Machine Learning Approach for Weather Forecasting

Yigezu Agonafir[1], University of Gondar

[1]Department of Information Technology, University of Gondar, Gondar, Ethiopia

[1]Email: yigezua@gmail.com

ABSTRACT

Weather predictive models are used to find potentially valuable patterns in the data, or to predict the outcome of any event. The choice and use of predictive technique to use becomes even harder, since no technique outperforms all others over a large set of problems. The main objective of this study is appraising the potential applicability of machine learning technology to predict rainfall using ensemble and single algorithms. The other contributions are improvements of a rule extraction technique, resulting in increased comprehensibility and more accurate result by ensemble machine learning. On the other hand, in this study the researcher used hybrid machine learning methodology. Also, the researcher used 9,543 instances with 8 selected on WEKA 3.8. In this paper, J48, PART, MLP and IBK algorithms are used with ensemble method. Furthermore, PART algorithm and J48 decision tree demonstrated on 10 fold cross validation method given best result rather than MLP and IBK. Ensemble PART algorithm & J48 Decision tree with selected attributes produced 95.46 % and 95.44% prediction accuracy respectively, on WEKA experimenter for one day advance prediction.

However, when we used the ensemble method, boosting ensemble was given better result rather than bagging and staking. Ensemble PART algorithm for one month advance prediction using selected attributes produced 95.35 % and 97.12% accuracy, on WEKA experimenter and explorer respectively. The researcher found temperature, humidity, wind speed, sunshine, month and year as the major variables to predict rainfall. Beyond this it's possible to extend the development of the model to a longer forecast such as ten days ahead and one year ahead. The researcher recommended other researcher's to predict other atmospheric variable like wind speed, humidity, and temperature. Furthermore, the researcher also recommended other researcher's to include association rule discovery to found strong internal relationship among meteorological variables.

Key Words: machine learning, weather forecasting, datamining

1. INTRODUCTION

Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. In addition to this weather information is important for human beings day to day activities, like aviation service, food production plan and water resource management. Specifically the occurrence of prolonged dry period or heavy rain at the development of crop may lead to reduce crop yield causes for drought and flood [1]. Rainfall prediction helps in proper aviation service; agricultural planning to make necessary arrangements for procurement, transport and distribution of food grains, if there is below normal rainfall [2].Thus to take preventive measure machine learning has tremendous advantage by forecasting the future weather condition. Although, meteorological departments obtain rainfall data using various methods like ground observation, satellite observation, observations taken from ships, aircraft and radar [3]. Meteorological department arrange these obtained dataset in form of sheets, graphs, and charts. However, Ethiopian National Meteorological Agency did not properly process and use the data. So, this research aims to find pattern among the data set and used for prediction purpose by using machine learning and by using conceptual machine learning. On the other hand, the Ethiopian Meteorological Agency uses a variety of information sources abroad, such as the US and UK, to prepare a seasonal forecast in Ethiopia.

However, as Teksande and Mohod [4], explored that simple long term summary of weather is not still a true picture of weather forecasting. Weather prediction performance can be improved by analysis of daily, monthly, and yearly meteorological data patterns using machine learning techniques.

Hence, the main aim of this research is to investigate the potential applicability of ensemble and single machine learning methods for rainfall prediction that supports metrology experts, planners and policy makers in Ethiopia.

1.1 STATEMENT OF THE PROBLEMS

Weather forecasting is using machine learning approach plays a tremendous advantage in today's information age. The underlying problems that necessitate this research are rainfall forecasting and early warning systems are the most important service for agricultural country like Ethiopia. In our country meteorological data are periodically gathered by Ethiopian Meteorology Agency. However, due to lack of appropriate data analysis tool, the available data are not practically used to alleviate the problems faced by planners, policy makers and decision makers. As the Guardian in 2015 reported that, Ethiopia is one of African developing country with nearly double digit economic growth for the last decade. However, the failed rain had devastating consequences for food supplies for its 96 million people [5]. "The Belg rain was much worse than the National Meteorology Agency prediction at the beginning of the year. As a result, food insecurity and malnutrition increased" [5]. Thus, supporting meteorology experts by developing an effective model can provide timely and accurate rainfall prediction has tremendous advantage.

On the other hand, useful knowledge can play important role in understanding the weather variability and weather prediction. However, effective use of analysis tools can provide the best possible information about the future [6]. In such situations, technologies like machine learning will be helpful in discovering hidden patterns for risk reduction rather than asking emergency assistance. Machine learning technologies are useful to extract rules and predict future event that can support private and government enterprise in planning their day- today operations and to enhance their quality of life. The primary goal of this thesis is to assess how a predictive model will be built by using machine learning technique and machine learning techniques.

