

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY FOR RAINFALL FORECASTING USING ENSEMBLE METHODS

By

YIGEZU AGONAFIR

A Thesis Submitted to the School of Information Science Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

ADDIS ABABA, ETHIOPIA

OCTOBER 2016

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

This is to certify that the thesis prepared by Yigezu Agonafir, entitled: Application of Data Mining Technology for Rainfall Forecasting using Ensemble Methods, and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Information Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner: Ato Getachew Jemaneh Signature _____ Date _____

Examiner: Martha Yifru (PhD) Signature _____ Date _____

Advisor: Temtim Assefa (PhD) Signature _____ Date _____

Chair of Department or Graduate Program Coordinator

DECLARATION

I declare that this thesis is my original work and has not been presented for a degree in any other University.

_____

Yigezu Agonafir

This thesis has been submitted for examination with my approval as university advisor.

_____

Dr. Temtim Assefa

Addis Ababa University

Addis Ababa, Ethiopia

October 2016

ACKNOWLEDGEMENT

First and foremost extraordinary thanks the almighty God and his Mother Saint-Marry. I would like also to thank my thesis advisor Dr. Temtim Assefa for his keen insight, guidance, and unreserved advising. The door to Dr. Temtim's office was always open whenever, I ran into a trouble spot or had a question about my research or writing. I am really grateful for his constructive comments and critical readings of this thesis. Also, special thanks to Dr. Martha Yifru for her special help and constructive comment.

Furthermore, I would also like to thank the experts who were involved in the result interpretation of this study. Also special thanks to Goytom, Yitbarek Abebe and all National Meteorological Agency experts who are contributed their passionate support and input; the result interpretation could not have been successfully completed without them.

Finally, I must express my profound gratitude to my father Agonafir Asaminew and my mother Tadelech Ayitenfisu, for their unfailing support and continuous encouragement throughout my years of study. The process of researching and writing of this thesis would not have been possible without them.

LIST OF ACRONYMS

ANFIS	Neuro-Fuzzy Inference System
BPN	Back Propagation Neural Network
CANN	Cluster based Artificial Neural Network
CRISP-DM	Cross Industry Standard Process for Data Mining
ENMA	Ethiopian National Meteorological Agency
FFNN	Feed Forward Neural Network FFNN
IBK	Instance Based K nearest Neighbor classifier
KEEL	Knowledge Extraction based on Evolutionary Learning
KNN	K-Nearest Neighbor
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MNLR	Multiple and Nonlinear Regression
MPR	Multi Variables Polynomial Regression
NMA	National Meteorological Agency
NMAS	National Meteorological Agency Service
PANN	Periodic Artificial Neural Network
PART	Projective Adaptive Resonance Theory
PCA	Performance Component Analysis
RMSE	Root Mean Square Error
SAS	Software Analysis Service
SEMMA	Sample, Explore, Model, Modify, Access
SVM	Support Vector Machine
TANN	Threshold based Artificial Neural Network
WEKA	Waikato Environment for Knowledge Analysis

Contents

ABSTRACT XI
CHAPTER ONE 1
1. Introduction 1
1.1 Statement of the Problems 2
1.2 Research Questions 4
1.3 Objectives of the Research 4
1.3.1 General Objective 4
1.3.2 Specific Objective 4
1.4 Scope of the Study 4
1.4.1 Limitation of the Study 5
1.5 Significance of the Study 5
1.6 Organization of the Thesis 6
CHAPTER TWO 7
2.0 Literature Review 7
2.1 Introduction 7
2.2 Economic Value of Improved Weather Information 8
2.2.1 Improved Weather Information Benefit 8
2.2.2 Weather Forecasting for Agriculture 9
2.2.3 Weather Forecasting for Aviation Service 9
2.3 Method of Weather Forecasting 9
2.3.1 Persistence 9
2.3.2 Synoptic 10
2.3.3 Statistical 10
2.3.4 Computer Modeling 10
2.4 Data Mining Techniques for Rainfall Forecasting 10
2.4.1 Artificial Neural Network 10
2.4.2 IBK (K - Nearest Neighbor) 12
2.4.3 Decision Tree 13
2.4.4 PART Algorithm 13
2.5 Ensemble Classifier 13
2.5.1 Bagging 14
2.5.2 Boosting 14
2.5.3 Blending 14
2.6 Data Mining Research Methodology 15

2.6.1 CRISP, SEMA & KDD Data Mining Methodology 15
2.6.1.1 Description of the KDD Process 15
2.6.1.2 CRISP Data Mining Methodology 17
2.6.1.3 SEMMA Process Model 19
2.6.1.4 Hybrid Methodology 21
2.7 Comparative Analysis of KDD, CRISP-DM, SEMMA & Hybrid Methodology 23
2.8 Data Mining Implementation in Weather Forecasting Related Work 25
CHAPTER THREE 35
3. Methodology 35
3.1 Study Area 35
3.1.1 Overview of Ethiopian National Metrology Agency 35
3.2 Hybrid Methodology 36
CHAPTER FOUR 39
4. Data Preparation and Model Building 39
4.1 Understanding of the Problem Domain 40
4.1.1 Task 1: Define the Problem Domain 40
4.1.2 Data Mining Goals 40
4.1.3 Task 2: Data Mining Tool Selection 41
4.1.4 Task 3: Identifying the key Domain Expert 45
4.1.5 Task 4: Learning about Current Solutions to the Problem 45
4.1.5.1 Using Ensemble Method for Rainfall Forecasting 46
4.2 Data Understanding 46
4.2.1 Data Source and Data Collection 46
4.2.2 NMA Database 46
4.2.3 Description of Raw Data Quality 46
4.2.4 Data Formatting 48
4.2.5 Attribute Selection 48
4.2.6 Attributes Rank with Information Gain 48
4.3 Preparation of the Data 49
4.3.1 Data Cleaning 50
4.3.2 Missing Values Treatement 50
4.3.3 Handling Outlier Values 51
4.3.4 Data Reduction 51
4.3.5 Data Transformation 52
4.4 Data Mining 56

4.4.1 Selection of Modeling Techniques 58
4.4.2 Experiment and Result Interpretation 58
4.4.2.1 Overview of Experimentation 58
4.4.3 Experimental setup for Mining 58
4.4.3.1 Experimental set up for Ensemble Method 59
4.4.4 Model Building 59
4.4.5 Model Performance Comparison for one day ahead Prediction 60
4.4.5.1 Model performance Comparison using all Attributes with single Algorithms 60
4.4.5.2 Model Performance Comparison using Selected Attributes with Single Algorithms 61
4.4.5.3 Model Performance Comparison using all Attributes with Ensemble Methods 63
4.4.5.4 Model Performance Comparison using Selected Attributes with Ensemble Methods 66
4.4.6 Model Performance Comparison for one Day ahead Prediction 69
4.4.6.1 Single Algorithms Performance using all Attributes for one Month advance Prediction 69
4.4.6.2 Ensemble Algorithms Performance for one Month ahead Prediction using all Attributes 71
4.4.6.3 Single Algorithms Performance for one Month ahead Prediction using Selected Attributes 73
4.4.6.4 Ensemble Algorithms Performance for one Month ahead Prediction using Selected Attributes 75
4.5 Selected Model Performance Evaluation 76
4.5.1 Evaluation Metrics 77
4.5.2 PART Algorithm Model with Selected Attributes for one Day ahead Prediction 79
4.5.3 PART Algorithm Model with all Attributes for one Month ahead Prediction 81
4.5.4 Discussion 82
4.5.5 Rule Extraction by using PART Algorithm for one Day advance Prediction 84
4.5.6 Rule Extraction by using PART Algorithm for one Month advance Prediction 87
4.6 Using Discovered Knowledge 88
4.6.1 Prototype development for Rainfall forecasting 88
CHAPTER FIVE 90
5. Conclusion and Recommendation 90
5.1 Conclusion 90
5.2 Recommendation 92
References 93
Annex 1: Outlier detection 97
Annex 2: Experimental Set up 99
Annex 3: Sample Experimentation with Selected attributes 104
Annex 4: Partial view of PART Algorithm rule for one day ahead Prediction 106
Annex 5: VB.NET Sample Code for Rainfall Forecasting 108

LIST OF TABLES

Table 1: Summary of KDD, CRISP-DM, SEMMA and hybrid Processes model 24
Table 2: Literature review summary on data mining techniques applied on rainfall prediction 33
Table 3: Performance Criteria 42
Table 3: Functionality Criteria 43
Table 5: Auxiliary task support 44
Table 6: List of attributes in the initial data set 47
Table 7: Attribute Rank on all input Data using information gain ranker 48
Table 8: Selected attribute from the NMA database 49
Table 9: Univariate Statistics of metrological dataset 51
Table 10: Experimental Setup on Addaboostm1 59
Table 11: Single algorithms with all attributes using cross validation for one day ahead prediction 60
Table 12: Single algorithms with all attributes using 75% split for one day ahead prediction 61
Table 13: Single algorithms with selected attributes using cross validation for one day ahead prediction 61
Table 14: Single algorithms with selected attributes using 75% split for one day ahead prediction 62
Table 15: Ensemble method with all attributes using cross validation for one day ahead prediction 63
Table 16: Ensemble methods with all attributes using 75% split for one day ahead prediction 65
Table 17: Ensemble methods with selected attributes using cross validation for one day ahead prediction 66
Table 18: Ensemble methods with selected attributes using 75% split for one day ahead prediction 68
Table 19: Single algorithms with all attributes using cross validation for one month ahead prediction 69
Table 20: Single algorithm with all attributes using 75% split for one month ahead prediction 70
Table 21: Ensemble method with all attributes using cross validation for one month ahead prediction 71
Table 22: Ensemble method with all attributes using 75% split for one month ahead prediction 72
Table 23: Single algorithms with selected attributes using cross validation for one month ahead prediction 73
Table 24: Single algorithms with selected attributes using 75% split for one month ahead prediction 74
Table 25: Ensemble methods with selected attributes & cross validation for one month ahead prediction 75
Table 26: Ensemble methods with selected attributes using 75% split for one month ahead prediction 76

LIST OF FIGURES

Figure 1: Artificial Neural network model 12
Figure 2: KDD process model 16
Figure 3: CRISP-DM process model 18
Figure 4: SEMMA process model 20
Figure 5: Hybrid DM process model 21
Figure 6: Cass attribute before balance for one day ahead prediction using all attributes 53
Figure 7: Class attribute after balance for one day ahead prediction using all attributes 53
Figure 8: Class attribute before balance for one day ahead prediction using selected attributes 54
Figure 9: Class attribute after balance for one day ahead prediction using selected attributes 54
Figure 10: Class attribute a before balance for one month ahead prediction using all attributes 54
Figure 11: Class attribute after balance for one month ahead Prediction Experiment using all attributes 54
Figure 12: Class attribute before balance for one month ahead prediction using selected attributes 55
Figure 13: Class attribute after balance for one month ahead prediction using selected attributes 55
Figure 14: General rainfall prediction work flow diagram 57
Figure 15: Single algorithms learning progress on all attributes and 75% split for a day ahead prediction 60
Figure 16: Single algorithms learning progress on all attributes and cross validation for a day ahead prediction 61
Figure 17: Single algorithms learning on selected attributes and cross validation for a day ahead prediction 62
Figure 18: Single algorithms learning progress on selected attributes and 75% split for a day ahead prediction 62
Figure 19: Ensemble learning progress on all attributes and cross validation for a day ahead prediction 64
Figure 20: Ensemble learning progress on all attributes and 75% split for a day ahead prediction 65
Figure 21: Ensemble learning progress on selected attributes and cross validation for a day ahead prediction 67
Figure 22: Ensemble method learning progress on selected attributes and 75% split for a day ahead prediction 68
Figure 23: Single algorithms learning on all attributes and cross validation for a month ahead prediction 69
Figure 24: Single algorithms learning progress on all attributes and 75% split for a month ahead prediction 70
Figure 25: Ensemble learning progress on all attributes and cross validation for a month ahead prediction 71
Figure 26: Ensemble learning progress on all attributes and 75% split for a month ahead prediction 72
Figure 27: Single algorithms learning on selected attributes and cross validation for a month ahead prediction 73
Figure 28: Single algorithms learning on selected attributes and 75% split for a month ahead prediction 74
Figure 29: Ensemble learning progress on selected attributes and cross validation for a month ahead prediction 75
Figure 30: Ensemble learning progress on selected attributes and 75% split for a month ahead prediction 76
Figure 31: Prototype development for rainfall prediction.....88

ABSTRACT

Rainfall is one the natural phenomena which must be known to use it as a resource or avoid its undesirable effects. Rainfall 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. It is even difficult to find the best parameter values for a specific technique, since these settings also problem dependent. Ensembles are more robust and powerful than individual models, in this thesis ensemble diversity used to solve wide range of rainfall prediction problems.

The main objective of this study is appraising the potential applicability of data mining technology and machine learning 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 data mining methodology. Also, the researcher used 9,543 instances with 8 selected and 10 all attributes on WEKA 3.8. In this thesis, J48, PART, MLP and IBK algorithms are used with ensemble method. In addition VB.NET 2010 is used for interface development to use the discovered knowledge. 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 all 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.

CHAPTER ONE

1. Introduction

Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Meteorological agencies have attempted to predict the weather by using different techniques like wolf model, data mining technology and expert system. 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 data mining 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 data mining 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. The methods of forecasting which are applied in preparing seasonal forecasts in the Ethiopian National Meteorological Agency are based on traditional forecasting method like analogue, trend analysis and statistical assessment [4].

Furthermore, ENMA uses wolf model for standard statistical and dynamic prediction. However, as Teksande and Mohod [5], 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 data mining methods for rainfall prediction that supports metrology experts, planners and policy makers in Ethiopia.

1.1 Statement of the Problems

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 [6]. "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" [6]. 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 [7].

In such situations, technologies like data mining will be helpful in discovering hidden patterns for risk reduction rather than asking emergency assistance. Data mining 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 data mining 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 data mining methodology and three single data mining 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 [5]. Dhanya & Kumar [9] conducted a research on "Data mining 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 [7].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. Also, the previous researcher's prediction duration covers only one day's ahead advance. Moreover, 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. Also, those previous researchers didn't develop interface to use the discovered knowledge. 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, extend prediction duration a month advance and develop a prototype system based on the discovered knowledge for the purpose of rainfall forecasting after one day and one month advance.

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

1.3 Objectives of the Research

1.3.1 General Objective

The general objective of this research is to investigate the potential applicability of data mining and machine learning for longer duration rainfall prediction by using ensemble methods.

1.3.2 Specific Objective

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

- To review the literature on data mining technology for rainfall prediction.
- To collect and prepare the data for analysis by cleaning and transforming the data.
- To identify appropriate data mining application software and access different algorithms that are more appropriate to the problem domain.
- To build the model using ensemble methods and single algorithms.
- To evaluate the model, and select the best algorithms for longer duration rainfall forecasting.

1.4 Scope of the Study

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

1.4.1 Limitation of the Study

This thesis is unable to include flood forecasting based on the predicted rainfall level. In this thesis, 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 thesis the researcher could not cluster based on similarity; and apply association rule discovery techniques to investigate the internal association exists among the different variables.

