A Comparative Analysis of Classification Algorithms on Weather Dataset using Data Mining Tool

ATHUL C R[1], Dr. Laksmi Mahesh[2]

[1]Scholar, SCMS, Cochin, Kerala, India
[2]Assistant Professor, SCMS, Cochin, Kerala, India

---***---

Abstract *-Data mining is the process of discovering insightful, interesting, and novel patterns, as well as understandable, descriptive, and predictive models from large-scale data. For analyzing the data which the software of Data mining to be used for analytical tools. There are different data mining tools are there, from these we selected one of the tools called Weka also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for analyzing the data and also predictive modeling. It involves algorithms and visualization tools that embed machine learning. The first phase called a training phase which analyzes the training data and the result of this analysis is a model that to make generalizations about how the attributes relate to the label. In the second phase called the testing phase at this time the model is applied to another, previously-unseen dataset where the labels are unknown. In a classification algorithm, the system tries to predict the label of each individual example. The Classification model tries to draw some output from the input values given for training. It will predict the class labels and categories for the new data. In our approach, we use classification algorithms like Decision Tree (J48), oneR, Naive Bayes, and IBk. Thereupon we examine the competency of these four classification algorithms. This research work aims to compare the performance of a few data mining algorithms for predicting wind direction using historical weather data of New Delhi, India, which is collected from http://www.wundergrounds.com website.*

Key Words: Weather, comparison, machine learning, Weka, Prediction

1. INTRODUCTION

Weather prediction is an important application in meteorology and has become one of the most scientifically and technologically challenging problem for meteorologists around the world. From the last few decades, the advancement and development in science and technology enable scientists to make better and precise weather prediction. The wind speed forecast in the wind energy sector is important for the following reasons like wind power system planning for unit commitment decision, load balancing decision, maintenance arrangement and energy storage capacity optimization. Wind speed is period arrangement information measured at various interims of time. Number of methods and techniques are used by the scientists to forecast rain; some of these techniques are more accurate than others.

Weather forecasting is the prediction of what the atmosphere will be like in a particular place by using technology and scientific knowledge to make weather observation. Long-established observations made at the surface of atmospheric pressure, temperature, wind direction, precipitation, humidity, etc are collected routinely from trained observers, automatic weather stations, or buoys. Now weather forecast has to rely on computer-based models that take many atmospheric factors into accounts. The input from the human is still required to select the best possible forecast model to base the forecast upon, which involves High-speed computers, wired and wireless sensors, metrological satellites, teleconnections, knowledge of model performance, etc. There is a huge amount of weather data available which is rich in information and can be used for weather prediction. Various data mining techniques are applied to the weather data to predict atmospheric parameters like temperature, wind speed, rainfall, meteorological pollution, etc. which tend to change from time to time and weather calculation varies with the geographical location along with its atmospheric parameters. Some commonly used data mining techniques for weather prediction are Decision Trees, Artificial Neural Networks (ANN), Naive Bayes Networks, Fuzzy Logic, Rule-based Techniques which includes Memory-based reasoning Techniques, and Genetic Algorithms.

2. METHODOLOGY

In data mining classification of large data set is a problem. Data mining has various techniques like classification, regression, clustering etc. This paper mainly focuses on the classification techniques having various algorithms which will help in classifying the records. The datasets contain instances or the classes and the attributes which helps in classifying the records. J48 Decision Tree, IBk, oneR and Naïve Bayes are the algorithms used for the analysis of the classification techniques. The research work mainly focuses on the comparative analysis of the classification algorithms which are Naïve Bayes, IBK, OneR and J48 on Weather forecast of wind direction dataset. The results of comparative analysis are anatomized to deduce best suited algorithm on the basis of definitiveness, execution time, correctly classified instances and incorrectly classified instances Classification is a data mining technique and is a supervised learning having broad application. Classification technique classifies each item of a set into a predefined set of classes or groups. Among all the techniques in the data mining the apex technique is classification. Dataset is being inspected by classification and each instances of the dataset is considered. Checking the instances and considered by the technique are appointed class such that there will be less error in the model. Models defining the influential data classes inlying in a particular dataset are withdrawn using classification technique. The two states of the classification include application of the algorithm to construct the model and afterwards constructed model is tested contrary to an already defined dataset to measure the performance and definitiveness(accuracy) of the model. In this research work we have analyzed Naïve Bayes, OneR, J48 and IBk algorithms on Weather forecast of wind direction dataset.