Taffese [4] conducted a research on weather forecasting by using artificial neural network. The researcher used 3 year's data set and explored the application of artificial neural network for temperature prediction one day ahead using single algorithm.
Ephrem [8] conducted on Application of data mining for weather forecasting by using 15 year's data set to predict rainfall after one day ahead. He used CRISP methodology and three single machine learning algorithm i.e. PART, J48, MLP. Beyond this, Taksande & Mohod proved that the combination of different algorithms on weather data, gives better performance higher than 90% prediction accuracy with several population size and crossover probability [4]. Dhanya & Kumar [9] conducted a research on "machine learning for evolution of association rules for droughts and floods in India using climate inputs". El-Halees & Kohail conducted a case study on meteorological data analysis and its implementation [6].The researchers applied outlier analysis, clustering, prediction, classification and association rules mining techniques. As many researcher's suggested that ensemble method can solve wide range of rainfall prediction problem.

However, weather condition varies from country to country by location to the equator. So, the developed model in other countries is not directly applied for our country without validation. Beyond this, in our country those previous studies were conducted by using single algorithm pattern mining. So, using single algorithm cannot improve the performance of the model to solve a wide range of rainfall prediction; or single algorithm cannot provide robust model. This implies that there is a gap for further research that deals about the prediction of rainfall by using ensemble method. In addition, for early warning and to improve the quality of life extending prediction duration a month advance plays significant role. So, to take safety measure and to solve wide range of rainfall prediction problems using ensemble method has tremendous advantage.

Research Questions

In order to accomplish the purpose of the research, the following guiding questions or lines of inquiries were listed as follows:

1. Which machine learning Algorithm is best to predict rainfall by using historical weather data set?
2. What are the major attributes to consider in applying machine learning technology for rainfall prediction?
3. Which ensemble method is robust in this study?

1.2 OBJECTIVES OF THE RESEARCH

1.2.1 GENERAL OBJECTIVE

The general objective of this research is to investigate the potential applicability of machine learning for rainfall forecasting by using ensemble methods.

1.2.2 SPECIFIC OBJECTIVE

In order to achieve the above general objective, the research work will carry out the following specific objectives:

- To build the model using ensemble methods and single algorithms.
- To evaluate the model,

- To select the best algorithms for longer duration rainfall forecasting.

1.3 SCOPE OF THE STUDY

The aim of this research is to appraise the potential applicability of machine learning and machine learning for rainfall forecasting. Thus, the scope of this research is strictly limited to appraising the potential applicability of machine learning for one day ahead rainfall prediction to support meteorology experts in Ethiopia.

1.3.1 LIMITATION OF THE STUDY

This study is unable to include flood forecasting based on the predicted rainfall level. In this study, the rainfall predictors like cloud cover and sea level pressure are not available in ENMA database. So, those attributes are not incorporated. If those attributes are incorporated the performance of the model is increased and more robust. Also, in this study the researcher could not cluster based on similarity; and apply association rule discovery techniques to investigate the internal association exists among the different variables.

1.4 SIGNIFICANCE OF THE STUDY

This research will contribute to an understanding of machine learning technology for rainfall forecasting and identify the best machine learning algorithm to predict rainfall. Moreover, accurate prediction of extreme rainfall or dry events can significantly aid in policy making and also in designing an effective risk management system. The finding of this study will help in extracting hidden predictive information or future trends and behaviors from the national meteorology agency database. It allows decision makers to make proactive knowledge driven decisions. Also the study will provide valuable help in developing models and identifying important variables to predict rainfall. On the other hand, the study will be used for the government authority's planners to take safety measure based on the status of rainfall. Beyond this, the proposed models can be used to predict daily and monthly rainfall. So, aviation and agricultural sector could be benefited from these predictions especially that income for many peoples in Ethiopia depend on agriculture.

1.5 LITERATURE REVIEW

Many researchers have tried to use machine learning technologies in areas related to weather prediction in general and particularly for rainfall prediction. Latha et al. [7] proposed a novel method to develop service oriented architecture for a weather information system and forecast weather by using data mining techniques. The researchers were used Support vector regression for atmospheric temperature prediction. The researchers conclude that weather forecaster web service can develop in .NET for getting the weather data status from the web.

Sethi et al. [8] explored that exploiting data mining techniques for rainfall prediction. The researchers are using multiple linear regression (MLR) technique for the early prediction of rainfall. Regression analysis includes parametric methods such as linear and logistic regression. Finally, the researchers were proposed a method for rainfall prediction after analysis of rainfall dataset which is derived from some data mining techniques like firstly apply correlation analysis then regression analysis.