1.5 Significance of the Study

This research will contribute to an understanding of data mining technology for rainfall forecasting and identify the best data mining 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.

1.6 Organization of the Thesis

This thesis is divided into five chapters. The first chapter is an introduction part, which contains background to the research work, statement of the problem , objective of the research, scope, limitation and Significance of the study. The second chapter deals about literature review on related data mining research for rainfall prediction and its application for meteorological data.

The third chapter is devoted to deal about overview of Ethiopian National Meteorological Agency, data mining methodologies/techniques and philosophies that are used in this study.

The fourth chapter provides discussions about the different data mining steps that were undertaken in this research work. This includes problem identification, data understanding, data preparation; model building and evaluating the results and using discovered knowledge are performed by using WEKA 3.8 software. The last chapter is devoted for the final conclusion and recommendations forwarded based on the research findings.

CHAPTER TWO

2. Literature Review

In this chapter, endeavor has been made to review the literature on the concepts and techniques about the state of the art in data mining technology applications for rainfall prediction which are aimed to provide potentially useful background about the models to be built.

2.1 Introduction

Data mining involves the use of various sophisticated data analysis tools for discovering previously unknown, potentially useful patterns and relationships in a massive data set. There are a number of applications for Machine Learning (ML), the most significant application of machine learning is data mining technology Vaithiyanathan et al. [10, 11]. Although assessment simulation and prediction of rainfall is essential for stakeholders and policy makers to plan or adopt the required policies. There are various techniques available in the literature to assess, simulate and predict rainfall. However, the selection of the techniques normally depends on the objectives of the study, availability of required input data, the quality of available models and some pre-defined assumptions.

On the other hand, rainfall prediction is a complicated task to solve this problem many techniques and methods are developed. One of these techniques is data mining technique Atole et al. [12]. Also the goal of data mining is to be clipped the data and converts them into useful data for further information. Data mining has connection between a several method like machine learning, database, artificial intelligence and statistics. In data mining technology, there is an empirical and dynamical method. The empirical method used for rainfall prediction is based on analysis of historical data of the rainfall and its relationship to a variety of atmospheric variables over different parts of the world. Dynamical approach is implemented for numerical weather forecasting method. The most widely used empirical approaches for climate prediction are regression, artificial neural network, fuzzy logic and group method of data handling Naing, [13, 14]. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses the kind of patterns that can be discovered depend upon the data mining task employed [15].

Furthermore, Data mining can discover unexpected patterns that were not under consideration when the mining process started. Data mining is helpful in taking quick decision regarding the chances of hitting and quick compilation and analysis of large volume of unstructured data from various sources helps to take timely decision making to better control over the system.

However, to facilitate an adequate level of accuracy, the developer has to be responsive to the characteristics of different methods and determine if a particular method is appropriate for the undertaken situation before embarking its usage in real application. As a result, the choice of a model is one of the important factors that will influence the forecasting accuracy.
In the following sections literature review on the application of various data mining algorithms for weather forecasting in general and particularly for rainfall prediction has been summarized.

2.2 Economic Value of Improved Weather Information

Now a day's public and private economic sectors are affected by weather and climate changes. Uncertainties in the scope and severity of weather changes pose financial and social risks for individuals, businesses, and government agencies. Hence, achieving more accurate weather and climate forecasts contributes to wellbeing and the economy by reducing risk and creating new opportunities [16, 17].

2.2.1 Improved Weather Information Benefit

- Improving civil government and military planning: Weather conditions have a major role in government planning for administering forests, grasslands, and other lands under federal management. Also Military operations like war or peacetime are affected by weather conditions [15]. In addition, accurate weather information reduces risk to personnel and gives them an information edge over adversaries. In peacetime, applying weather forecasts to logistics and field operations reduces operational costs by improving routing and timing of deliveries.

- Improving natural hazard mitigation, response and recover: More accurate prediction of severe weather can help substantially reduce the costs to society of weather-related disasters [16]. Better information induces governments, businesses, and individuals to invest in loss-reduction activities; it can also reduce economic costs from unnecessary loss- education activities that derive from uncertainty about adverse weather.

2.2.2 Weather Forecasting for Agriculture

Weather plays an important role in agricultural production. It has a profound influence on the growth, development and yields of a crop, incidence of pests and diseases. Weather aberrations may cause for physical damage to crops and soil erosion. The quality of crop produce during movement from field to storage and transport to market depends on weather. Bad weather may affect the quality of produce during transport and viability and vigor of seeds and planting material during storage [16].

2.2.3 Weather Forecasting for Aviation Service

Weather forecasting play tremendous role for Pilot Briefing Service (PBS). PBS is the provision of consultation on meteorological and aeronautical information to assist pilots in flight planning and includes a fully interpretive weather briefing service [18].

2.3 Method of Weather Forecasting

2.3.1 Persistence

Persistence forecasting is based on the concept that current weather conditions can reveal clues to tomorrow's weather. Meteorologists who rely on this forecasting method predict that the current conditions will persist, or continue unchanged. They make observations using thermometers and barometers to assess the weather, and then theorize that the next few days will feature similar weather patterns. This forecasting technique works best in areas with predictable weather patterns, such as a tropical zone or an arctic region [19].

2.3.2 Synoptic

Synoptic or analogue forecasting is a method of predicting the weather based on accepted theories and principles of meteorology. This technique requires some skill and training, and incorporates weather maps, radar and satellite images. Forecasters combine these tools with information about atmospheric pressure, air flow and temperatures to come up with a forecast [20].

2.3.3 Statistical

Statistical or climatological forecasting allows meteorologists to make predictions based on historical trends. It assumes consistent weather patterns over time. Forecasters examine historical information about average, high and low temperatures to estimate future temperature ranges. They also examine historical storm records and precipitation amounts and use those as a basis for forecasting. For example, a statistical forecaster may state that the next month will bring rain and cold temperatures because that is considered the normal condition for this area at this time of year [18, 21].

2.3.4 Computer Modeling

Computer modeling forecasts represent the most advanced method of predicting the weather. This method relies on mathematical formulas that are designed to model atmospheric and weather conditions. By inputting current weather data, the meteorologist can calculate future conditions [18, 21].

2.4 Data Mining Techniques for Rainfall Forecasting

2.4.1 Artificial Neural Network

ANNs provide a methodology for solving many types of non-linear problems that are difficult to solve by traditional techniques. Most meteorological processes often exhibit temporal and spatial variability, and are further plagued by issues of non-linearity of physical processes, conflicting spatial and temporal scale and uncertainty in parameter estimation [22].

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze.

A neural network consists of a number of neurons that are arranged in layers and are linked to every neuron in the previous layer by connections with different strengths or weights associated to them. The learning adapts weights at each point of iteration and provides the system of a method to learn by example. Multi-Layer Perceptron (MLP) is currently the most popular type of neural network. It is a simplified model of the human mind elaboration process, and works by simulating and elevated number of simple elaboration units that resemble abstract versions of neurons.

As illustrated in figure 1, an input layer represents the input fields, an output corresponds to the output fields and one or more hidden layers represent the propagation from each neuron to each other. The experts can be used to provide projections of new situations to answer "what if" questions. Other advantages include:

- Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
- MLP/Neural networks do not make any assumption regarding the underlying probability density functions or other probabilistic information about the pattern classes under consideration in comparison to other probability based models.

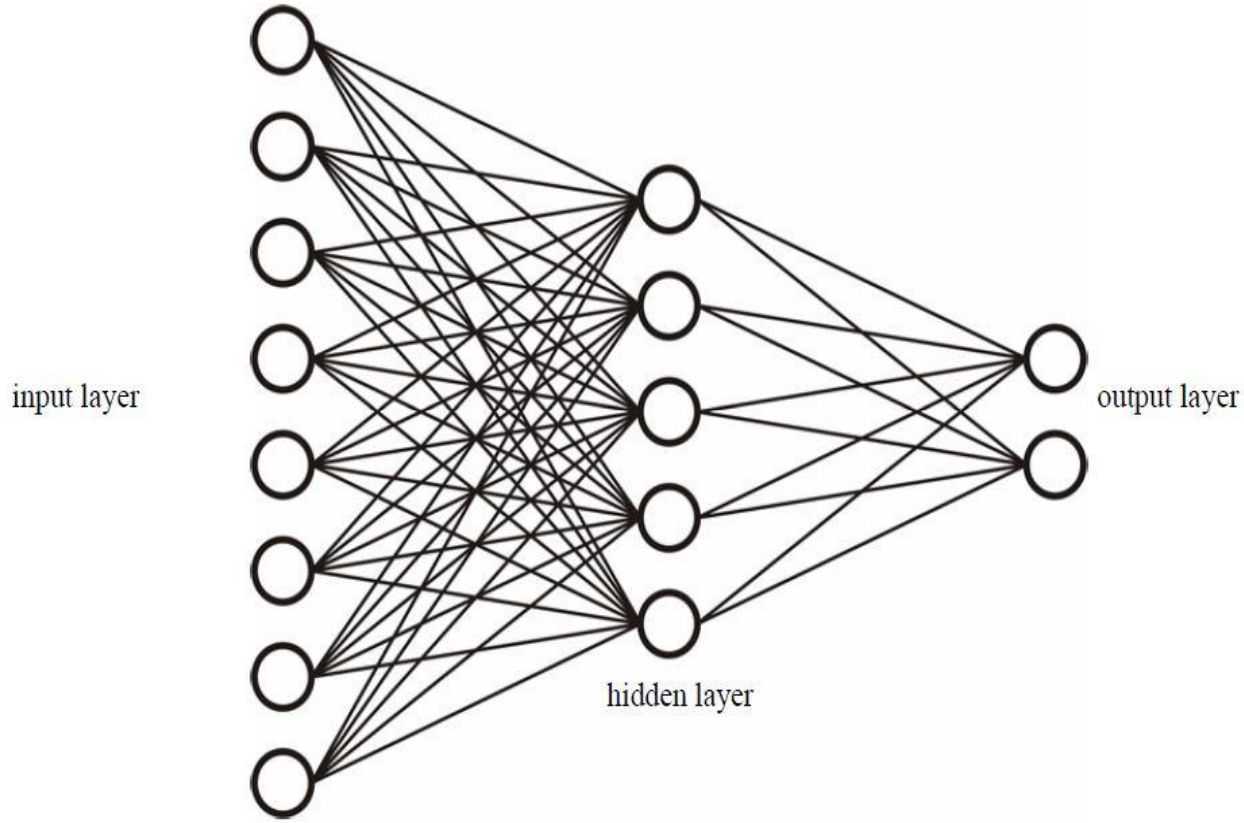


Figure1: Artificial Neural network model

2.4.2 IBK (K - Nearest Neighbor)

IBK is a k-nearest-neighbor classifier that uses the same distance metric. The number of nearest neighbors can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbors.

A linear search is the default but further options include KD-trees, ball trees, and so-called “cover trees” [23].

2.4.3 Decision Tree

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances [24].Decision trees constitute a method able to forecast or classify future observations according to decision rules. If the information is divided in classes, it is possible to utilize the data to generate rules able to classify previous cases and new cases with absolute precision. The decisional process behind the model appears clear when observing the tree, with a clear advantage with respect to other techniques whose internal logic is difficult to interpret.

Furthermore, the process includes automatically in the rule only the attributes relevant to the decision, as the irrelevant attributes are ignored and the data dimensions are reduced. Decision trees are converted into "if-then" rules to enhance the comprehension of the model when the relationship among the elements of a group is relevant.

2.4.4 PART Algorithm

PART stands for Projective Adaptive Resonance Theory. The input for PART algorithm is the vigilance and distance parameters. The rule based classification algorithms namely Part algorithms is used for generating rules [25].

2.5 Ensemble Classifier

An ensemble classifier is a machine learning method which uses or combines multiple classifiers to improve robustness as well as to achieve an improved classification performance from any of the constituent classifiers [26]. Furthermore, this technique is more resilient to noise compared to the use of a single classifier. This method uses a 'divide and conquer approach' where a complex problem is decomposed into multiple sub-problems that are easier to understand and solve wide range rainfall forecasting problems.

2.5.1 Bagging

Bagging, which means bootstrap aggregation, is one of the simplest but most successful ensemble methods for improving unstable classification problems [27]. For example, weak classifiers such as decision tree algorithms can be unstable, especially when the position of a training point changes slightly and can lead to a very different tree. This method is usually applied to decision tree algorithms, but it can be used with other classification algorithms such as MLP, IBK and PART algorithm. The most well-known independent method is bagging (bootstrap aggregating). In this study this method aims to increase accuracy by creating an improved composite classifier and by amalgamating the various outputs of learned classifiers into a single prediction.

2.5.2 Boosting

Boosting is an ensemble method for boosting the performance of a set of weak classifiers into a strong classifier. This technique can be viewed as a model averaging method and it was originally designed for classification, but it can also be applied to regression. Boosting provides sequential learning of the predictors. The first one learns from the whole data set, while the following learns from training sets based on the performance of the previous one. The misclassified examples are marked and their weights increased so they will have a higher probability of appearing in the training set of the next predictor [42].In this study boosting method aims to support decision maker's preferences, thus increasing performance and comprehensibility.

2.5.3 Blending

Stacking combines several classifiers using the stacking method [41]. It can do classification or regression. We can choose different meta-classifiers and the number of stacking folds. We can choose different classifiers, different level-0 classifiers and a different meta-classifier. In order to create multiple level-0 models, we can specify a meta-classifier as the level-0 model. In this study staking aims to achieve the highest generalization accuracy.

2.6 Data Mining Research Methodology

2.6.1 CRISP, SEMA & KDD Data Mining Methodology

2.6.1.1 Description of the KDD Process

Knowledge Discovery and Data Mining (KDD) is a multi-stage process [29]. The overall methodology will follow the main stages of KDD process, as depicted in Figure 2.

1. Developing and understanding of the application domain

Developing and understanding of the application domain is the first stage of KDD process in which goals are defined from customer's view point and used to expand and understanding about application domain and its prior knowledge.

2. Creating a target data set

Creating target data set is the second stage of KDD process which focuses creating on target data set and subset of data samples or variables.

3. Data cleaning and pre-processing

This step is composed of several sub-steps and requires understanding of both the domain to which the KDD is applied and of statistical methods for improving data quality. Hence, the cooperation of experts from various disciplines is most crucial. Data cleaning and preprocessing is the third stage of KDD process which focuses on target data cleaning and pre-processing to complete and consistent data without any noise and inconsistencies. In this stage strategies are develop for handling such type of noisy and inconsistent data.

4. Data Transformation

This is the fourth stage of KDD process which focuses on transformation of data from one to another so that data mining algorithms can be implemented easily. For this purpose different data reduction and transformation methods are implemented on target data. Data transformation aims to manipulate the data so that its content and its format are most suitable for the data mining process. The transformation process effects the distribution of the various features and the structure in which their values are stored. Various data mining techniques present different requirements regarding these characteristics of the data. The requirements of each technique should be taken into account prior to its application.