Finally, it chooses the attribute that offers rules with minimum error and constructs the final decision tree

2.1 Weka

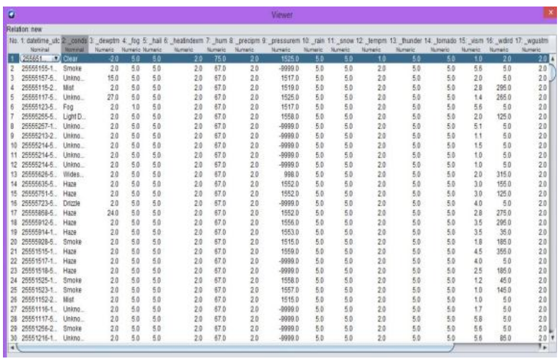
WEKA known as Waikato Environment for Knowledge Analysis which is constructed in New Zealand in the University of Waikato. This machine learning software is written in Java. WEKA is a collection of visualization tools and algorithms for the predictive modeling. Different types of data mining algorithms can be tested using different type of datasets. The techniques which are supported by the WEKA are Data Processing, Classification, Clustering, Visualization Regression and Feature Selection. There are 5 interfaces in the tool and main user interface is explorer with which we work but all other interfaces provide same functionality just as the explorer.

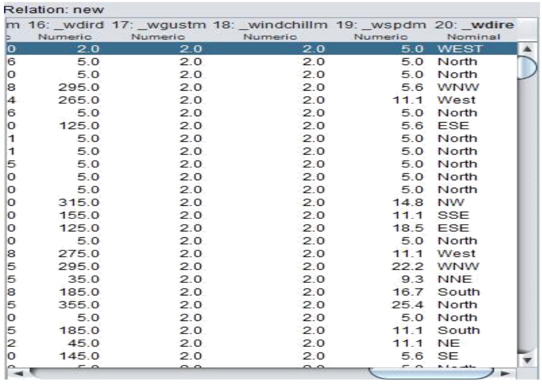
2.2 Dataset used

2.2.1. Weather dataset

In this research work we have used Weather forecast of wind direction dataset. The main focus of this research is performance and evaluation of Naïve Bayes, IBk, J48, OneR algorithms. This dataset contains 83320 instances and 20 attributes. For analyzing the performance of the classification algorithms WEKA data mining tool is used.

- Number of attributes: 20
- sum of weights: 83320
- Total Size (in Mb): 7 Mb





Relation: new

m	16: _wdird Numeric	17: _wgustm Numeric	18: _windchillm Numeric	19: _wspdm Numeric	20: _wdire Nominal
0	2.0	2.0	2.0	5.0	WEST
6	5.0	2.0	2.0	5.0	North
0	5.0	2.0	2.0	5.0	North
8	295.0	2.0	2.0	5.6	WNW
4	265.0	2.0	2.0	11.1	West
6	5.0	2.0	2.0	5.0	North
0	125.0	2.0	2.0	5.6	ESE
1	5.0	2.0	2.0	5.0	North
1	5.0	2.0	2.0	5.0	North
5	5.0	2.0	2.0	5.0	North
0	5.0	2.0	2.0	5.0	North
0	5.0	2.0	2.0	5.0	North
0	315.0	2.0	2.0	14.8	NW
0	155.0	2.0	2.0	11.1	SSE
0	125.0	2.0	2.0	18.5	ESE
0	5.0	2.0	2.0	5.0	North
8	275.0	2.0	2.0	11.1	West
5	295.0	2.0	2.0	22.2	WNW
5	35.0	2.0	2.0	9.3	NNE
8	185.0	2.0	2.0	16.7	South
5	355.0	2.0	2.0	25.4	North
0	5.0	2.0	2.0	5.0	North
5	185.0	2.0	2.0	11.1	South
2	45.0	2.0	2.0	11.1	NE
0	145.0	2.0	2.0	5.6	SE