Atole et.al [9] explored that for the better crop productivity rainfall prediction is necessary and required. Data mining technique is used to calculate or analyze the rainfall prediction. In statistical modeling clustering algorithm are used for cluster like K-means and k-medoid. Also the researchers used Multi Linear Regression (MLR) and Multi Polynomial Regression (MPR).

The researchers conclude that the prediction of rainfall is very complex topic and can't be predicted easily. The researchers predict rainfall using seven year data that include minimum temperature; humidity, wind speed. Also the researcher develops a model that can predict the event this day is sunny or rainy. The researchers prove that Multi variables polynomial regression (MPR) & Multi linear regression technique is reliable for prediction.

Sumi et al. [10] investigated that optimal data-driven machine learning methods for forecasting an average daily and monthly rainfall. This comparative study is conducted discusses on three aspects modeling inputs, modeling methods and pre-processing techniques. The comparison between linear correlation analysis and average mutual information is made to find an optimal input technique. For the modeling of the rainfall, a novel hybrid multi-model method is proposed and compared with its constituent models. The models include artificial neural network, multivariate adaptive regression, the *k*-nearest neighbor, and radial basis support vector regression. Each of these methods is applied to model the daily and monthly rainfall, coupled with a preprocessing technique including moving average and principal component analysis.

Moreover, the researchers were exploring the use of several machine learning methods and particularly suggests to employ a hybrid multi-model method coupled with model ranking and selection for improving two rainfall forecasting problems. The rainfall series include the daily and monthly rainfall. For reasonable evaluation of the performance of the hybrid

method, its constituent models (ANN, *k*-NN, MARS and SVR) are separately constructed and used for the purpose of comparison. Finally, the researchers conclude that PCA can be assessed as a more effective and efficient method among the two input techniques due to the simplicity in computation and superior capability of forecasting. The researcher's experimental result showed that PCA improves the hybrid method performance. The hybrid method produces more accurate forecast than the single models for the daily rainfall series. Among the single models, SVR performs better and produced a better forecast for monthly rainfall series.

Kanth et al. [11] investigated that Machine Learning Algorithms are better than the existing techniques / methodologies/traditional statistical methods. Hence the development of the new Hybrid SVM (Support vector machines) model is required for effective weather prediction by analyzing the given weather data and to recognize the patterns existing in it. SVM comes under the set of supervised learning methods for classifications & regression. It will be yielding good results in predicting the weather than the existing machine learning programming techniques. The researchers conclude that the correlation between mean precipitation and mean cloud cover its value is 0.754384589. Likewise, there is a correlation between the average temperature and mean vapor pressure is 0.780409787. There is a correlation between mean temperature and mean potential evapotranspiration is 0.686833118. There is a negative correlation between Mean ground frost frequency and mean wet day frequency and its value is -0.840915846.

The mean average temperature has correlation with all the parameters, namely mean cloud cover, temperature range, ground frost frequency, precipitation, vapor pressure, wet day frequency, potential evapotranspiration and reference crop evapotranspiration. It has the highest correlation with reference crop evapotranspiration and negative correlation with ground frost frequency.

Ephrem [9] investigated that the application of data mining technology for weather forecasting. He used 15 years data set from (2000-2014) with CRISP data mining methodology. The researcher applied three modeling single algorithms using WEKA 3.6 like MLP, PART and J48 decision tree ,the researcher achieved (83.88, 86.03, 86.65) prediction accuracy respectively. Finally, the researcher conclude that the best performance achieved with J48 decision tree with pruned techniques i.e. 86.65 prediction accuracy. Also the researcher conclude that month, temperature, sun hours and relative humidity are the most determinant factors for rainfall prediction.

1.6 METHODOLOGY

This section provides information regarding the KDD process and the machine learning approach applied in this study are explored.

1.6.1 STUDY AREA

1.6.2 OVERVIEW OF ETHIOPIAN NATIONAL METROLOGY AGENCY

As Ethiopian Meteorology Agency revealed that at the end of the 19th century missionaries entered in Ethiopia were taking meteorological observation in Addis Ababa.

In addition to this, meteorological stations were established in 1890 and 1986 at Adamitulu and Gambela respectively. However, due to the growing demands of meteorological information for safe operations of the air transport and day to day activities, which handles meteorological activity was also established in the early fifties under the civil aviation department. Finally, economic and social sectors began to realize the importance of meteorological services then metrological station was changed with the responsibility of giving assistance to non-aviation activities. NMA had its present status. As Ethiopian National Meteorological Agency revealed that government of Ethiopia officially established the National Meteorological Services Agency in December 31, 1980 under proclamation no 201 0f 1980. Now a day's the Ethiopian National Meteorological Agency using bullets from US and UK. Also NMA using dynamic and standard statistical method for weather forecasting. The National Meteorological Agency used Wolf model, it works based on the mathematical equation developed by meteorology expert. However, they didn't use intelligent machine learning tools. Hence, to improve prediction performance intelligent machine learning tools and data mining plays tremendous advantage.

