by a model.

III. RESULTS AND DISCUSSION

In this paper, WEKA software was used for implementing Machine Learning Algorithms. The dataset is loaded into WEKA explorer. The J48, Random Forest, RandomTree, Reptree, Hoeffding Tree were implemented in WEKA. The data were transformed into WEKA Data Mining software as acceptable formats and is listed below:

Table-1

ATTRIBUTE	TYPE
Date	Date
MinTemp	Numerical
MaxTemp	Numerical
Precipitation	Numerical
Windspeed	Numerical
Visibility	Numerical

The data was in the Comma Seperated Value(CSV) in MS Excel and later it is converted in to Attribute Relation File Format (ARFF) using ARFF converter and then classified using WEKA and finally result is produced. The 10 fold Cross Validation is selected under "Test Options" for evaluation approach.

The following Metrics are used to verify the performance of model Accuracy, Kappa Statistic, MAE, RMSE and ROC area.

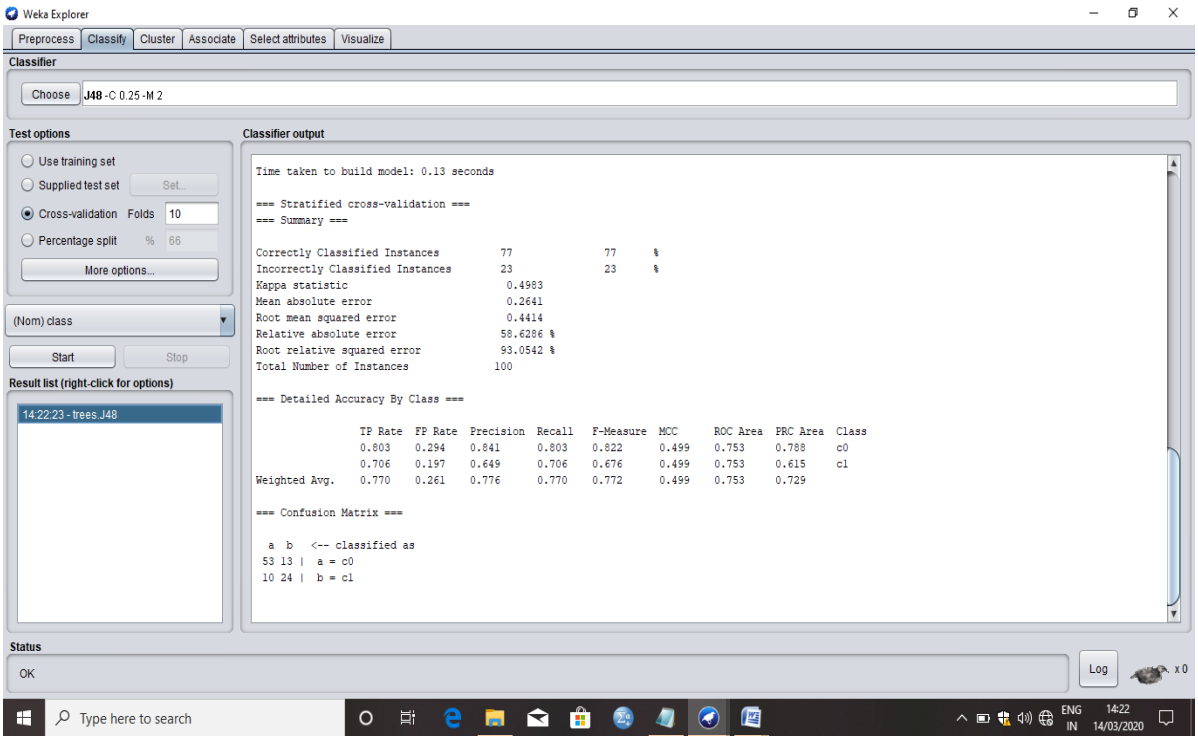


Fig 1 J48 Output

From the Fig.1, J48 output gives accuracy measure values like Kappa Statistic 0.4993, 0.2641 is Mean Absolute Error, Root Mean Square error is 0.4414, Relative Absolute error is 58.6286. Root relative squared error is 93.0542. It also give True positive Rate, False Positive Rate, Precision, F-measure, ROC Area, PRC Area for confusion matrix C_0 with 77 and C_1 with 23 classified instances.