Figure 3.1: Instances and Attributes of Weather forecast of Wind direct

3. RESULTS

This research work analyses different classification algorithms accomplishment for Weather forecast of wind direction dataset. Comparison of classifiers for Weather forecast of wind direction datasets done using criteria accuracy, correctly classified instances, incorrectly classified instances, error rate, and execution time to analyses the performance of the classification algorithms and its application domain is also discussed. Models for each algorithm are constructed using two methods mainly Cross-Validation with 10 folds out of which training set uses 9 folds and 1 fold for testing and Percentage Split in which 60% of the dataset is used for the training and 40% is used for the testing and output is given according to it. Figures are shown for the comparison of the different classifiers for Weather forecast of wind direction dataset using a 10-fold cross-validation testing bed. Applications are also discussed of these classifiers in the table. According to the table and research the execution time taken by the One Rule algorithm is least with 0.4 seconds followed by Naïve Bayes with 0.68 seconds, J48 algorithm with 3.28 seconds and Instance-Based Learner took much more time for execution which is 159.3 seconds. Accuracy of both J48 and One Rule classifier have 95.07%, Instance-Based Learner with 73.61%, and naïve Bayes with 43.18%. The naive Bayes classifier algorithm of accuracies have difference in comparison to others. Hence according to the data J48 and OneR algorithms are most accurate in case of 10-fold cross-validation method.

3.1. Final test accuracy of cross-validation

Analysis for Weather dataset of Wind Direction using percentage Cross-validation is done and this is as below. The final value represents the average accuracy of all the training and testing steps. This value is significant because it represents the accuracy or in other words rate of successful classification by the resulting model. The following figures from Figure 5.1 to 5.5 represents the final test accuracy of the various classification algorithms

3.1.1 J48

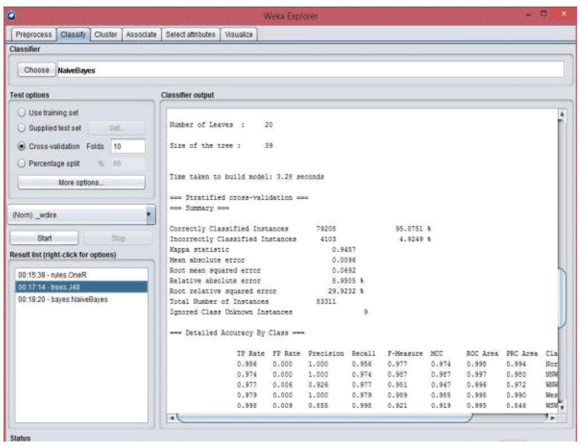


Figure 3.2: Ttest Accuracy of cross-validation of J48

3.1.2. Naivebayes

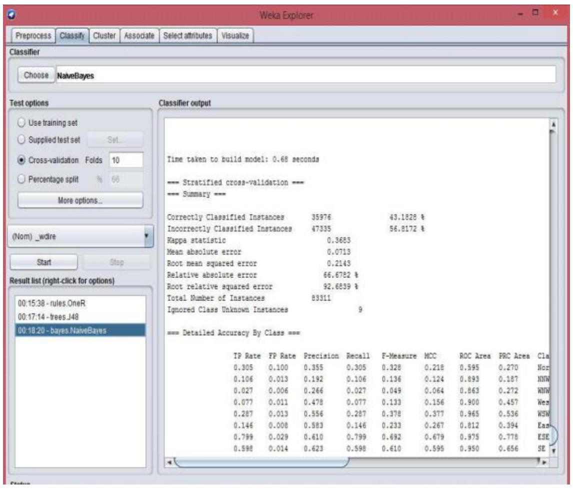


Figure 3.3: Accuracy of cross-validation Naive Bayes.