1.7 HYBRID METHODOLOGY

In this study Hybrid machine learning methodology used, because the hybrid machine learning methodology has five iterative steps but CRISP has only three feedback mechanism. In addition SEMMA and KDD have no feedback mechanism. Moreover, Hybrid machine learning methodology is more general research-oriented description of the steps and introducing 'using discovered knowledge' step instead of the 'deployment' in CRISP. As a result to encourage knowledge discovery process for this research domain hybrid methodology used to predict rainfall.

In addition, CRISP-DM and SEMMA mostly company oriented especially SEMMA that is used by SAS enterprise miner and integrate with their software. However, CRISP-DM is more complete as compare to SEMMA. Furthermore, all these

process models guides and helps the people and experts to know that how they can apply machine learning into practical scenarios. However, hybrid methodology is the extension of CRISP-DM so, hybrid machine learning methodology is more complete and better than others and this methodology is used for this research domain.

1.8 DATA PREPARATION AND MODEL BUILDING

1.8.1 DATA SOURCE AND DATA COLLECTION

To achieve this study, we use historical data records of thirty years period (1985-2015) recorded from Addis Ababa Bole station. The obtained record include the daily average relative humidity, average daily minimum and maximum temperature, wind speed, elevation, sunshine, humidity and rainfall observation.

1.8.1.1 DATA FORMATTING

At this step the researcher changes the data into a format which was suitable for the machine learning tool or algorithms. The preprocessing of the data performed in WEKA 3.8 and SPSS and MSExcel. The final dataset was also in MS-Excel format; however, the selected tool WEKA 3.8 doesn't accept the data in Excel format. So, the researcher first convert the data in comma delimited (CSV) text file in ARFF (Attribute Relation File Format). Comma delimited applied for a list of records where the items are separated by commas, whereas ARFF is an extension of a file format that the WEKA software can read.

1.8.1.2 ATTRIBUTE SELECTION

In machine learning technology not all attributes are relevant. Selection of relevant attribute is necessary for machine learning, among all original attributes. Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid over fitting and improve model performance and to provide faster and more cost-effective models Beniwal [12]. As a result all attributes are not necessary for the experiment, Due to this reason in this thesis information gain ranker is used for attribute selection.

1.8.1.3 ATTRIBUTES RANK WITH INFORMATION GAIN

The effect of the attributes on the model performance was investigated. The full training set containing a total of 9553 instances and 10 attributes. The attributes are selected by using information gain ranker.

Table 1: Attribute Rank on all input Data using information gain ranker

Relative Importance	Attribute
0.205	Year
0.13345	month
0.10529	Maxtemp
0.10248	Mintemp
0.02567	rhumidity
0.00846	sunshine
0.00685	Wind speed
0	day
0	Elevation

Based on information gain ranker result the selected attributes for this thesis are listed in the following table

Table 2: Selected attribute from the NMA database.

№	Field Name	Data Type	Description
01	Year	date	Year considered
02	Month	date	Month considered
03	rainfall	nominal	Rain fall
04	Relative humidity	Numeric	Relative humidity
05	Sunshine	Numeric	Sunshine
06	Temp Max	Numeric	Max temp
07	Temp Min	Numeric	Min temp
08	Wind speed	Numeric	Wind speed

1.8.1.4 PREPARATION OF THE DATA

Data about Meteorology in Ethiopia provided by Ethiopian National Meteorology Service Agency. This research utilized Ethiopian meteorological Agency databases for meteorological data analysis purpose, i.e. the data source contained raw

data about metrological variables. Assembling and cleaning the data were the initial task in the metrological data analysis with machine learning techniques.

In this step, the researcher concerns which data will be used as input for DM methods in the subsequent step. It involves sampling, running and significance tests and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values are performed. In order to, improve our classification accuracy we must first analyze the raw data. Before running any classification algorithms on the data, the data must first be cleaned and transformed in what is called a pre-processing stage. During this pre-processing stage, several processes take place, including evaluating missing values, eliminating noisy data such as outliers, discretization, aggregation, and balancing unbalanced data was performed.

1.8.1.5 DATA CLEANING

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data Onwubolu [13].In this study detecting and correcting errors in the data are done. In order to provide access to accurate and consistent data, missing value replacement, outlier detection, data reduction and data transformation are performed.