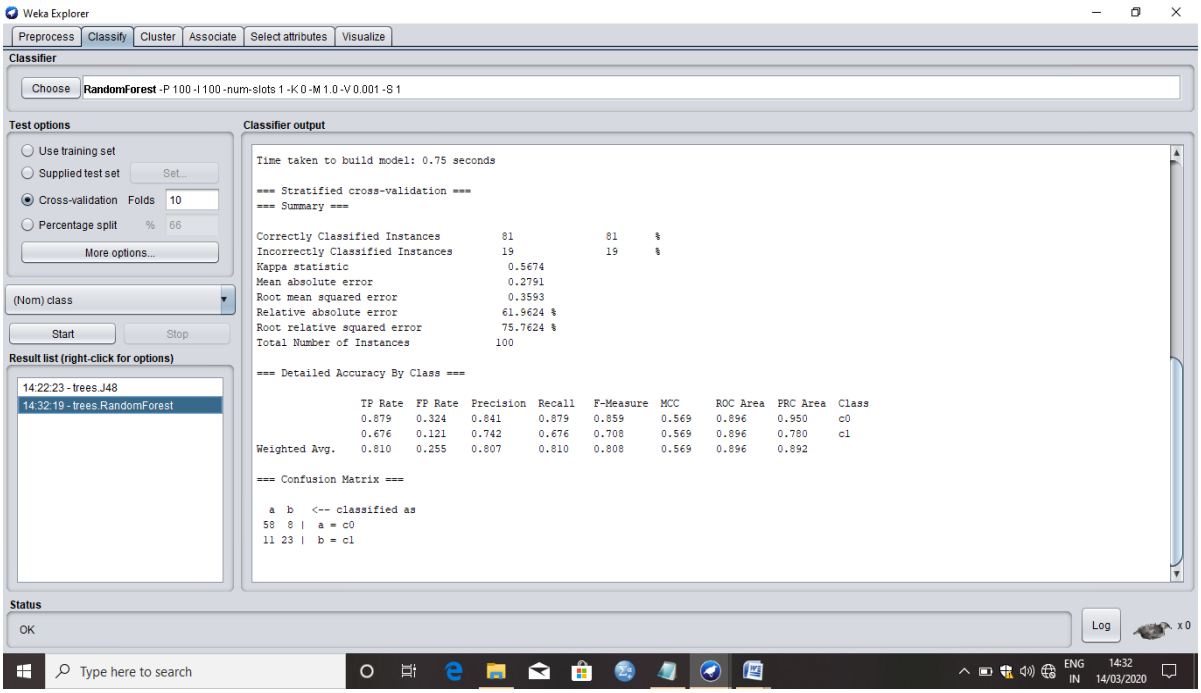


Fig 2. Random Forest Output

From the Fig.2, Random Forest gives accuracy measure values like Kappa Statistic 0.5674, 0.2791 is Mean Absolute Error, Root Mean Square error is 0.3593, Relative Absolute error is 61.9624. Root relative squared error is 75.7624. It also gives True positive Rate, False Positive Rate, Precision, F-measure, ROC Area, PRC Area for confusion matrix C_0 with 81 and C_1 with 19 classified instances.