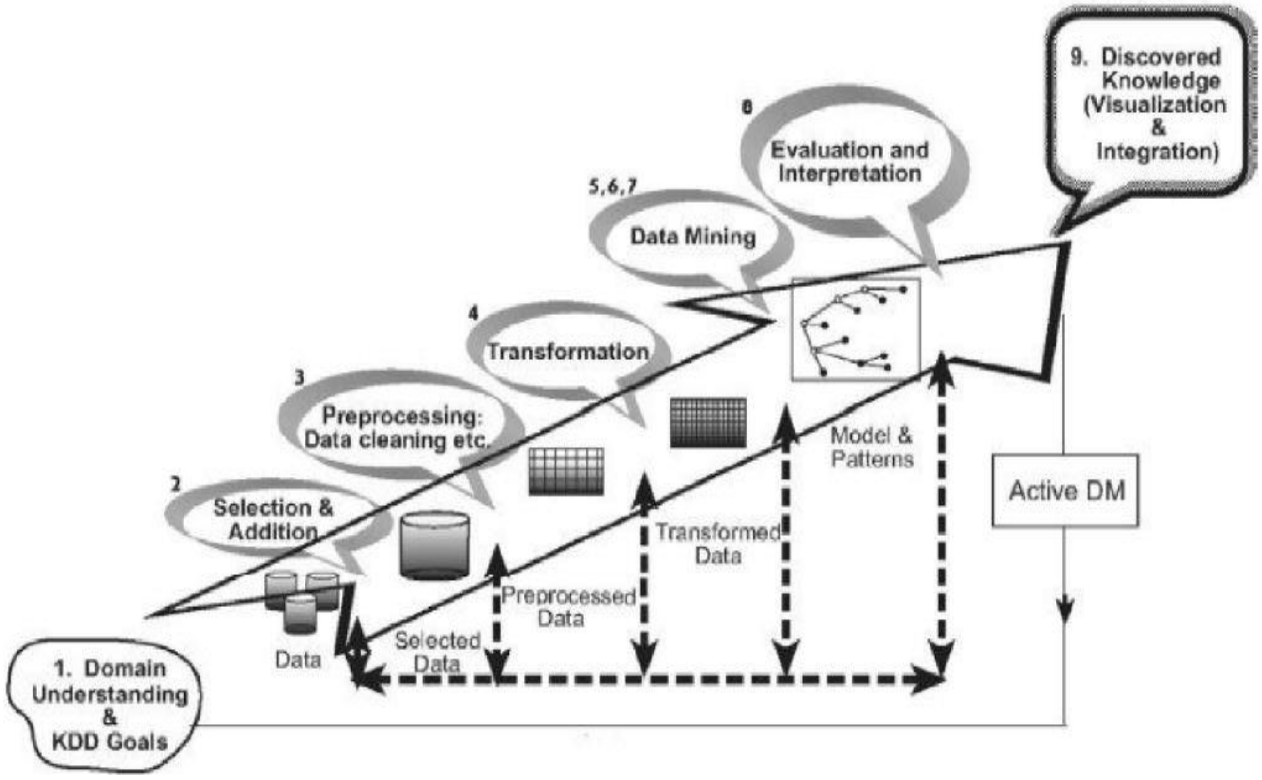


Figure2: KDD process model
Adopted from Shafique and Qaiser[61]

Each one of the above KDD steps has to be performed efficiently and thoroughly, in order to ensure the quality of the next step and the final outcome of the entire process.

5. Choosing the suitable data mining task

Choosing the suitable data mining task is the fifth stage of KDD process in which appropriate data mining task is chosen based on particular goals that are defined in first stage. The examples of data mining method or tasks are classification, clustering, regression and summarization.

6. Choosing the suitable data mining algorithm

Choosing the suitable data mining algorithm is the sixth step of KDD process in which one or more appropriate data mining algorithms are selected for searching different patterns from data. There are a number of algorithms present today for data mining but appropriate algorithms should be selected to produce quality knowledge.

7. Employing data mining algorithm

This is the seventh step of KDD process in which selected algorithms are implemented.

8. Interpreting mined patterns

Interpreting mined patterns are the eighth step of KDD process that focuses on interpretation and evaluation of mining patterns. This step involves extracted pattern visualization.

9. Using discovered knowledge

This is the last and final step of KDD process in which the discovered knowledge is used for different purposes. The discovered knowledge can also be used interested parties or can be integrate with another system for further action.

2.6.1.2 CRISP Data Mining Methodology

CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem [30]. The CRISP-DM KDP model consists of six steps, which are summarized below:

1. Business understanding. This step focuses on the understanding of objectives and requirements from a business perspective. It is further broken into several sub steps, namely, determination of business objectives, assessment of the situation, determination of DM goals, and generation of a project plan.
2. Data understanding. This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data and detection of interesting data subsets. Data understanding is further broken down into collection of initial data, description of data, exploration of data and verification of data quality. In addition, once business objectives and the project plan are established, data understanding considers data requirements. In addition, this step can include initial data description, data exploration, and the verification of data quality. Data exploration such as viewing summary statistics (which includes the visual display of categorical variables) can occur at the end of this phase. Models such as cluster analysis can also be applied during this phase, with the intent of identifying patterns in the data.
3. Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tools in the next step. It includes table, record, attribute selection, data cleaning, data transformation and construction of new attributes.

Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. Data exploration at a greater depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding.

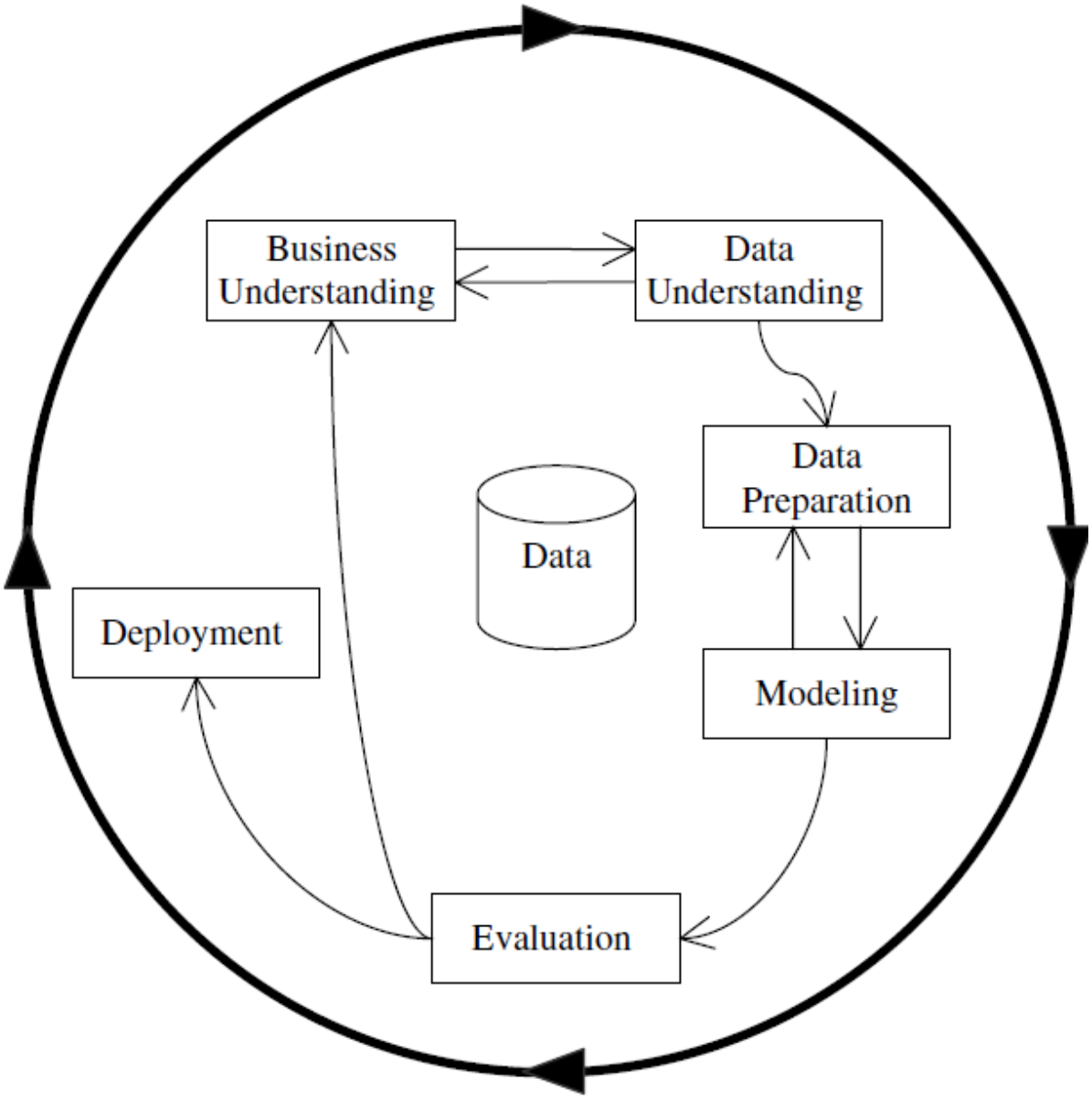


Figure 3: CRISP-DM process models

Adopted from The CRISP-DM KD process model (source: http://www.crisp-dm.org)

4. Modeling

This is the fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set and different models are built for same data mining problem. At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into selection of modeling techniques, generation of test design, creation and assessment of generated models.

5. Evaluation

This is the fifth stage of CRISP-DM, after one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached. The key sub steps in this step include evaluation of the results, process review and determination of the next step.

6. Deployment. Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating report or as complex as implementing a repeatable KDP. This step is further divided into plan deployment, plan monitoring and maintenance, generation of final report, and review of the process sub steps.

2.6.1.3 SEMMA Process Model

SEMMA stand for (Sample, Explore, Modify, Model, and Access) is data mining method developed by SAS institute [29]. It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically a logical organization of the functional tools for them. It has a cycle of five stages or steps.

Sample

This is the first and optional stage of SEMMA process which focuses on sampling of data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.

Explore

This is the second stage of SEMMA process which focuses on exploration of data. This can helps in gaining the understanding and ideas as well as refining the discovery process by searching for trends and anomalies.

Modify

This is the third stage of SEMMA process which focuses on modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also looks for outliers and reducing the number of variables.

Model

This is the fourth stage of SEMMA process which focuses on modeling of data. The software for this automatically searches for combination of data. There are different modeling techniques are present and each type of model has its own strength and is appropriate for specific situation on the data for data mining.

Access

This is the fifth and final stage for SEMMA process focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

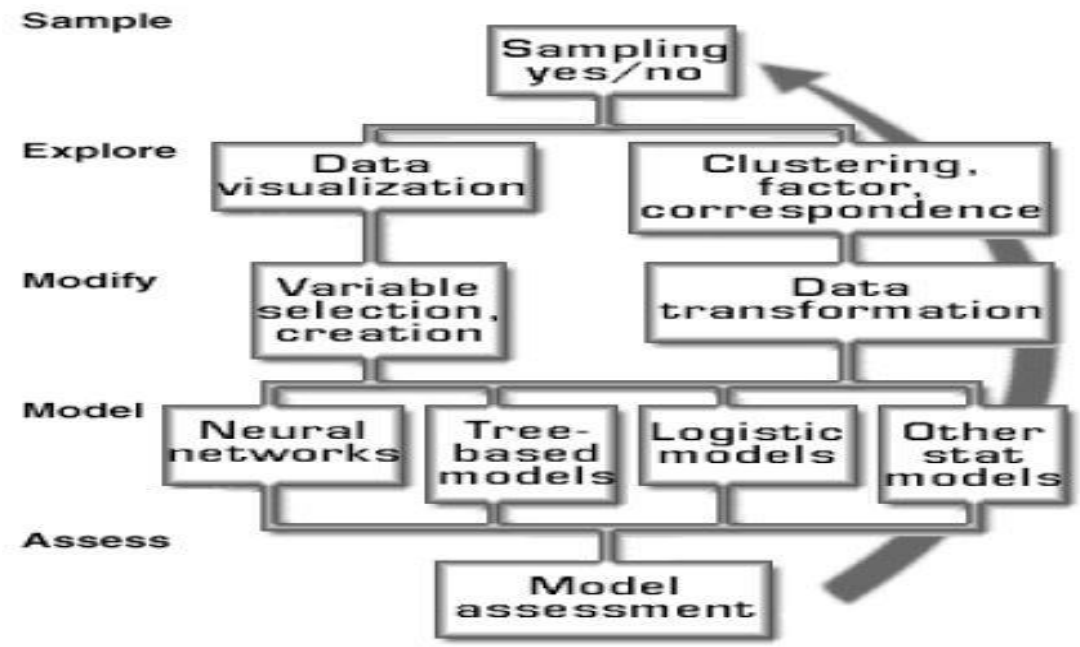


Figure 4: SEMMA process model by SAS institute

2.6.1.4 Hybrid Methodology

Hybrid process is characterized by providing more general, research oriented description of the steps [31]. The hybrid model also encourages the application of knowledge discovery for a particular domain in other domains and it has a six step process as depicted in Figure5.

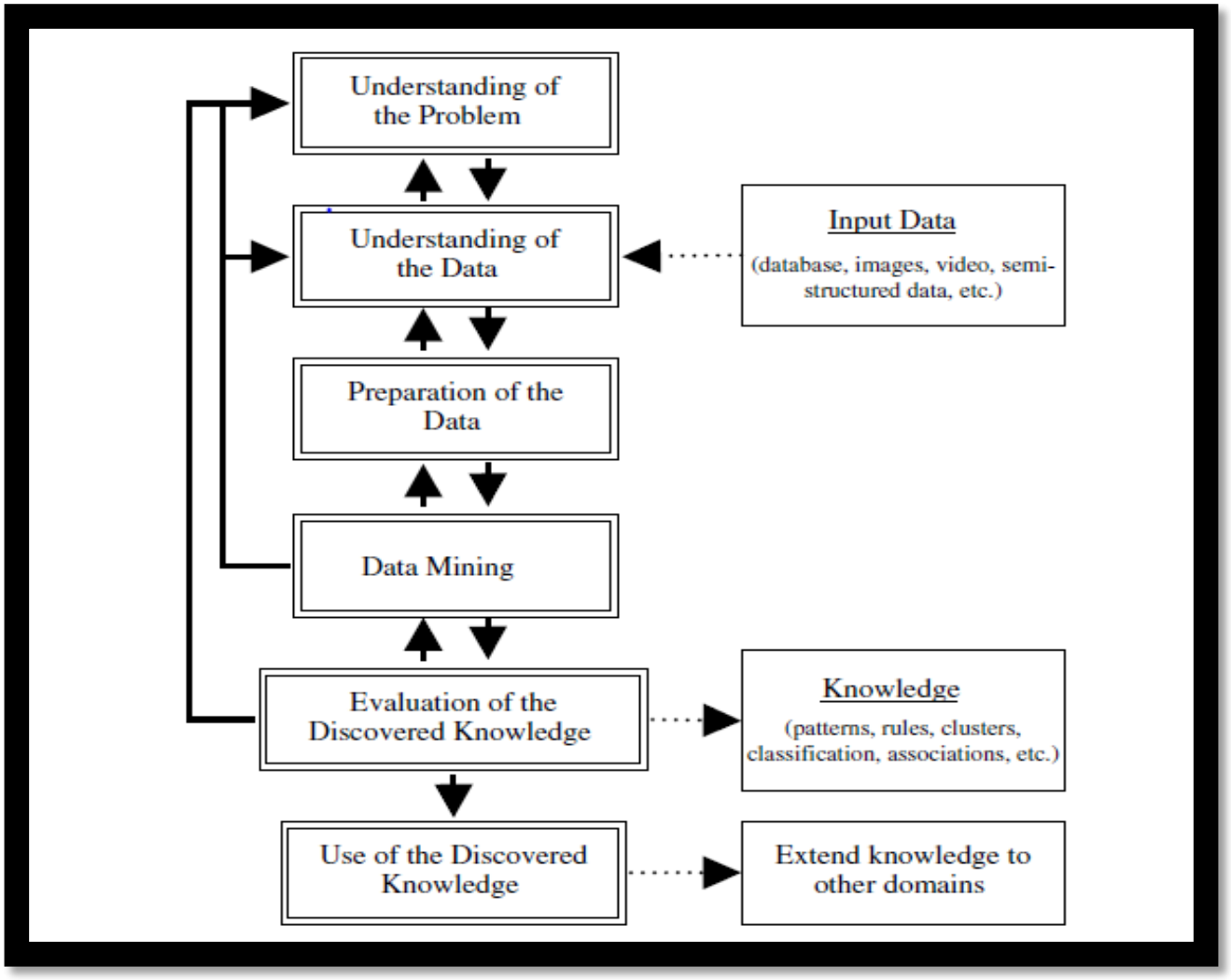


Figure 5: Hybrid DM process model Changala et.al. 2015

Description of the six step Knowledge Discovery Process

1. Understanding of the problem domain: This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem.

   It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, research goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.
2. Understanding of the data: This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.
3. Preparation of the data: This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms to reduce dimensionality by derive new attributes by discretization, and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.
4. Data mining: The data miner uses various data mining methods to derive knowledge from preprocessed data.
5. Evaluation of the discovered knowledge*:* Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.
6. Use of the discovered knowledge. This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.7 Comparative Analysis of KDD, CRISP-DM, SEMMA & Hybrid Methodology

There are nine, six, five and six stages for KDD, CRISP-DM, SEMMA and hybrid process model respectively. By examining all the three data mining process models they clearly shows that they are in somehow equivalent to each other but SEMMA is directly linked to the SAS enterprise miner software[43].