1.8.1.6 MISSING VALUES TREATMENT

The treatment of missing values is an important task in KDD process. Especially, while the dataset contains a large amount of missing data, the treatment of missing data can improve the quality of KDD dramatically. Selection of missing values using series mean method is highly depends on given data set, structure of attributes and missing data mechanism. Unfortunately missing data mechanism is usually unknown Kaiser [14].There are several strategies that could be used to handle missing values. Instances with missing values could be removed, missing values can be replaced with a certain value not present data can be replaced with a value that is representative for the data set. However all strategies have their own flaws and which one to choose has to be decided from case to case. A common method for continuous attributes is to replace the missing value with the mean value of instances with no missing values. In the same way nominal missing values can be replaced with the serious mean value (the most common class) and also in this study, this method was applied.

Table 3: Univariate Statistics of metrological dataset

Attribute	Total number	Mean	Std. Deviation	Missing	
				Count	Percent
Year	9553	2002.47	7.509	1	.0
month	9553	6.50	3.453	1	.0
date	9553	15.81	8.843	1	.0
Rainfall	**9446**	**3.036**	**7.0350**	**108**	**1.1**
Max temp	9300	23.970	2.8685	254	2.7
Min temp	9400	9.996	2.7404	154	1.6
humidity	9349	60.634	20.9179	205	2.1
sunshine	9058	7.441	3.2784	496	5.2
Wind speed	8729	1.357	2.1208	825	8.6
elevation	9553	2354.00	.000	1	.0

1.8.1.7 HANDLING OUTLIER VALUE

Outliers are unwanted entries which affects the data in one or the other form and distorts the distribution of the data and may mislead the algorithm[15,16] So, higher and lower extreme values are discard from independent variables. Finally after the outliers are detected only 9543 instances with 8 attributes are used for the experiment.

1.8.1.8 DATA REDUCTION

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results Kaiser [15].In this study there is 30 years total data set due to the heterogeneity problem of the data, data reduction is applied and only 25 years data are used for this study. Moreover, two attributes (date and elevation) are discarded based on information gain ranker result.

1.8.1.9 DATA TRANSFORMATION

Now, we have reliable data we can make it more efficient. The uses of feature selection methods to reduce dimensionality combine features into new ones are implemented at this point. Also, as discussed before the data file saved in Commas Separated Value (CSV) file format which is appropriate for WEKA and the datasets are discretized and aggregated to reduce the effect of scaling on the data i.e. discretizing continuous value of rainfall in two class. If the value of rainfall is above 0 which is categorized under 'yes' class and the value of rainfall is 0 which is categorized under 'no'. On the other hand, to prepare the data for one month's ahead prediction daily data was aggregated. Finally, WEKA class balancer used to balance the positive and negative class.

1.9. SELECTION OF MODELING TECHNIQUES

Selecting appropriate model depends on the main goal of the problem to be solved and the structure of the available data Gibert [16]. Consequently, to attain the objectives of this research four classification techniques has been selected for model building. The analysis was performed using WEKA experimenter environment and Explorer. Among the different available classification algorithms in WEKA, MLP, J48, IBK and PART algorithm are used for experimentation of this study. The researcher selected the above algorithms, easy of understanding and interpretation of the result of the model and appropriateness for weather forecasting.

1.9.1 EXPERIMENT AND RESULT INTERPRETATION

In this study different experiments were conducted using various machine learning methods to derive knowledge from preprocessed data to predict rainfall from metrological data. According to the methodology of this study after preparation of the data, the next task is the mining process. As it has been stated in the previous sections, after preprocess are completed from the total of 30 years data set 25 years data set with 9,543 instance and 8 selected attributes applied for experimentation.

1.9.2 MODEL BUILDING

In this study, to build the model, the experimentations are performed on WEKA 3.8 experimenter and explorer by using single algorithm and ensemble method. Also those experiments was tested with cross validation and 75 percentage split by using MLP,IBK, J48 decision tree and PART algorithm classifier. To get the best performance of the model, the researcher conducted different cross validation and percentage split values on the experiment schemes as depicted in comparison table.

Moreover, the performance of the model in this study was evaluated using the standard metrics of the accuracy, mean absolute error and root mean absolute error, ROC curve analysis and confusion matrix.

1.9.3 MODEL PERFORMANCE COMPARISON FOR ONE DAY AHEAD PREDICTION

1.9.4 EVALUATION METRICS

Some error measures are more useful than others. In selecting the best algorithms and parameters generated the best model for rainfall forecasting , the following performance metrics are used.

Evaluation on single Algorithm & Ensemble Method for one Day ahead Prediction using Selected Attributes

Method	Learning Algorithm	MAE	RMSE
Single Algorithm	PART	0.12	0.26
Boosting	PART	0.05	0.22

The above table showed that boosting classifiers has led to considerable decrease of prediction error from (MAE=0.12) to (MAE = 0.05) compared to single algorithm with selected attributes. Also, the prediction error in boosting decrease from (RMSE=0.26) to (RMSE=0.22) compared to single algorithm.