3.1.3. Onerule

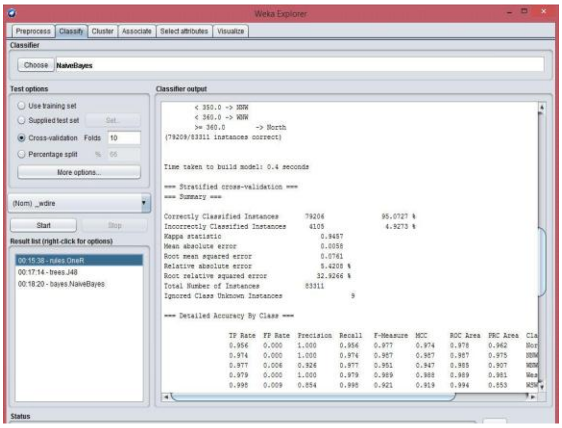


Figure 3.4: Accuracy of cross-validation OneR.

3.1.4. Instance-based learner

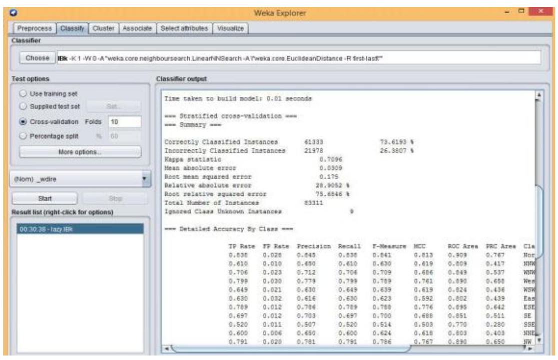


Figure 3.5: Ttest Accuracy of cross-validation of IBK.

3.1.5. Tabulated summary of results of cross-validation

For Weather dataset of Wind Direction				
Classifier	**Naïve Bayes**	**OneR**	**IBK**	**J48**
Testing Bed	Cross Validation	Cross Validation	Cross Validation	Cross validation
Applications	Text classification, Spam filtering, Online Application, Hybrid recommender system	Text document categorization by term association[32]	Random projection in dimensionality reduction: Applications to image and text data[17]	Emotion recognition, Verbal column pathologies.
Execution Time	0.68 seconds	0.4 seconds	0.01 seconds	3.28 seconds
Accuracy	43.18%	95.07%	73.61%	95.07%

Table 1: Comparison of classifiers for Weather dataset of Wind Direction using cross-validation testing bed

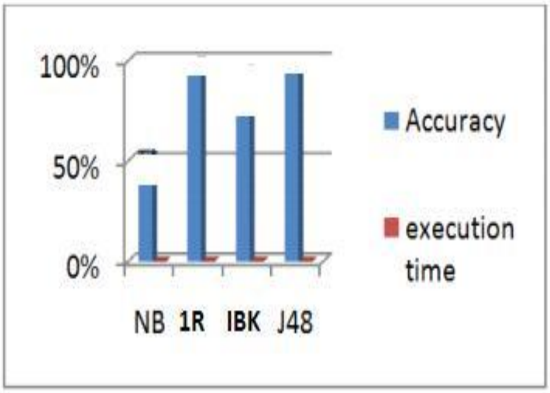


Figure 3.6: Graphical representation of different algorithms accuracy and execution time using cross-validation method.

In the graph the abbreviation NB stands for Naïve Bayes, 1R for OneRule, IBK for Instance-Based Learner. The number of correctly classified instances in Naïve Bayes is 35976, One Rule with 79206, j48 with 79208 and instance-Based Learner with 61333. The incorrectly classified instances by Naïve Bayes is 47335, OneRule with 4105, j48 with 4103 and Instance-Based Learner with 21978.

3.1.6 Final test accuracy of percentage split

Now analysis for Weather dataset of Wind Direction using percentage split method is done. The final value represents the average accuracy of all the training and testing steps. This value is significant because it represents the accuracy or in other words rate of successful classification by the resulting model. The following figures from Figure 5.6 to 5.9 represents the final test accuracy of the various classification algorithms below.