The comparison between them shows that:

- The KDD process step "Developing and Understanding of the Application Domain" can be identified with "Business Understanding" phase of CRISP-DM process.
- The KDD process steps "Creating a Target Data Set" and "Data Cleaning and Pre-processing" can be identified with "Sample" and "Explore" stages of SEMMA respectively, and these can be identified with "Data Understanding" phase of CRISP-DM.
- The KDD process stage "Data Transformation" can be identified with "Data Preparation" stage of CRISP-DM and "Modify "stage of SEMMA process respectively.
- The three stages of KDD "Choosing the suitable Data mining task", "Choosing the suitable Data Mining Algorithm" and "Employing Data Mining Algorithm" can be identified with "Modeling" phase of CRISP-DM and "Model" stage of SEMMA process respectively.
- The KDD process step "Interpreting Mined Patterns" can be identified with "Evaluation" phase of CRISP-DM process and "Assessment" stage of SEMMA process respectively.
- The KDD step "Using Discovered Knowledge" can be identified with "Deployment" phase of CRISP-DM process.
- The KDD step "developing and understanding of the application" can be identified with "understanding of the problem domain" phase of hybrid methodology.
- CRISP step "modeling" can be identified with "data mining" phase of hybrid methodology.
- CRISP step "deployment" can be identified with "using discovered knowledge" phase of hybrid methodology.

Data mining process model	KDD	CRISP DM	SEMMA	Hybrid
No. of steps	9	6	5	6
	Developing and Understanding of the Application	Business Understanding	---------	Understanding of the problem domain
	Creating a Target Data Set	Data Understanding	Sample	Understanding of the data
	Data Cleaning and Pre-processing		Explore	
	Data Transformation	Data Preparation	Modify	Preparation of the data
	Choosing the suitable Data Mining Task	Modeling	Model	Data mining
	Choosing the suitable Data Mining Algorithm			
	Employing data mining algorithm			
	Interpreting Mined Patterns	Evaluation	Assessment	Evaluation
	Using Discovered Knowledge	Deployment	----------	Use of the discovered knowledge

Table 1: Summary of KDD, CRISP-DM, SEMMA and hybrid processes model

2.8 Data Mining Implementation in Weather Forecasting related work

Many researchers have tried to use data mining technologies in areas related to weather prediction in general and particularly for rainfall prediction. Latha et al. [32] 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. [33] 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 [12] 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. [34] 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. [35] 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.

Khandelwal et al. [36] investigated that climate assessment of drought using data mining technology. The researchers explored that when the fluctuation occurs in the climate, it makes the greatest impact on the world's economy and social life of humans. The researchers consider climatic factors for weather prediction, i.e. rainfall. Rainfall is a major issue for nature, food productions and water resource management. Then the researchers apply data mining techniques like regression analysis of the rainfall dataset. The researchers are using multiple regression analysis to predict rainfall of a year by using climatic factors like temperature, humidity, pressure and sea level. For selecting these factors the researchers apply correlation analysis.

Finally, the researchers proposed a model after analysis of the Jaipur rainfall dataset which is derived by data mining techniques like firstly applying correlation analysis secondly regression analysis and lastly applying statistical analysis. Then the researchers predict rain in the future year by knowing climate factors and then use this dataset for calculating drought possibility. The researchers conclude that the topic climate is very highly complex due to climate conditions which are rapidly changing.

Hemalatha [37] was investigated the implementation of data mining methods for guiding the path of the ships by using decision tree. The implementation of the researchers study was used Global Positioning System (GPS) which helps in identifying the area in which the ship is currently navigating. The weather report on that area is compared with the existing database and the decision is made in accordance with the output obtained from the Data Mining technique. This decision about the weather condition of the navigating path is then instructed to the ship.

In addition, the researchers describe some statistical themes and lessons that are directly relevant to data mining and attempts to identify opportunities where close cooperation between the statistical and computational communities might reasonably provide synergy for further progress in data analysis. Finally, the researcher concludes that the weather report of the ships location is made to traverse the Decision Tree and the corresponding decision is passed to the ship for its safe navigation.

Kumar [38] explored that decision tree evaluation can be quantified and it is simple to use, as a result decision tree has been used by the researcher to predict the event like fog, rain and thunder by inputting average temperature, humidity and pressure. The researcher was used Decision tree (Decision stump) and it has been implemented in WEKA to facilitate the forecasting of weather.

The researchers conclude that to tap the potential of huge amount of data, decision tree can be used in predicting the dependent variable like fog and rain. Software equipped with decision tree can provide artificial intelligence to the machine. Such software can be used by trekkers, mountaineers and drivers, which can facilitate them in decision making, because many location exists where telemetric data does not exist and if it exists, it is too much location sensitive, so generalizing to a vast geographical are still remains doubtful. In this information age it is quite cost effective to equip a machine with the sensors and artificial intelligence software so that machine exhibits intelligence.

Taffese [4] investigated the application of feed forward neural network with back propagation learning algorithm to forecasting temperature. The researcher used Matlab 6.5 release 13 software for ANN module building. The ANN model generates the 3 hourly temperature of Addis Ababa one day ahead. He was used three years data set collected from National Meteorological service agency of Ethiopia i.e (2001-2003). The model is then tested with an independent data set. As the researcher conclude that the temperature is well predicted by the forecast model as, in 60% of the case of the errors are within 1 degree centigrade and about 90% of the cases are within 2 degree centigrade.

Ephrem [8] 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.

Mandale and Jadhawar [39] explored that the use of data mining technology in weather forecasting by using Artificial Neural Network and Decision tree Algorithms. The researchers used ANN to configure a particular application, such as pattern recognition or data classification through a learning process. The artificial neuron is an information processing unit that is fundamental to the operation of neural network. Also the authors used decision tree structure to provide an explicit set of if then rules (rather than abstract mathematical equations), making the results easy to interpret.

The researchers conclude that ANN's can detect the relationships between weather parameter to predict future weather conditions. Also the researchers discussed ANN is important to climatic change studies because the variation in weather conditions in term of temperature, rainfall and wind speed can be studied using these data mining techniques.

Meganathan et al. [40] developed neural network model for monthly rainfall prediction. The training and testing patterns are prepared as a time-series data of ten months. The numbers of training and testing patterns are 372 and 96, respectively. In the training step, the neural network gives 99.6 % of accuracy and 96.9 % of accuracy in the testing step. The researcher's experimental result shows that it is possible to predict annual rainfall one year ahead with acceptably accuracy. Also According to the researcher's experimental result predictions of the summary rainfall data using back propagation neural network has acceptable accuracy.

El-Shafie et al. [41] explored application of neural network for weather forecasting. The researchers used Artificial Neural Network ANN model and Multi Regression MLR model. A Feed Forward Neural Network FFNN model was developed and implemented to predict the rainfall on yearly and monthly basis. In order to make the comparison between the two models the researchers were used statistical parameters. These parameters include the Root Mean Square Error RMSE, Mean Absolute Error MAE, Coefficient Of Correlation CC and BIAS. The FFNN model has shown better performance than the MLR model. The MLR model revealed a humble prediction performance. The linear nature of MLR model estimators makes it inadequate to provide good prognostics for a variable characterized by a highly nonlinear physics. On the other hand, the ANN model is a nonlinear mapping tool, which potentially is more suitable for rain (nonlinear physics) forecasts.

The results of the researchers study were showed that, the one-day-ahead forecasting network is successful in forecasting the rainfall occurrence of the next day with a success rate of 74.3%. The one-year-ahead forecasting network is highly successful in forecasting the annual rainfall depth of the next year, with a success rate of 80.0%. As the researchers explored that the feed-forward neural network artificial neural networks have been found to be a powerful tool for solving different problems in a variety of engineering applications ranging from pattern recognition to system optimization.

The researchers conclude that The MLR model revealed a humble prediction performance. The linear nature of MLR model estimators makes it inadequate to provide good prognostics for a variable characterized by a highly nonlinear physics. On the other hand, the ANN model is a nonlinear mapping tool, which potentially is more suitable for rain (nonlinear physics) forecasts. The analyzed study cases suggest that, ANN provide better results than the MLR model regarding the statistical criteria used to make the comparison between ANN and MLR models.

Gelder et.al [42] investigated that three forms of hybrid artificial neural networks (ANNs) are used as time series models, namely, the threshold-based ANN (TANN), the cluster-based ANN (CANN), and the periodic ANN (PANN). For the purpose of comparison of forecasting efficiency, the normal multi-layer perceptron form of ANN (MLP–ANN) is selected as the baseline ANN model. Having first applied the MLP–ANN models without any data-grouping procedure, the influence of various data preprocessing procedures on the MLP–ANN model forecasting performance is then investigated. The researchers mainly discussed on selection of the best ANN model. Overall, among the three variations of hybrid ANNs tested, the PANN model performed best Compared with the MLP–ANN fitted to the deseasonalized data, the PANN based on the soft seasonal partitioning performed better for short lead times (%3 days), but the advantage vanishes for longer lead times.

As the researchers proved that the model evaluation results indicate that, among three types of hybrid ANN models, the PANN model performs best. Furthermore, the PANN based on soft seasonal partitions performs better than the PANN based on hard seasonal partitions.

Vaithiyanathanet al. [10] investigated a performance evaluation of J48, Multilayer Perceptron, Naïve Bayes Updatable, and BayesNet classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The researcher sets out to make comparative evaluation of classifiers J48, Multilayer Perceptron, Naive Bayes Updatable, and BayesNet in the context of Labour, Soyabean and Weather datasets. The experiments are carried out using weka 3.6 of Waikato University. The researcher's results showed that the efficiency of J48 and Naive bayes is good.

Taksande et al. [5] investigate the potential applicability of data mining technology to support weather forecasting providers. The researchers were used frequent pattern growth Algorithm and Genetic algorithm. In addition, the FP Growth Algorithm was used to generate decision trees and rules for classifying weather parameters such as maximum temperature, minimum temperature, rainfall, humidity and wind speed in terms of the month and year. Also the authors were applied FP Growth algorithm with the evaluation of MAE, MSE and SD. The researchers were concluding that these calculations are accurate more than the existing model Neural Network (NN) and FP Growth Algorithm, it shows that correct monthly rainfall prediction than Neural Network. Finally the authors prove that FP growth algorithm and genetic algorithm can gives better prediction with higher than 90% accuracy with several population size and crossover probability.

The researchers proved that, the MLP(multilayer perceptron) has performed best ANNs are being used increasingly for the prediction and forecasting of a number of water resources variables, including rainfall, flow, water level and various water quality parameters. The researchers proposed the complex multi-dimensional behaviors of rainfall for a number of stations using several soft computing techniques.

Khan and Hayat [43] explored that weather forecasting for different period of prediction pattern. As the researchers described that some models carry predictions of weather forecasting from real time to Annual period. The researchers were proposed a system that carries out climate prediction using previous state of weather having attribute (Date, Time, Hour, Temperature, Dew Point (DP), Relative Humidity, Wind Speed (WS), Wind Direction (WD), Wind Chill (WC) and Pressure.
The researchers were used different data mining technique to predict weather including K-Nearest Neighbor, Decision Trees and naive Bayes.

Decision Tree has achieved quite promising performance among the other algorithms. The researchers were achieved overall accuracy of 82.62%, sensitivity of 0.8260, FP Rate of 0.029 and MCC of 0.992. The researcher's empirical results show that the proposed model can be very effective and a robust computation model in hourly based weather prediction.

Gaikwad and Nikam[44] proposed general data mining rainfall prediction model. The researchers investigate different rainfall prediction models like weather forecasting, seasonal climate forecasting, global data forecasting and general data mining rainfall prediction model.

The researchers were mainly discussed about different numerical weather prediction model and general data mining techniques for rainfall prediction. As the researchers discussed that data mining tasks provide a very useful and accurate knowledge in a form of rules, models, and visual graphs. This knowledge can be used to obtain useful prediction and support the decision making for different sectors. The researchers conclude that possible to use different data mining techniques on meteorological data set to predict the rainfall on the basis of previous year (Historical) Weather data set.

Chaudhari et al. [15] discussed the application of different data mining techniques applied in various ways to predict, associate, classify and cluster the pattern of meteorological data. The researchers conclude that data mining techniques are the important techniques utilized in all application area related to meteorological data for the prediction and decision making by discovering interesting rules or patterns or groups that indicate the relation between variables.

Mahajan and Mazumdar [45], proposed prediction model for time series rainfall data using 5th order non-linear predicting code. The researchers were used steepest descent algorithm for extraction of appropriate coefficient of rainfall time series data. These coefficients are reinforced in model learning process. The Model has been tested on rainfall data for different training sets by the researchers. The proposed model is capable of forecasting yearly rainfall 1 year in advance. Rainfall estimation accuracy is above 85%. The researchers conclude that the proposed method of N^{th} order polynomial can be useful for rainfall forecasting using previous year's average rainfall data. The trend of rainfall prediction is up to 99% accuracy. Quantitative rainfall prediction error of the model is below 10%.