Evaluation using ROC area and Confusion Matrix for one Day ahead Prediction using Selected Attributes

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -weka.classifiers.rules.PART -- -B -M 2 -C 0.25 -Q 1

Instances: 9543

Attributes: 8

Test mode: 10-fold cross-validation

PART decision list

Correctly Classified Instances 9129 95.6617 %

Incorrectly Classified Instances 414 4.3383 %

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.973	0.103	0.971	0.973	0.972	0.984	no
	0.897	0.027	0.904	0.897	0.900	0 .984	yes
Weighted Avg.	0.957	0.086	0.957	0.957	0.957	0.984	

=== Confusion Matrix ===

a	b	<-- classified as
7259	199	a = no
215	1870	b = yes

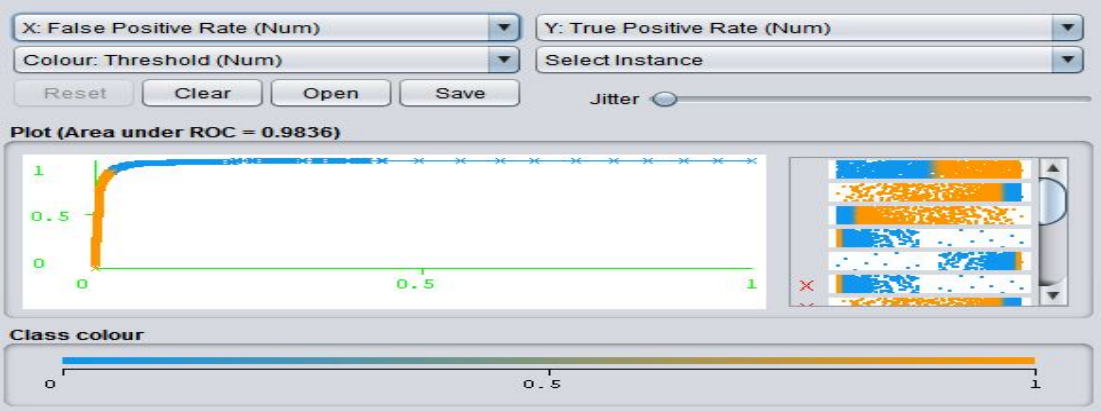
1.9.5 PART ALGORITHM MODEL FOR ONE DAY AHEAD PREDICTION

In PART algorithm predictive model when the model is built on WEKA 3.8 Explorer, the predictive performance of the model is 95.66% i.e. **9129** instances were classified correctly while 414 (4.34%) instances were wrongly misclassified to other class. The model classified **1870** instances as 'yes' out of **2085** instances; in fact they are yes tested on the test data or which are classified correctly in the class of 'yes'. The remaining **215** instances were misclassified to another class as "no" actually they are "yes".

The model classified 7259 instances as " no" out of 7458 instances that in fact " no" and wrongly classified 199 instances to other class as "yes" while actually they are "no".

On the other hand, ROC curves measure the relationship between hit rate and false alarm rate and are based on estimations of the probability that an instance belongs to a certain class. ROC has an especially attractive property in that they are insensitive to changes in class distributions.

When we come to the ROC curve of the selected model i.e. PART algorithm, the true positive case (sensitivity) and false positive case (specificity) are represented with 'X' and 'Y' axis. The y-axis expresses the number of true positives and the false positives are shown on the x-axis. Both axes show the percentage of all true and false positives to give the unit square area the sum of 1. Also instances are predicted as 'yes' actually there exist rain and predicted as 'no' actually there were no rain. The following Figure shows the ROC area of PART algorithm model with selected attributes.



In the above ROC curve, initially moves sharply up from zero, it shows that the model is better in detecting true positive rather than false positive. At the end of the curve trade off and become more horizontal showing that from the point where the curve starts to bend to onwards, false positivity outweighs true positivity i.e. the more the curve bend to the right, the more the false positivity rate and the less true positivity rate. The area under the model PART algorithm model is 0.9836 which is closer to 1 showing that the class value 'yes' gives ROC accuracy of 98.4%