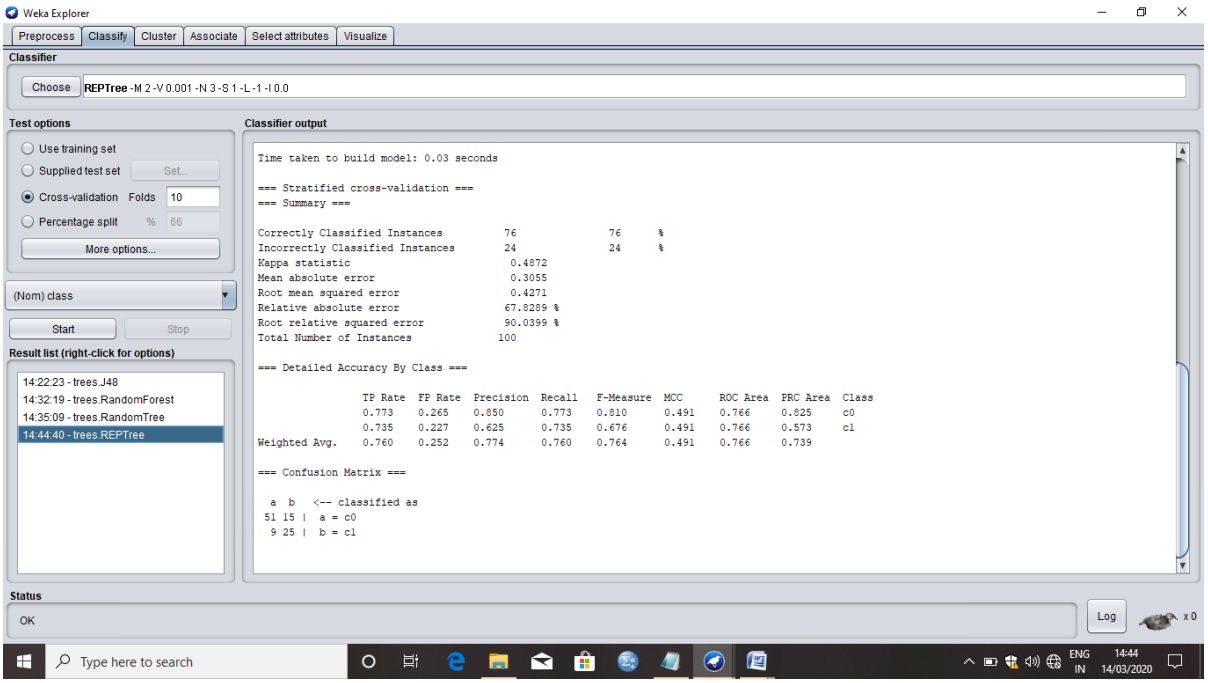


Fig 3. Reptree output

From the Fig.3 Reptree output gives accuracy measure values like Kappa Statistic 0.4872, 0.3055 is Mean Absolute Error, Root Mean Square error is 0.4271, Relative Absolute error is 67.8289. Root relative squared error is 90.0399. It also gives True positive Rate, False Positive Rate, Precision, F-measure, ROC Area, PRC Area for confusion matrix C_0 with 60 and C_1 with 40 classified instances.