3.2.1. Naivebayes

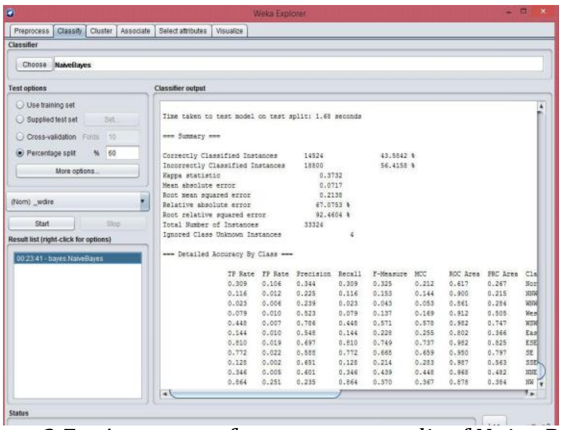


Figure 3.7: Accuracy of percentage split of Naive Bayes.

3.2.2. J48

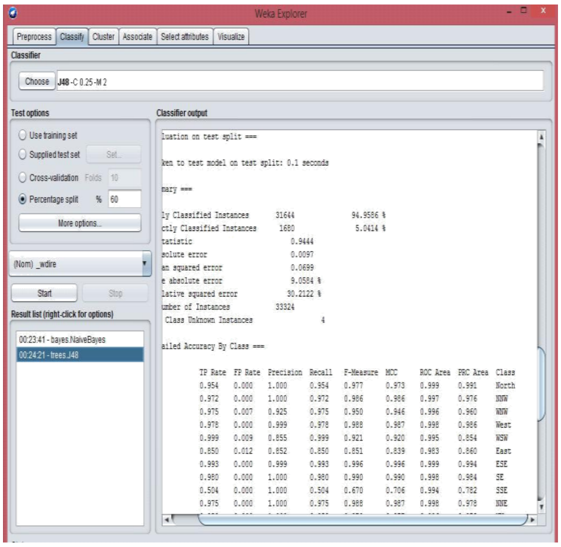


Figure 3.8: Accuracy of percentage split of j48.

3.2.3. ONERULE

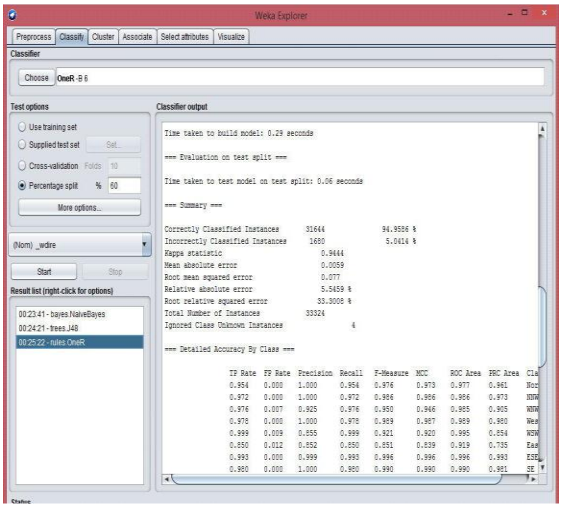


Figure 3.9: Test Accuracy of percentage split of OneR..

3.2.4. Instance-based learner

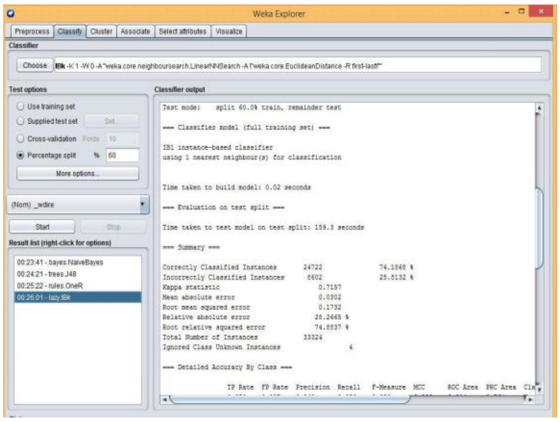


Figure 3.10:Test Accuracy of percentage split of IBK

3.2.5. Tabulated summary of results of percentage split

Weather dataset of Wind direction				
Classifier	Naïve Bayes	OneR	IBK	J48
Testing Bed	Percentage Split	Percentage Split	Percentage Split	Percentage Split
Execution Time	1.68 seconds	0.06 seconds	159.3 seconds	0.01 seconds
Accuracy	43.58%	94.95%	74.18%	94.95%