1.9.6 DISCUSSION

In this study, comparison of classification techniques with ensemble method and single algorithm are performed. PART algorithm model was selected using selected attributes with 95.66% accurate prediction. Furthermore, different experiments are done using 75 % split and cross validation. The experimental result shows that with 75% split on single algorithm decision tree and IBK achieved better result. IBK algorithm produces 94.10 % prediction accuracy followed by decision tree with 92.65%, prediction accuracy for one days advance prediction using selected attributes. Boosting ensemble using 75% split PART algorithm and J48 decision tree produced 95.0% and 94.84 prediction accuracy respectively for one days advance prediction using selected attributes. Besides, J48 decision tree produced 94.49% prediction accuracy for one days advance prediction using all attributes and 10 fold cross validation. Also, in this study, ensemble method was given better prediction accuracy rather than single algorithm with 75 percentage split and cross validation. On the other hand, the experiments after one month's ahead showed that the prediction accuracies achieved by single algorithm, IBK gives better result rather than Decision tree, MLP and PART algorithm which are demonstrated on 10 fold cross validation. IBK algorithm using all attributes produces 93.94 % prediction accuracy followed by J48 decision tree and PART algorithm with 92.82%, prediction accuracy. Boosted PART algorithm and J48 decision tree produced 95.03% and 95.18%, prediction accuracy respectively by using selected attributes. Moreover, PART algorithm model selected with 95.35% and 97.12% prediction accuracy, on WEKA experimenter and explorer respectively using all attributes for one month's advance prediction.

Also, the experimental results after one month's advance using selected attributes, on 75% split J48 and IBK achieved better results rather than MLP and PART algorithm, i.e. IBK algorithm produced 94.24% prediction accuracy followed by decision tree with 92.43%, prediction accuracy. Although, in this study, boosting has led to considerable decrease of prediction error, from (MAE=0.12) to (MAE = 0.05). We achieved this result when single PART algorithm and boosted PART algorithm are compared respectively, with selected attributes, for a day's advance prediction. Also, when we compared single PART algorithm and boosted PART algorithm. The prediction error decreased from (RMSE=0.26) to (RMSE=0.22) respectively, for one day's advance prediction using selected attributes. On the other hand, for a month's advance prediction boosting classifiers has led to considerable decrease of prediction error from (MAE=0.08) to (MAE = 0.05). This result is achieved when we compared single PART algorithm and boosted PART algorithm with all attributes. Hence, PART algorithm selected for further rainfall prediction. In general, based on information gain ranker experiment, year, month, maximum temperature, minimum temperature, humidity, sunshine and Wind speed are the major attributes to predict rainfall. However, based on domain expert suggestion, year is not necessarily relevant for rainfall forecaster model building rather its relevant for trend analysis. Beyond this, the data and method used in this study can utilize as baseline in future related researches.

2. CONCLUSION

In general, ensemble methodology imitates several opinions before making a crucial decision. The core principle is to weigh several individual pattern classifiers, and combine them in order to reach a classification that is better than the one obtained by each of them separately. An ensemble is largely characterized by the diversity generation mechanism and the choice of its combination procedure. This study attempted to explore machine learning technology on rainfall predictive modeling using data base of ENMA. The hybrid, iterative methodology, was employed in this study which consists of six basic steps such as problem domain understanding, data understanding, data preparation, machine learning, evaluation and use of the discovered knowledge. In this study, we compare the effectiveness and performances of several machine learning techniques such as decision tree, PART algorithm, MLP and IBK for rainfall prediction. The novelty of this study is comparison of classification techniques with ensemble method and single algorithm. To conclude that in this study year, month, maximum temperature, minimum temperature, humidity, sunshine and Wind speed are the major attributes to predict rainfall. Experimental results, after one day's ahead prediction showed that better accuracies are achieved by Ensemble method. Although, IBK algorithm using all and selected attributes produced 94.31 % and 94.83% prediction accuracy respectively, on WEKA experimenter. Boosted J48 decision tree produced better result using all attributes in 10 fold cross validation and bagging 75 percentage split i.e. 94.49% and 94.17% respectively, on WEKA experimenter. Besides, boosted PART algorithm and J48 decision tree produced better prediction accuracy with selected attributes which is 95.46% and 95.44%, prediction accuracy respectively on WEKA experimenter. However, PART algorithm produced 95.66% prediction accuracy using selected attributes, on WEKA explorer for one day's advance prediction.

On the other hand, the results of this study using 75% split on single algorithm IBK and achieved better results rather than J48, MLP and PART algorithm. IBK algorithm produced 93.62 and 94.10 % prediction accuracy with all and selected attributes respectively, on WEKA experimenter. Boosting ensemble method on 75% split PART algorithm and J48 decision tree produced better prediction accuracy which are 95.0% and 94.84, prediction accuracy respectively, on WEKA experimenter. Furthermore, In this study, ensemble method were given better prediction accuracy rather than single algorithm with 75 percentage split and cross validation testing method. Also in this study, boosting ensemble gives better prediction accuracy with 75% split and cross validation testing method rather than Bagging and staking ensemble. In addition 75 percentage split is more efficient in case of time to build the model rather than cross validation. Also, PART algorithm with 10 fold cross validation and selected attributes produced best result for one day's advance prediction. Moreover, J48 algorithm with all attributes produced better results rather than MLP, IBK and PART algorithm i.e.92.63% and 92.44 with 10 fold and 75 percentage split respectively. PART algorithm and J48 decision tree using boosting ensemble method on 75% split produced better prediction accuracy which are 94.62% and 94.23% respectively, on WEKA experimenter. Interestingly, IBK performs better than J48 decision tree, PART algorithm and MLP when algorithms are compared with out ensemble method. The algorithms performance are increased on WEKA explorer rather than experimenter. In this study, ensemble methods are always more efficient than the individual algorithm. For this dataset, boosting ensemble works best than bagging and staking but bagging and staking are better than single algorithm.