Related work review Summary

Author	Year	Subject	Data Mining Technique
Chatur et al.	2012	Prediction of future weather	Back propagation & ANN
Mandale&Jadhawar	2015	Data mining technology for weather prediction	ANN and decision tree
Mohod et al.	2013	Hybrid approach for weather prediction	FTP and genetic algorithm
Sethi et al	2014	Data mining technology for rainfall prediction	MLR
Majenthan et al.	2004	Monthly rainfall prediction	ANN
Rupa& Jain	2015	Adaptive inference system for rainfall prediction	ANFIS and ANN
Gaikwad&Nikam	2013	General rainfall prediction model	K-NN, Naïve Bayesian and Multiple Regression and ID3
Atole&Kapgate	2015	Rainfall prediction for crop productivity	K-means and MPR
Sumi et.al	2012	Machine learning for rainfall forecasting	KNN,ANN,MARS and SVR

Kumar	2013	Decision tree for rainfall prediction	Decision tree
Ephrem	2015	Data mining for weather forecasting	J48,PART,MLP
Mahajan&Mazumdar	2014	Time serious rainfall prediction one year ahead	Steepest descent algorithm
Kanth et al.	2014	Hybrid machine learning for weather prediction	Hybrid SVM
Hayat et al.	2015	Real time weather forecasting	K-Nearest Neighbor, Naive Bayes. And Decision Tree
Pitale et al.	2013	Performance evaluation of different data mining algorithm	J48, Multilayer Perceptron and Naïve Bayes
Gelder	2005	Hybrid artificial neural network	MLP
Tafesse	2004	ANN for weather forecasting	ANN
Ephrem	2015	Data mining for weather forecasting	J48, MLP and PART

Table 2: literature review summary on data mining techniques applied on rainfall prediction

CHAPTER THREE

3. Methodology

This chapter provides information regarding the KDD process and the data mining methodology applied in this study. Furthermore, some details regarding the options in data mining techniques are explored.

3.1 Study Area

3.1.1 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. After that from 1946-1949 some preliminary meteorological tasks were carried out by government offices for agricultural sector only especially for locust control. 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.

3.2 Hybrid Methodology

In this thesis Hybrid data mining methodology used, because the hybrid data mining methodology has five iterative steps but CRISP has only three feedback mechanism. In addition SEMMA and KDD have no feedback mechanism. Moreover, Hybrid data mining 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 data mining into practical scenarios. However, hybrid methodology is the extension of CRISP-DM so, hybrid data mining methodology is more complete and better than others and this methodology is used for this research domain.

On the other hand, the development of academic and industrial models has led to the development of hybrid data mining methodology. Hybrid methodology is a six-step KDP model developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research i.e. hybrid data mining methodology is recommended for academic research. In this thesis based on the six step hybrid data mining methodology the following steps with detailed task are performed.

1. Understanding the problem domain: The first step in this process is the identification and the definition of the problem itself. In this study to identify the problem domain the researcher reviewed relevant documents and conducted with domain experts. Also, in this step, research goal is changed to data mining goal. So, to achieve data mining goals the researcher identify the appropriate mining tool. The data mining tools are selected based performance, functionality and auxiliary support. Besides, the researcher proposed ensemble methods as a solution for wide range of rainfall prediction problem.

2. Data Understanding: This is the second step of hybrid data mining methodology. In this study, the researcher used background knowledge to understand the data completeness and data heterogeneity quality problem. Also, in this step the researcher selected the attributes by using information gain ranker.
3. Preprocessing: This step is composed of several sub-steps and requires understanding of both the domain to which statistical methods are applied to improving data quality. In this study, missing value treatment, outlier detection, data reduction and data transformation are performed to preparing the data for mining. Moreover, in this step before mining the pattern's, the researcher balancing the negative and positive class by using WEKA class balancer.
4. Data Mining: In hybrid data mining methodology the fourth step is data mining. In this step, the researcher performed model building by using single algorithm and ensemble method. Also, the model building process performed on Explorer and Experimenter environment on WEKA 3.8. In rainfall forecasting, data mining can be used to build a model that can predict rainfall based on metrological historical data. To build this kind of model, the first step is machine learning algorithms to learn the training dataset, which contains metrological historical data. This learning phase results in a model that can be used to predict rainfall. There are many possible algorithms that can be used in rainfall prediction. In this study based on the objective of the study MLP,J48,IBK and PART algorithm are used; our objective is improve the prediction performance of wide range of rainfall prediction problem. So, MLP is feed forward artificial neural network which is appropriate for nonlinear physics like rainfall.

   On the other hand, IBK is the strongest K-nearest neighbor classifier and J48 decision tree are week machine learning algorithm, to making stronger classifier using ensemble method IBK and J48 decision tree used. PART algorithm is also internally work's based on mathematical equations and appropriate to solve a wide range of rainfall prediction problem.
5. Evaluation: This is the fifth step of hybrid data mining methodology. In this study, the researcher used MAE, RMSE, confusion matrix, ROC curve analysis and accuracy to evaluate the model. MAE and RMSE are used in this study to evaluate the magnitude average error of the models In addition, the researcher extracted potentially useful comprehensive rules for one days advance and one month's advance prediction.

Finally, the cooperation with meteorological prediction experts are most crucial. As a result, the results are interpreted by meteorology prediction experts.

6. Using discovered Knowledge: In this step the researcher developed a prototype to using the discovered knowledge based on the extracted rules. In this study, VB.NET 2010 used for the development of the interface to using the discovered knowledge. The interface is developed based on the concomitant attributed that are used for rainfall forecasting.

CHAPTER FOUR

4. Data Preparation and Model Building

In this Chapter, metrological data sets used in this study were introduced, preprocessed, and analyzed to gain a superficial insight into their attributes. Also for each data set, the various predictive data-mining techniques will be used to build models in order to compare their predictive abilities with each other.

A lot of research has been performed in the rainfall forecasting domain, often with great success. Also using ensembles for weather forecasting is both more accurate and more reliable compared to single model forecasts. ensemble yields more accurate forecast in testing datasets than other methods Lusia & Suhartono [46].This is considered a requirement for weather forecasts, but it is clearly an advantage in all decision making, since a forecast with associated uncertainties is more informative.

The techniques used when performing rainfall forecasting are often a combination of machine learning algorithms and statistical knowledge of weather phenomena. Ensemble approach is presented by Bushara & Abraham [47]. In this study ensemble method has been constructed based on Meta classifier MLP, IBK, J48 and PART algorithm .The models have been evaluated by using mean absolute error and root mean-squared error, accuracy, ROC curve analysis and confusion matrix as performance metrics. Also the researcher evaluate time taken to build the model on percentage split and cross validation to compare and differentiate among the model result.

The aim of this research is to develop new models to forecast rainfall, based on average daily and monthly available meteorological variables data in Addis Ababa. In this study ,we investigate how the technique proposed in ensemble method, could be adapted to solve a wider range of rainfall forecasting problems.

4.1 Understanding of the Problem Domain

Understanding problem domain is working closely with domain experts to define the problem and determine the research goals, identifying key people, and learning about current solutions to the problem changala et.al [31].

4.1.1 Task 1: Define the Problem Domain

Statistical modeling methods have already been developed for producing empirical predictions and also for the calibration of deterministic predictions produced by physically derived dynamical models Coelho [48].

Although, as in the first chapter revealed by Taksande et.al [5] weather is measured by thermometers, rain gauges, barometers, and other instruments but the study of climate relies on statistics. Nowadays, such statistics are handled efficiently by computers. Simple long term summary of weather prediction is not still a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns using data mining technology have tremendous advantage. So far, Ethiopian national metrological agency was not using analysis tools for weather prediction system [4].

However, with the availability of ensemble climate predictions from different coupled models there is clearly the need for development of proposed model for the enhancement of improved rainfall predictions. Thus, the main aim of this research is to investigate the potential applicability of data mining technology in rainfall forecasting using ensemble method.

4.1.2 Data Mining Goals

The first data mining goal for this study investigated the potential application of data mining technology for rainfall forecasting by using ensemble and single algorithm method. To effectively accomplish this task, identifying important variables from the total data Collected was necessary. The identified variables were in turn used to build a model by using classification techniques.

The data preparation and data analysis phases help the identification of important attributes that could serve as an input for the model building. Classification model done by using the most appropriate data mining algorithm for rainfall forecasting, i.e. is decision tree, IBK, MLP and PART algorithm. The classification technique the one which has performs better accuracy will be selected.

4.1.3 Task 2: Data Mining Tool Selection

Nowadays software are flooded in the market, there is a need for selection criteria of software package that should be available for individuals and organizations Bhargava [49]. As the number of software continues to grow and new features added to the new software the selection of the most suitable software package is becoming has tremendous advantage. Wrong or immature decision would result in great loss of time and money.

In this thesis WEKA, R-programing, ORANGE and KEEL are compared to select the best appropriate data mining tool. WEKA is java based open source data mining tool which has collection of data mining algorithms such as lazy, rules, decision trees and so on. On the other hand, R-project is very similar to Matlab and it uses same logic and user can easily load CSV format file and work with functions.

KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. Also, Orange is a component based data mining and machine learning software suite, featuring a visual programming front end for explorative data analysis and visualization, and Python bindings and libraries for scripting.

However, based on performance KEEL is not efficient and ORANGE is not good for report capabilities but R-programing and WEKA are good. On the other hand, based on functionality WEKA is better than KEEL, R-programing & ORANGE. Finally, based on auxiliary support WEKA, ORANGE, KEEL are better than R-programing. Generally, WEKA is better based on performance, functionality and auxiliary support when we compare to ORANGE, KEEL and R-programing. So, WEKA 3.8 is selected to achieve the objective of this research. The description of selection criteria's are described below in the following tables.

№	Criteria	Criteria group	Criteria meaning	WEKA	R-Programing	ORANGE	KEEL
1	Sturdiness	reliability	Capability: run consistently Without crushing.	WEKA can run consistently.	Reliable: can run consistently without crush.	Capability: run consistently Without crushing.	Capability: run consistently Without crushing.
2	Time behavior	Efficiency	Ability: produce results in reasonable amount relative to data size.	Efficient: can produce reasonable output based on data size.	Efficient: can process result reasonable amount of time	can produce reasonable output based on data size.	Efficiency is restricted
3	Report	output	Capability: Standard and customized report from the software package	Standard and customized from software package.	To use R-programing the package must be downloaded	Reporting capabilities are limited to exporting visual representations of data models	Standard and customized from software package

Table 3: Performance Criteria

As, we have seen from the above table WEKA and R-programing is more reliable, efficient and generating the output with reasonable amount of time. However, ORANGE is also reliable and efficient but reporting capabilities are limited or not robust like WEKA. Also, KEEL is reliable and good for reporting but KEEL is not efficient like WEKA and ORANGE.

№	Criteria	Criteria meaning	WEKA	R-Programing	ORANGE	KEEL
1	Algorithmic variety	Availability of adequate variety of mining and algorithms available in the software and customization	WEKA has variety of machine learning and data mining algorithm.	Used for only Naïve bayes ,confusion matrix and data summary	Limited list of machine learning algorithms	Limited algorithms
2	Adaptability	Possible customization in general and for the specific company	Possible to customize	Possible to customize	Possible to customize	Possible to customize
3	Data type Flexibility	Variety of data types that are supported	WEKA can support CSV and ARFF.	R-programing support CSV	CSV	CSV

Table 4: Functionality criteria

In the above table, WEKA is better in algorithmic variety, familiarity to the researcher and the ability to support different data types. However, R-programming, ORANGE and KEEL are not good for algorithmic variety. They have limited list of machine learning and data mining algorithms.

№	Criteria	Criteria meaning	WEKA	R-Programing	ORANGE	KEEL
1	Data Cleansing	Capability to modify spurious values in the data	WEKA have different data cleansing mechanisms	R-Programing cannot used for data cleansing or preprocessing.	ORANGE have different preprocessing mechanisms	KEEL can support preprocessing
2	Data filtering	Capability for selection of subsets based on user defined selection criteria	WEKA used for feature selection	-	Capability to data filtering	Capability to data filtering
4	Record deletion	Capability to delete record which may be biased from entire population of records	WEKA used for deletion of unnecessary records	-	Capability to support deletion of unnecessary records	Capability to support deletion of unnecessary records
5	Handling blanks	Capability to handle blanks and replace the entries	WEKA can handle blanks and replace missing value	-	Capability to missing value replacement	Capability to missing value replacement

Table 5: Auxiliary task support

As we compare WEKA, R-programing, ORANGE and KEEL based on auxiliary task support WEKA, ORANGE and KEEL are better for data preprocessing. However, R-programing is not supporting data preprocessing. In general, based on the above three criteria WEKA is better and selected for this research domain.

4.1.4 Task 2: Identifying the key Domain Expert

Domain experts are consulted to have insight into the problem domain. As a result, prediction experts are the key domain expert and the researcher constitute from National meteorological Agency experts on the basis of the insight gained from discussion and review of relevant documents, a clear understanding of the data was achieved. In addition, perdition experts are consulted for result interpretation.

4.1.5 Task 3: Learning about Current Solutions to the Problem

In today's information age, rainfall prediction has drawn much attention for research community because it helps in safeguarding human life and their wealth. Beyond that, it is useful in effective prediction of natural calamities, agricultural yield growth, air traffic control, marine navigation, forests growth & military purposes. Also literatures shows that machine learning Algorithms proved to be good than the existing techniques traditional statistical methods N.Rajasekhar [50].

In addition, weather forecasting plays a significant role in meteorology and which remains a formidable challenge because of its data intensive and frenzied nature. Generally two methods are used to forecast weather: the empirical approach and the dynamical approach. The first approach is based on the occurrence of analogues and it is often referred to as analogue forecasting. This approach is useful in predicting local scale weather if recorded cases are plentiful. The second case is based upon equations and forward simulations of the atmosphere and is often referred to as computer modeling.

In this study, Artificial Neural Networks (ANN), PART algorithm, IBK and Decision Trees (DT) were used to analyze meteorological data gathered in-order to develop classification rules for the application of data mining techniques in weather prediction. Weather parameters over the study period and for the prediction of future weather conditions using available historical data. The targets for the prediction are those weather changes that affect us daily like changes in minimum and maximum temperature, rainfall, sunshine and wind speed.

4.1.5.1 Using Ensemble Method for Rainfall Forecasting

In statistics and machine learning, ensemble classifiers are used to improve predictive results obtained by using its constituent algorithms. An ensemble combines a number of 'weak' learners and aggregates their results into one 'strong' learner. An ensemble is a 'supervised' learning algorithm itself. This is because it can be trained on labeled data and later used to predict labels for previously untested data. Ensembles thus work in a two-step process. Firstly, a set of classifier methods are learnt. The results of these methods are combined to obtain higher accuracy. Ensemble many algorithms has many advantages over using a singular classification method. Most important advantage is that the results of the classification are less dependent on the peculiarities of a single algorithm.

As a result in this study we, investigate the potential applicability of data mining technology using ensemble method to solve wider range of rainfall forecasting problems. So basically, ensemble ensures a more generic inspection of data at hand. Another advantage is that ensemble different methods makes the entire model more expressive and unbiased than a single model[46].

4.2 Data Understanding

According to changala et al. [31] data understanding focus on collecting sample data and deciding which data type and format size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, missing values, plausibility of attribute values. Finally, in this step includes verification of the usefulness of the data with respect to the DM goals.