Table 2: Comparison of classifiers for Weather dataset of Wind Direction using cross-validation testing bed

Comparison of classifiers for weather dataset of wind direction using percentage split method According to this test method that is percentage split it can be concluded that Naïve Bayes took 1.68s, Instance-Based Learner took 159.3s One Rule took 0.06s seconds and, j48 took 0.01s for execution. Accuracy of the J48 algorithm comes out to be 94.95% while same as that of One Rule, Naïve Bayes with 43.58% accurate and, Instance- Based Learner with 74.18% accurate. The number of correctly classified instances in Naïve Bayes is 14524, Instance-Based Learner with 24722, One Rule with 31644 and J48 with 31644. Number of incorrectly classified instances in Naïve Bayes is 18800, Instance-Based Learner with 5602, OneRule with 1680 and J48with 1680.

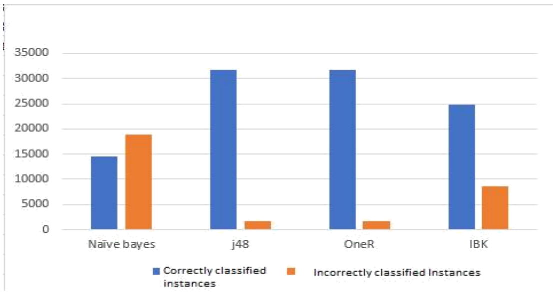


Figure 5.12: correctly and incorrectly classified instances in case of Percentage Split

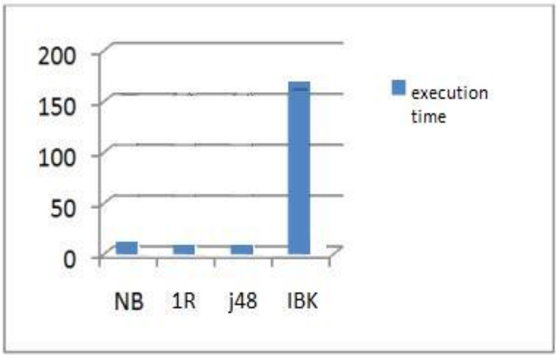


Figure 5.10: Graphical representation of different algorithms execution time in percentage split

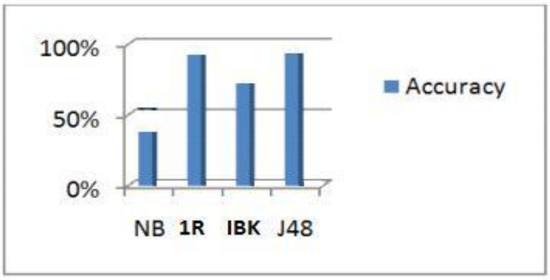


Figure 5.11: Graphical representation of different algorithms accuracy in percentage split

4. DISCUSSIONS

Graphical representation of different algorithms accuracy in percentage split method. The abbreviations in the chart stands for Naïve Bayes, Instance-Based Learner, One Rule, j48.Graphical representation of correctly and incorrectly classified instances by the classifiers are:

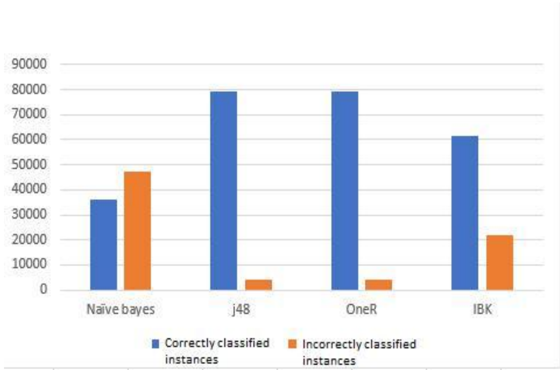


Figure 5.13: correctly and incorrectly classified instances in case of Cross-Validation