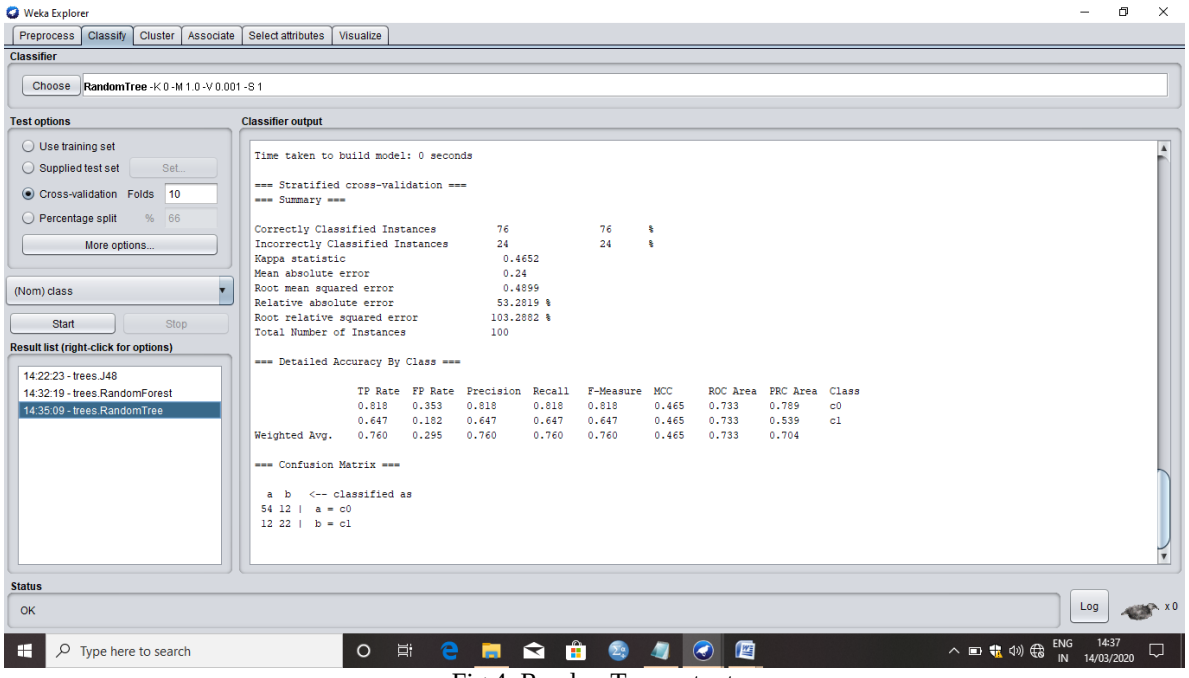


Fig 4. RandomTree output

From the Fig.4, Random tree output gives accuracy measure values like Kappa Statistic 0.4652, 0.24 is Mean Absolute Error, Root Mean Square error is 0.4899, Relative Absolute error is 53.2819. Root relative squared error is 103.2882. It also gives True positive Rate, False Positive Rate, Precision, F-measure, ROC Area, PRC Area for confusion matrix C_0 with 66 and C_1 with 34 classified instances.