4.2.1 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.

4.2.2 Description of Raw Data Quality

4.2.2.1 NMA Database

This section addresses the description of the structure of the data, by introducing exploration and verification of the data quality and illustrating selection of data sets.

Ethiopian National Meteorological Agency gather information regarding meteorological information in Ethiopia. One of the initial problems in checking the data quality was the records from 1985-1990 have not the attributes sunshine, wind speed and humidity but the record from 1990-2015 have those attributes. With this data heterogeneity problem in mind and with the contemporary objective of analyzing initially thousands of meteorological records from NMA database records from (1990 – 2015) with 9553 instances and 10 attribute were selected for the initial runs of the data mining methods. In this period the coding system was consistent and allowed avoiding potential problems related to the heterogeneity of the data.

Furthermore, the source data employed for this research purpose is NMA data, which was collected from bole station based on different meteorological variables. The aim of this study is to create a model based on secondary data that was selected on the base meteorological data in the district of Bole station in Addis Ababa. The entire attributes in the original dataset were not concerned for this experimentation. Thus, only relevant attributes were considered to achieve the objectives of the study. Sample data was collected from the whole dataset on which feature selection and preprocessing is conducted, the next task is comprehensive assessment of distal and proximal determinants based on their significance level with respect to rainfall prediction.

Number	Field Name	Data Type	Description
1	Elevation	Numeric	Elevation considered
2	Year	date	Year considered
3	Month	date	Month considered
4	day	date	Day considered
5	rainfall	Numeric	Rain fall of the day
6	Relative humidity	Numeric	Relative humidity of the day
7	sunshine	Numeric	Sunshine of the day
8	Temp Max	Numeric	Max temp of the day
9	Temp Min	Numeric	Min temp of the day
10	Wind speed	Numeric	Wind speed of the day

Table 6: List of attributes in the initial data set

4.2.3 Data Formatting

At this step the researcher changes the data into a format which was suitable for the data mining 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.

4.2.4 Attribute Selection

In data mining technology not all attributes are relevant. Selection of relevant attribute is necessary for data mining, 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 [51]. 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.

4.2.5 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.

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

Table 7: Attribute Rank on all input Data using information gain ranker

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

Number	Field Name	Data Type	Description
01	Year	date	Year considered
02	Month	date	Month considered
03	rainfall	nominal	Rain fall of the day
04	Relative humidity	Numeric	Relative humidity of the day
05	Sunshine	Numeric	Sunshine of the day
06	Temp Max	Numeric	Max temp of the day
07	Temp Min	Numeric	Min temp of the day
08	Wind speed	Numeric	Wind speed of the day

Table 8: Selected attribute from the NMA database

4.3 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 data mining 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.

4.3.1 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 [52].

In this thesis 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.

4.3.2 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 [53].

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.

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

Table 9: Univariate Statistics of metrological dataset

4.3.3 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[54] So, higher and lower extreme values are discard from independent variables Appendex1. Finally after the outliers are detected only 9543 instances with 8 attributes are used for the experiment.

4.3.4 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 [53].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.

4.3.5 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.

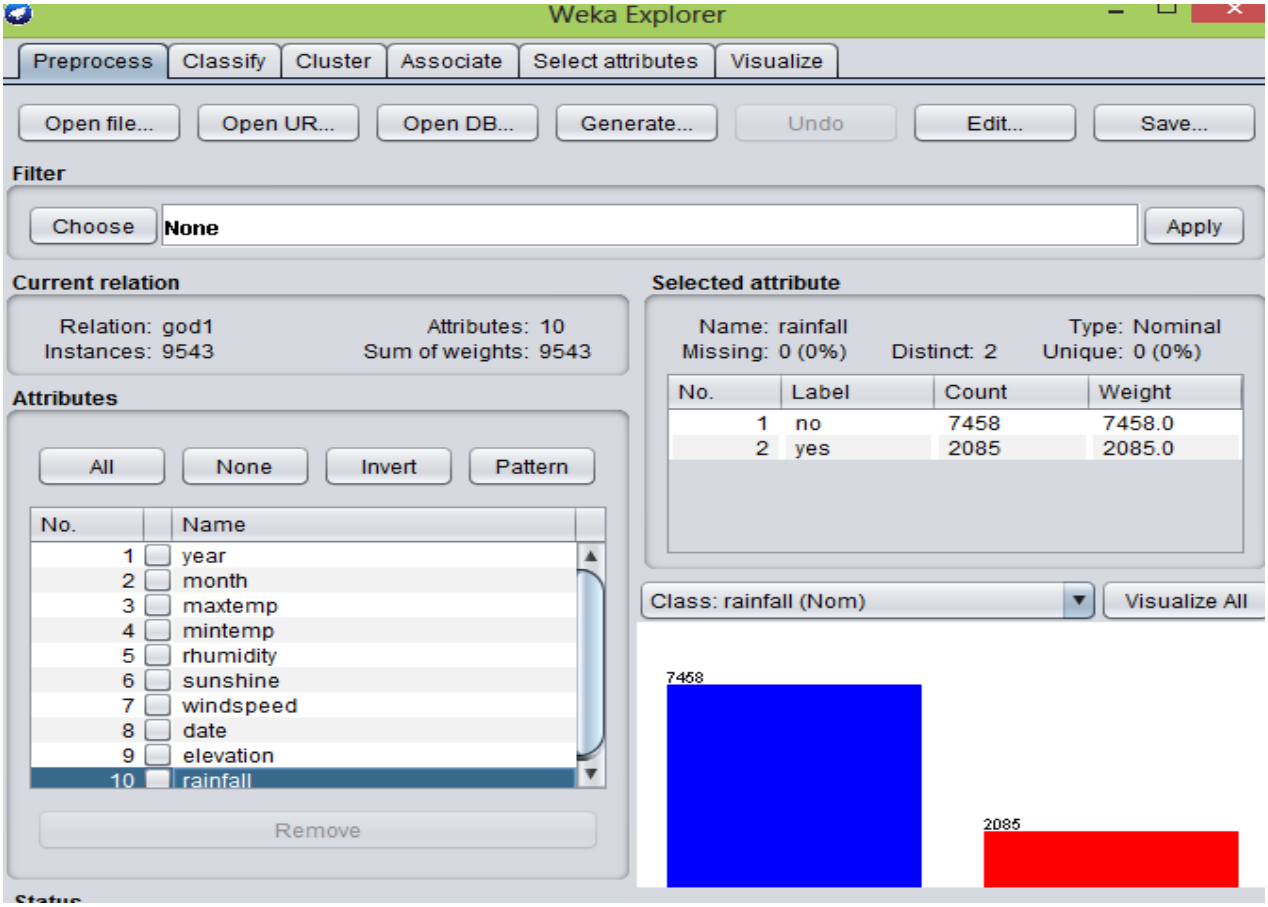


Figure 6: Class attribute before balance for one day ahead prediction using all attributes

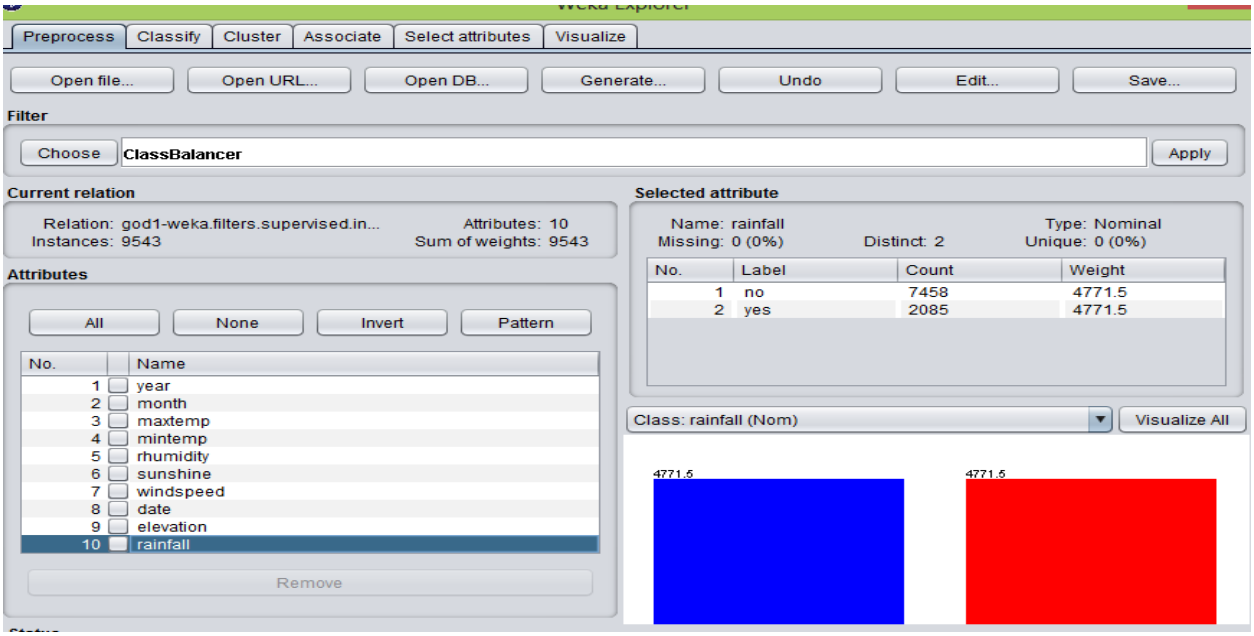


Figure 7: Class attribute after balance for one day ahead prediction using all attributes

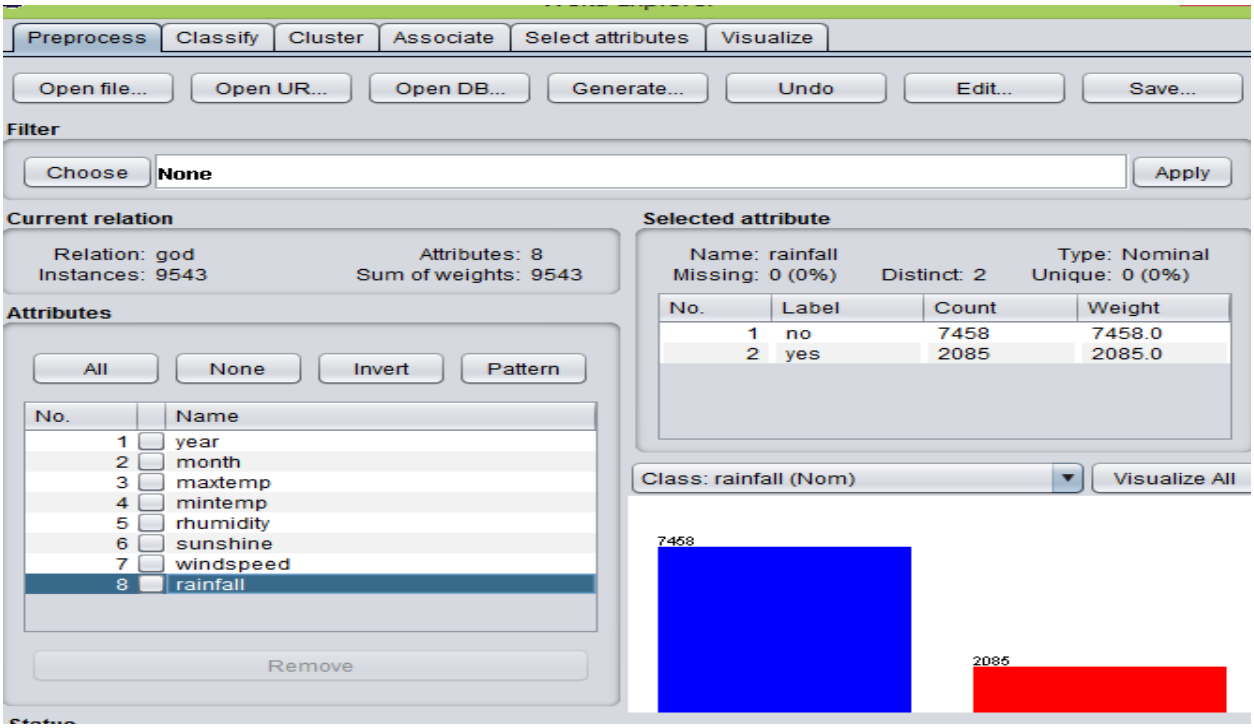


Figure 8: Class attribute before balance for one day ahead prediction using selected attributes

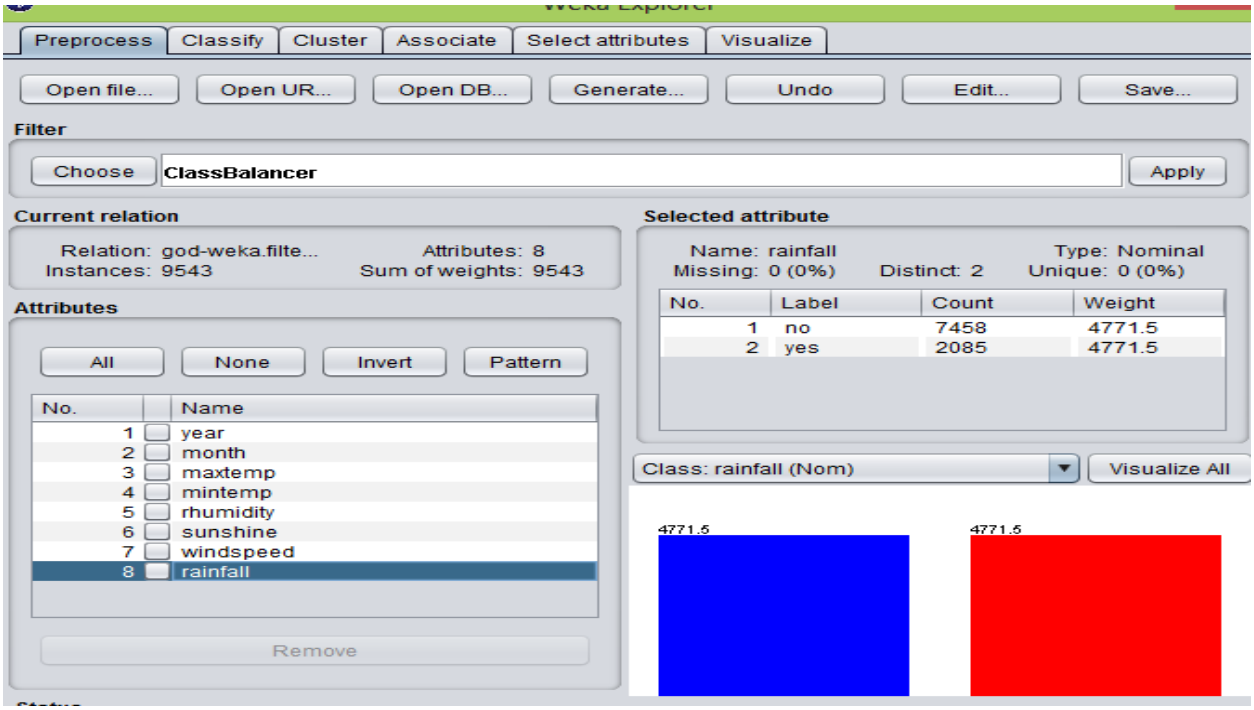


Figure 9: Class attribute after balance for one day ahead prediction using selected attributes