From the graphs it is analyzed that there is no such difference between the performance of the classification algorithms they have significant performances for the Weather dataset of Wind direction but on the basis of graph analysis J48 and One Rule classifiers is most accurate when using cross-validation method and percentage split and also Naive Bayes classifier has less accurate while compared to other classifier.

5. CONCLUSION

Comparison and investigations of the accomplishment of various classification algorithms is done using different criteria which are accuracy, execution time, correctly classified instances, incorrectly classified instances and error rate. In this work we carried out an experimental work to compare popular classification algorithms for wind direction prediction using various performance measures over weather data of New Delhi, India. The different measuring attributes play a pivotal role in giving precise weather forecast of wind prediction. According to the result evaluation it can be concluded that J48 and ONERULE are both have most accurate with 95.07% when 10 folds cross-validation method is applied for weather forecast of wind prediction dataset and for Percentage Split method J48, and ONERULE algorithms both have most accurate with 95.95% accuracy. In our case, J48, and ONERULE approach proves to be an efficient and acceptable method for wind direction prediction. The level of accuracy and prediction highly depends on the data being used as input for classification and prediction. Every algorithm has its advantages and limitations; it is difficult to choose the best algorithm. The prediction accuracy of the model can be increased by developing a hybrid prediction model where multiple machine learning algorithms are put to work together. For our weather dataset, it was concluded after analyzing various models of supervised learning that the J48, and OneR classification algorithms has appreciable level of accuracy and acceptance.

6. REFERENCES

1) Olaiya, Folorunsho and Adesesan Barnabas Adeyemo, "Application of data mining techniques in weather prediction and climate change studies", International Journal of Information Engineering and Electronic Business, Vol. 4, No. 1, pp. 51, 2012.

2) Nikam, B. Valmik and B. B. Meshram, "Modeling rainfall prediction using data mining method: A Bayesian approach", Computational Intelligence, Modelling and Simulation (CIMSim), 2013 Fifth International Conference on. IEEE, 2013.

3) A. Geetha, and G. M. Nasira, "Data mining for meteorological applications: Decision trees for modeling rainfall prediction", Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE, 2014.

4) Ahmed, Bilal, “Predictive capacity of meteorological data: Will it rain tomorrow?”, Science Information Conference (SAI), 2015, IEEE, 2015.

5) D. Hand, H. Mannila and P. Smyth, “Principles of data mining”, MIT, 2001.

6) R. Mohammad, M. Butt Ahmed and M. Baba Zaman, “Predictive Analytics: An Application Perspective”, International Journal of Computer Engineering and Applications, Vol. 9, No. 8, Aug. 2017

7) M. Kannan, S. Prabhakaran and P. Ramachandran,“Rainfall forecastingusing’ data mining technique”, 2010

8) Bingham, Ella; and Mannila, Heikki;"Random projection in Dimensionality reduction: applications to image and text data", in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001

9) R. Mohammad, M. Butt Ahmed and M. Baba Zaman, “Tools for Predictive Analytics: An Overview”, International Journal of Scientific Research Engineering & Technology (IJSRET), Vol. 6, No. 7, pp. 748-750, 2017.

10) Dutta, Pinky Saikia and Hitesh Tahbilder, “Prediction of rainfall using data mining technique over Assam”, IJCSE, Vol. 5, No. 2, 2014, pp. 85-90.

11) P.Langley, W.Iba and K. Thompson,“ An analysis of Bayesian

12) L.Breiman, “Random forests”, Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001

13) Dubey and D. Akash, “Artificial neural network models for rainfall prediction in Pondicherry”, International Journal of Computer Applications, Vol. 120, No. 3, 2015