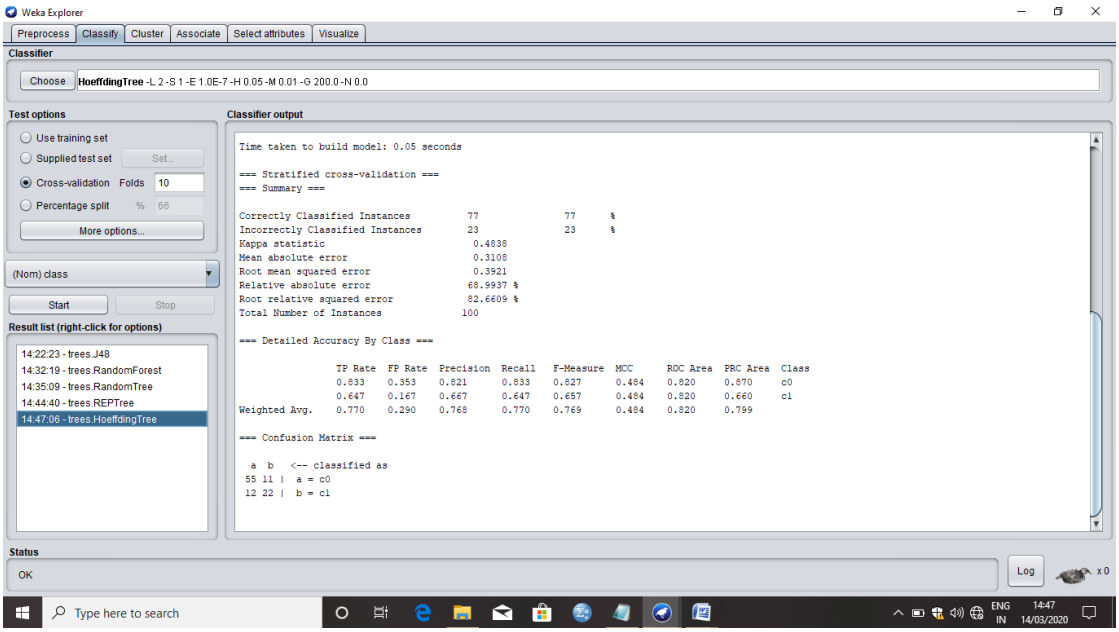


Fig 5. Hoeffding Tree Output

From the Fig.5, Hoeffding tree output gives accuracy measure values like Kappa Statistic 0.4838, 0.3108 is Mean Absolute Error, Root Mean Square error is 0.3921, Relative Absolute error is 68.9937. Root relative squared error is 82.6609. It also gave True positive Rate, False Positive Rate, Precision, F-measure, ROC Area, PRC Area for confusion matrix C_0 with 67 and C_1 with 33 classified instances.

IV. SUMMARY AND CONCLUSIONS

In this paper, we studied the performance of five different classifiers. The study is done with Weather Dataset. We got different results for each classifier. This is performed to identify the best classifier. The classifiers like J48, Random Forest, Reptree, Random Tree and Hoeffding Tree are used to identify the best relative appropriate classifiers among them. It is observed that Random Forest performs better in many ways when compared to others. In the aspect of accuracy, ROC area, Kappa Statistic and RMSE are the evidence for identifying better performance among classifiers. According to these criteria, We propose Random Forest classifier is effective and showing a good performance as listed in table-2

Table-2

Type	Accuracy	Kappa statistic	RMSE	MAE	ROC
J48	0.772	0.4983	0.4414	0.2641	0.753
Random Forest	0.808	0.5674	0.3593	0.2791	0.896
Reptree	0.764	0.4872	0.4271	0.3055	0.766
RandomTree	0.760	0.4652	0.4899	0.24	0.733
Hoeffding Tree	0.769	0.4838	0.3921	0.3108	0.820

REFERENCES

[1]. Md. Nurul Amin and Md. Ahsan Habib, Comparison of different Classification Techniques using Weka for Hematological data, American Journal of Engneering Research, 4(3) (2015)55-61.
[2]. Larissa E. Back and Christopher S. Bretherton, the relationship between Windspeed and Precipitation in the Pacific ITICZ, Journal of climate,18(20)(2005)4317-4328.
[3]. Vidyullatha Pellakrui, D.Rajeswara Rao and Lakshmi Narayan, Applying Regression Techniques on Environmental Data by WEKA, International Journal of Advances in Engineering and Research, 8(III)(2014)38-45.
[4]. Gianluca Bontempi, Souhaib Ben Taieb, and Yann-AeT Le Borgne, Machine Learning Strategies for Time series Forecasting, eBISS Business Intelligence (2012) 62-77, Springer-Verlag.
[5]. Begum Cigsar and Deniz Unal, Comparison of Data Mining Classification Algorithms in determining the default Risk, Scientific Programming, Volume 2019, https://doi.org/10.1155/2019/8706505.
[6]. Chinnaiyan Ponnuraja, Babu C Lakshmanan and Valarmathi Srinivasan, Performance Accuracy between Classifiers in Sustain of Disease Conversion for Clinical Trial Tuberculosis Data: Data Mining Approach, IOSR Journal of Dental and Medical Sciences (IOSR-JDMS), 15, Issue 4(I)(2016)105-111.
[7]. Razeef Mohd1, Muheet Ahmed Butt2 and MajidZaman Baba3 Comparative Study of Rainfall Prediction Modeling Techniques (A Case Study on Srinagar, J&K, India), Asian Journal of computer science and technology, 7(3)(2018)13-19.
[8]. WWW.Visualcrossing.com