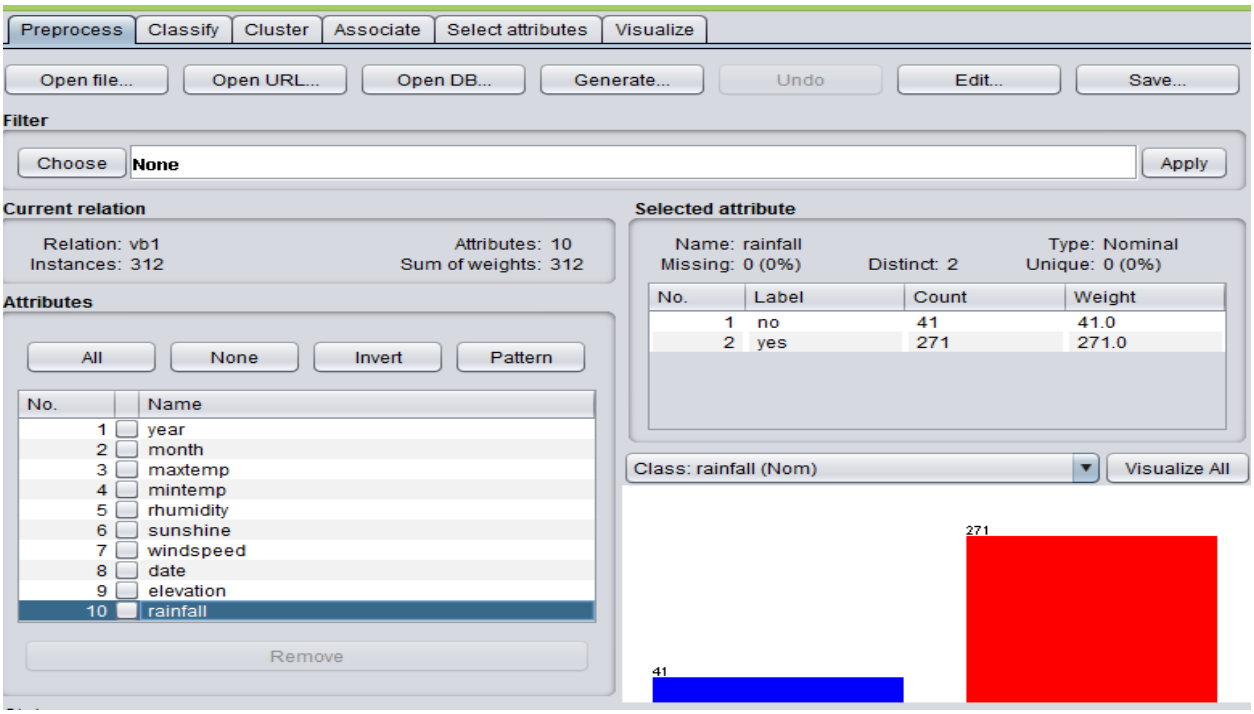


Figure 10: Class attribute a before balance for one month ahead prediction using all attributes

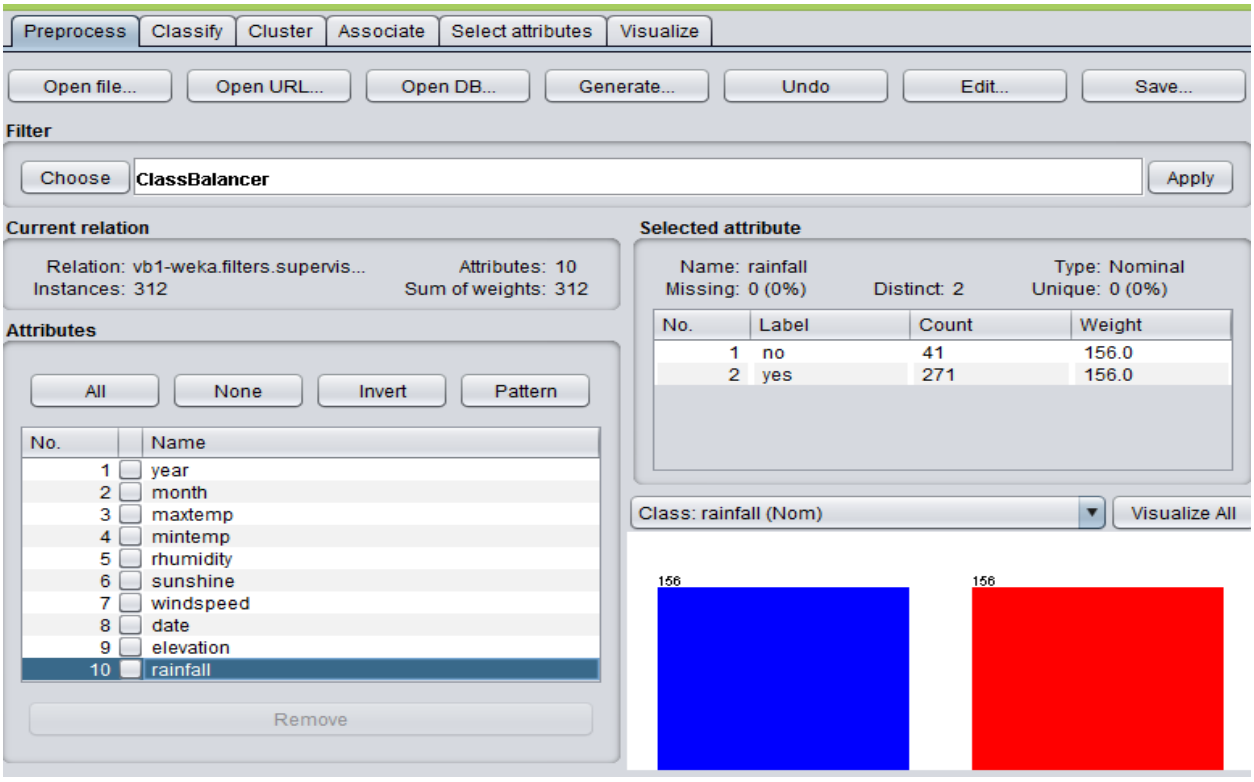


Figure 11: Class attribute after balance for one month ahead Prediction Experiment using all attributes

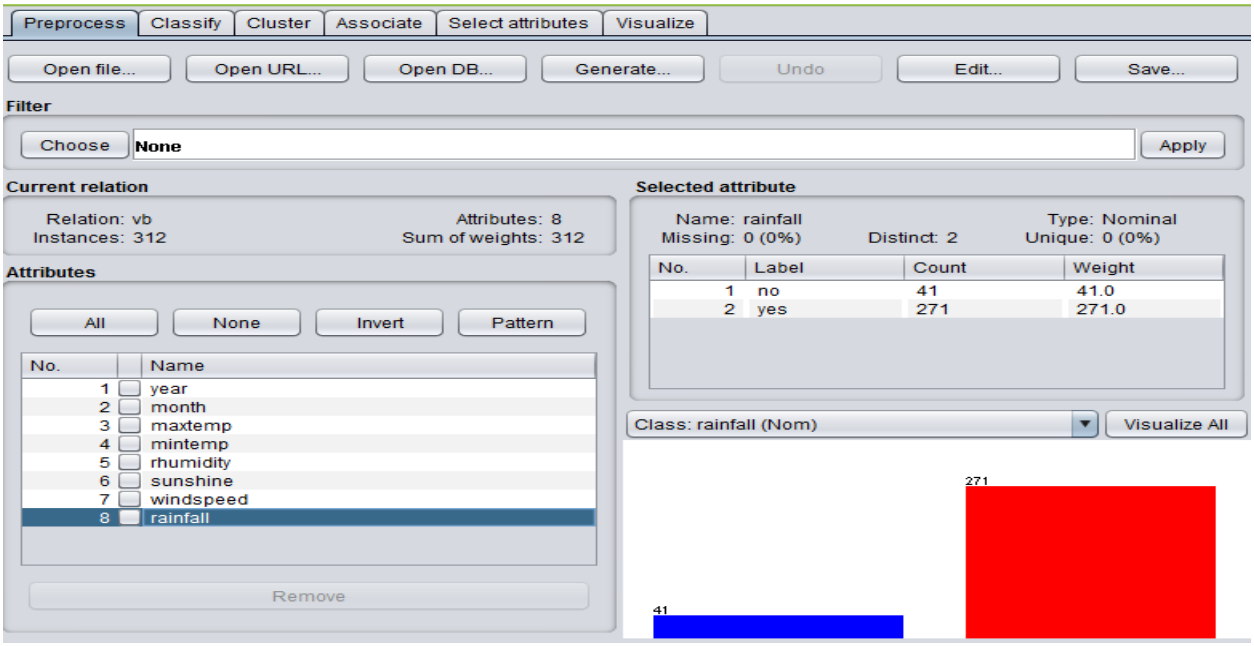


Figure 12: Class attribute before balance for one month ahead prediction using selected attributes

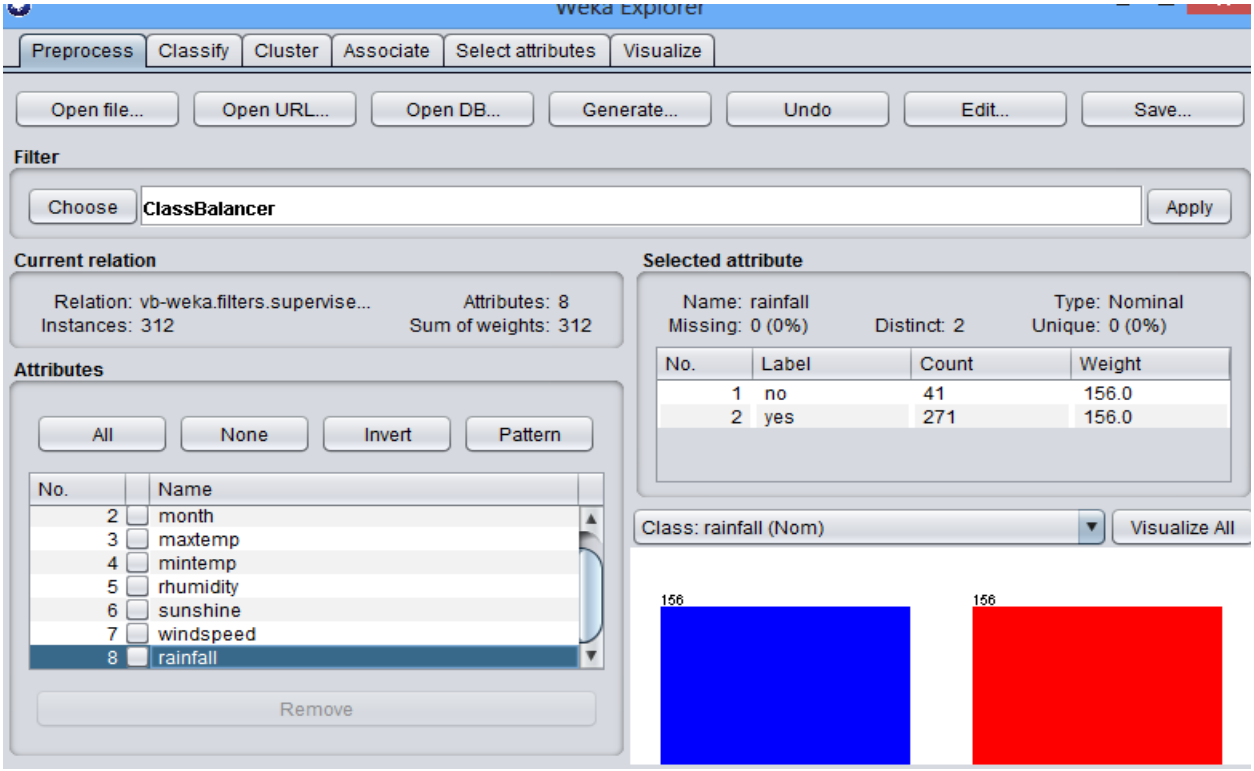


Figure 13: class attribute after balance for one month ahead Prediction Experiment using selected attributes

4.4 Data Mining

The data miner uses various data mining methods to derive knowledge from preprocessed data.

There are different tasks performed during the data mining phase, the major once are selection of modeling techniques and building a model.

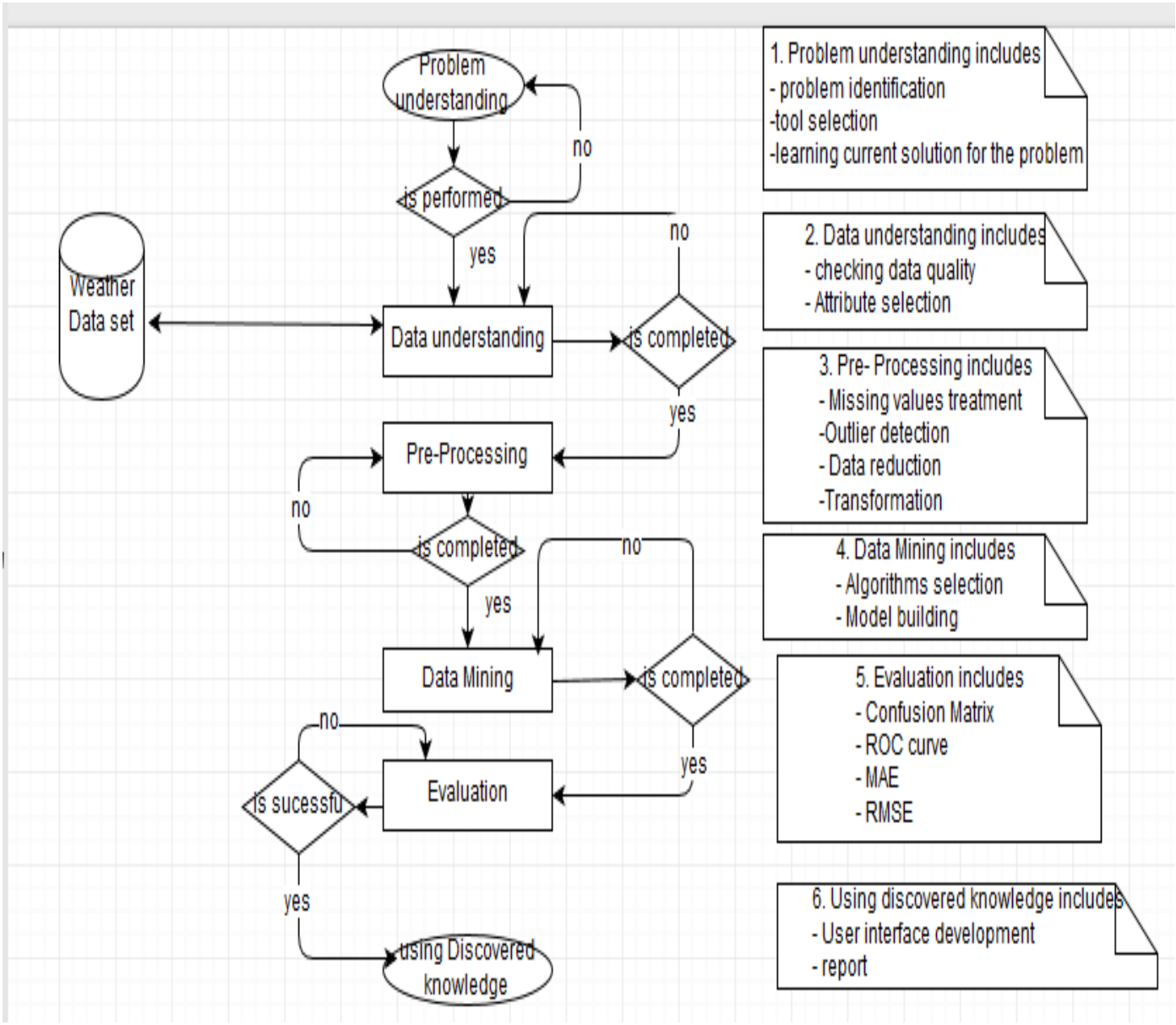


Figure 14: General Rainfall Prediction work flow diagram

4.4.1 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 [55]. 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.