2.1 RECOMMENDATION

According to the presented conclusions, this report intends to provide some guidelines for improving current practice and initiate further research. Thus, the following recommendations could be made considering as they are important issues for further research directions in weather forecasting strategies. In this research attempts were made to explore machine learning technology to build rainfall predictive modeling based on predefined classes (yes, no). Results found from this research should be given attention, so as to have a better decision making in the Ethiopian metrological agency service particularly should give special attention to best attribute selected as rainfall predictors such as temperature, wind speed, sunshine, relative humidity, year and month. Although J48 decision tree, IBK, PART algorithm and MLP approaches resulted in an encouraging output, still performance improvement is expected. Hence, ensemble other classification algorithms such Bayesian network (Belief network) and SVM which have also been proved to be important techniques in the rainfall forecasting should be tested in order to investigate their applicability to the problem domain in the program by using the entire dataset. The possibility of incorporating the findings of this study with web based system using ASP.NET needs further study; because ASP.NET is more appropriate for web based weather report generator by fetching the data directly from NMA online database.

REFERENCES

[1]. P. SaikiaDutta et.al, "Prediction of Rainfall using Data mining technique over ASSAM",Indian Journal of Computer Science and Engineering (IJCSE), vol. 5,pp. 85-90, May 2014.
[2]. P.Vyas "To Predict Rain Fall in Desert Area of Rajasthan using Data Mining Techniques",International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 3, pp. 323-327, May 2015.
[3]. O.F. Oyediran and A.B Adeyemo , "Performance Evaluation of Neural Network MLP and ANFIS models for Weather Forecasting Studies", vol. 6, pp.147-164, March 2013.
[4]. A. Taksande and P. S. Mohod "Applications of Data Mining in Weather Forecasting Using Frequent Pattern Growth Algorithm",International Journal of Science and Research (IJSR), vol. 4, pp. 3048-3051, June 2013.
[5]. The guardian, "Global Development about Food Security", December, 2015,http://www.theguardian.com
[6]. N. Kohail& M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis",International Journal of Information and Communication Technology Research, vol. 1, pp. 96-100, July 2011.
[7]. N.Sethi et.al, "Exploiting Data Mining Technique for Rainfall Prediction", International Journal of Computer Science and Information Technologies, vol. 5(3), pp. 3982-3984, 2014.
[8]. S. MoniraSumi et.al, "Rainfall forecasting method using machine learning models and its application to the fukuoka city case",Int. J. Appl. Math. Comput. Sci, vol. 22, no. 4, pp. 841- 854, 2012.
[9]. T. Ephrem, "Application of Data Mining for Weather Forecasting", Msc. Thesis, Dept. Information Science, Addis Ababa University, Addis Ababa, 2015.
[10]. R. Kanth et.al, "analysis of indian weather data sets using data mining techniques"Computer Science & Information Technology (CS &IT),pp. 89–94, 2014.
[11]. N.Khandelwal and R.Davey, "Climatic Assessment Of Rajasthan's Region For Drought With Concern Of Data Mining Techniques", International Journal Of Engineering Research and Applications (IJERA) , vol. 2, pp.1695-1697, October 2012.
[12]. Beniwal and Arora, "Classification and Feature Selection Techniques in Data Mining", International Journal of Engineering Research & Technology (IJERT), vol. 1, pp. 1-6, August 2012.
[13]. C. Onwubolu et.al, "Self-organizing data mining for weather forecasting", IADIS, European Conference on Data Ming, 2007.
[14]. J.Kaiser, "Dealing with Missing Values in Data", Journal of Systems Integration, vol. 1, pp. 42-51, 2014.
[15]. Vijendra and Shivani, class lecture, Topic: "Robust Outlier Detection Technique in Data Mining: A Univariate Approach" Faculty of Engineering and Technology, Mody Institute of Technology and Science, Lakshmangarh, Sikar, Rajasthan, India.

[16]. K. Gibert et.al, "On the role of pre and post-processing in environmental data mining", International Congress on Environmental Modeling and Software, Modeling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada, 2008.