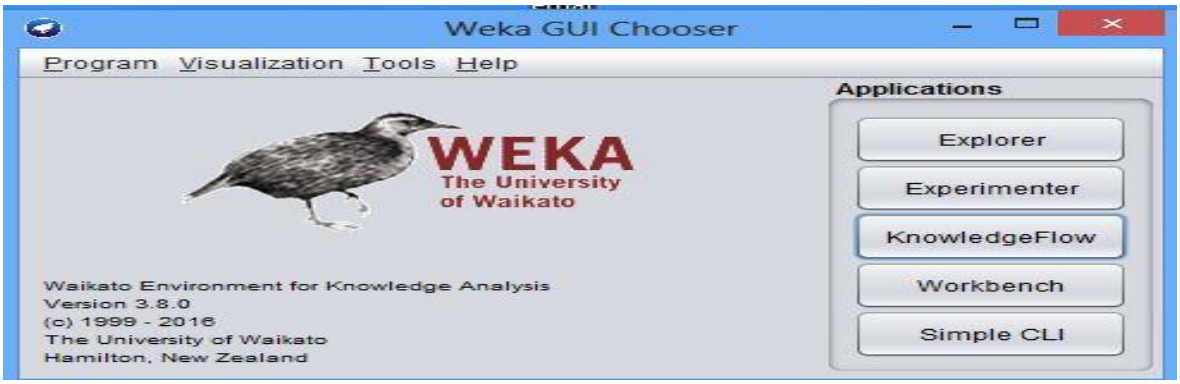
4.4.2 Experiment and Result Interpretation

4.4.2.1 Overview of Experimentation

In this study different experiments were conducted using various data mining 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 and 10 all attributes applied for experimentation.

4.4.3 Experimental setup for Mining

Start WEKA. This may involve finding it in program launcher or double clicking on the weka.jar file. This will start the WEKA GUI chooser.



During experimentation binary split = "yes" number of iteration = 10 with different number of folds and 75% split. On the other hand, all the default parameters that already set in the WEKA were used.

4.4.3.1 Sample Experimental set up for Ensemble Method

Parameters	Descriptions	Parameter type
Bach size	The preferred number to preprocess if batch prediction is performed.	numeric
debug	If set to true, classifier may out put additional info. To the console.	Boolean
Do not check capabilities	If set classifier capabilities are not checked before classifier is built.	Boolean
Num. decimal places	The number of decimal places used for the output number in the model	numeric
Num. iterations	The number of iterations to be performed.	numeric
seed	The seed used for randomizing the data when reduced-error pruning is used.	numeric
Use resampling	Weather resampling instead of wreathing	Boolean
Weight threshold	Weight threshold for weight pruning.	numeric

Table 10: Experimental Setup on Addaboostm1

4.4.4 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.

4.4.5 Model Performance Comparison for one Day ahead Prediction

4.4.5.1 Model Performance comparisons using all attributes with single Algorithms

Algorithm	Accuracy of models with different Number of folds								
	2	3	4	5	6	7	8	9	10
J48	91.43	92.22	92.65	92.79	93.07	93.18	93.24	93.39	93.31
IBK	91.25	92.71	93.37	93.70	93.88	94.07	94.20	94.30	94.31
MLP	89.47	89.81	89.92	90.02	90.09	89.99	90.17	90.09	90.23
PART	90.29	90.26	90.61	90.60	90.86	90.61	90.64	90.76	90.87

Table 11: Single algorithms with all attributes using cross validation for one day ahead prediction

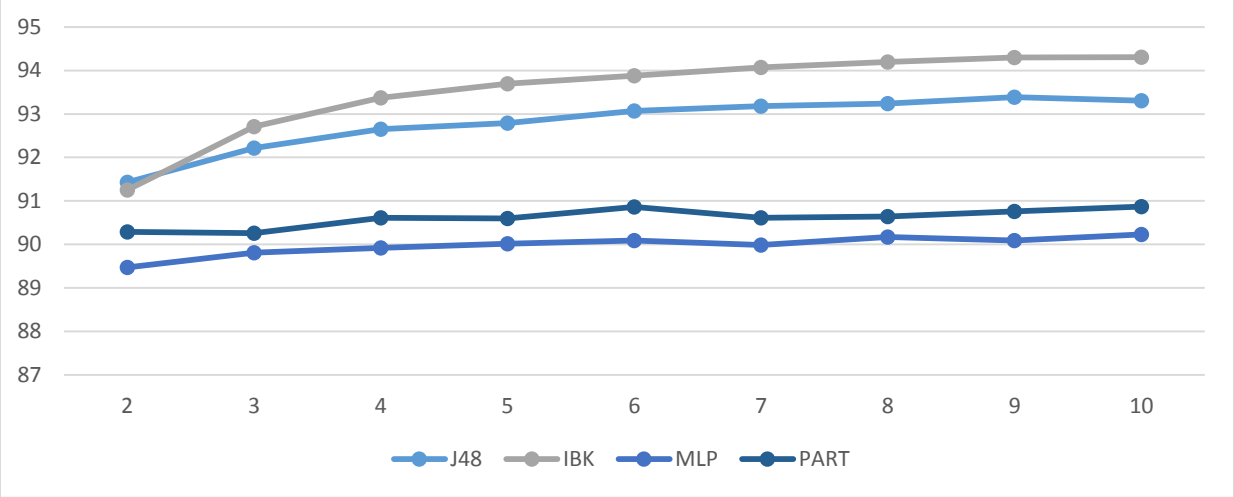


Figure 15: Single algorithms learning progress on all attributes and 75% split for a day ahead prediction

Algorithm	Accuracy of models with 75% split
J48	92.70
IBK	93.62
MLP	90.00
PART	90.93

Table 12: Single algorithms with all attributes using 75% split for one day ahead prediction

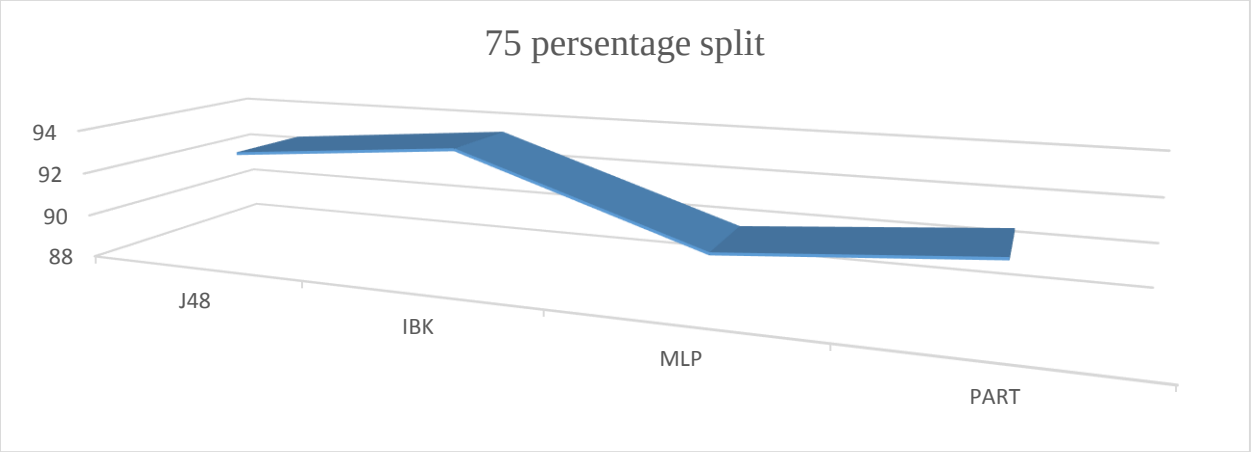


Figure 16: Single algorithms learning progress on all attributes and cross validation for a day ahead Prediction

4.4.5.2 Model Performance Comparisons using Selected Attributes with Single Algorithms

Algorithm	Accuracy of models with different Number of folds								
	2	3	4	5	6	7	8	9	10
J48	91.29	92.07	92.51	92.59	92.92	92.89	93.04	93.06	93.09
IBK	91.81	93.26	93.85	94.21	94.41	94.56	94.67	94.78	94.83
MLP	89.32	89.73	89.75	89.77	89.88	89.99	89.99	89.60	89.91
PART	90.05	90.06	90.21	90.34	90.57	90.48	90.64	90.66	90.54

Table 13: Single algorithms with selected attributes using cross validation for one day ahead prediction

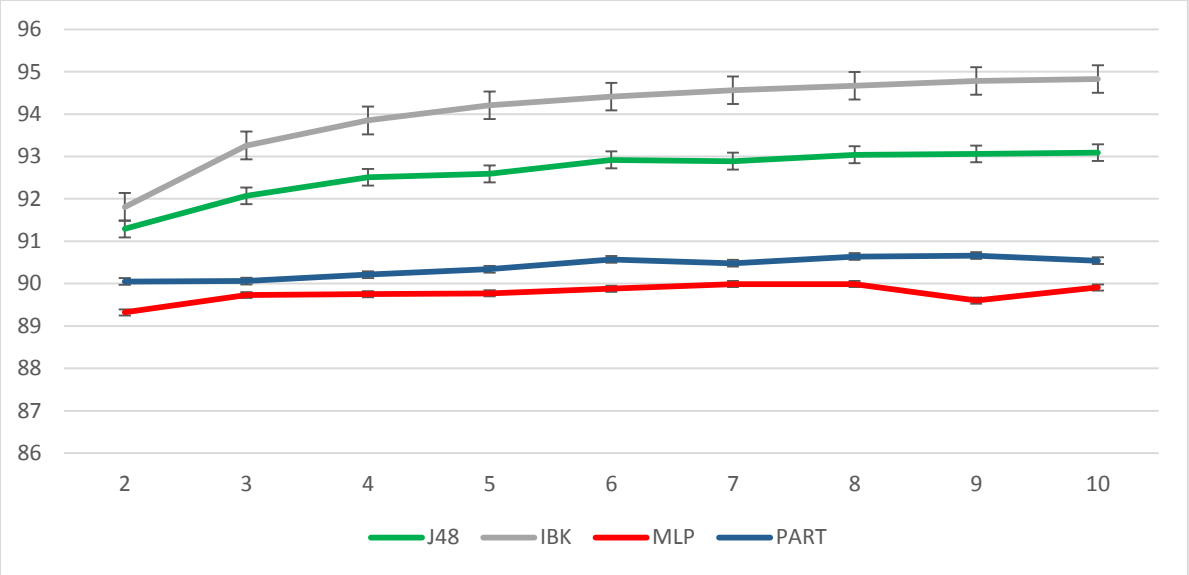


Figure 17: Single algorithms learning on selected attributes and cross validation for a day ahead prediction

Algorithm	Accuracy of models with 75% split
J48	92.65
IBK	94.10
MLP	90.02
PART	90.26

Table 14: Single algorithms with selected attributes using 75% split for one day ahead prediction

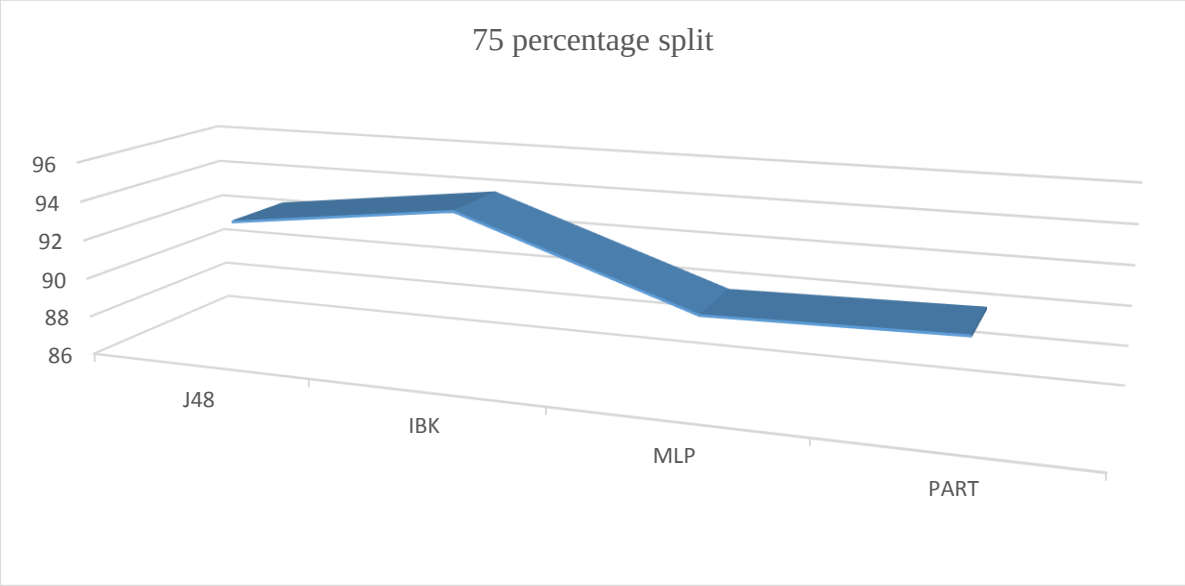


Figure 18: Single algorithms learning progress on selected attributes and 75% split for a day ahead prediction

4.4.5.3 Model Performance Comparison using all Attributes with Ensemble Methods

Method	Algorithm	Accuracy of models with different Number of folds								
		2	3	4	5	6	7	8	9	10
Boosting	J48	91.43	92.22	92.65	92.79	95.19	93.18	93.24	93.39	94.49
	IBK	91.25	92.71	93.37	93.70	93.88	94.07	94.20	94.30	94.31
	MLP	89.47	89.81	89.92	90.02	90.11	89.99	90.17	90.09	90.23
	PART	90.29	90.26	90.61	90.60	95.11	90.61	90.64	90.76	90.87
Bagging	J48	92.87	93.70	94.23	94.43	94.51	94.64	94.79	94.75	94.85
	IBK	90.73	92.13	92.72	93.03	93.19	93.35	90.67	93.57	93.63
	MLP	90.09	90.40	90.55	90.63	90.68	90.63	93.52	90.64	90.67
	PART	92.20	92.84	93.17	93.28	93.36	93.55	93.60	93.62	93.66
Staking	J48 & IBK	91.56	92.73	93.30	93.59	93.77	93.95	94.07	94.23	94.14

Table 15: Ensemble method with all attributes using cross validation for one day ahead prediction

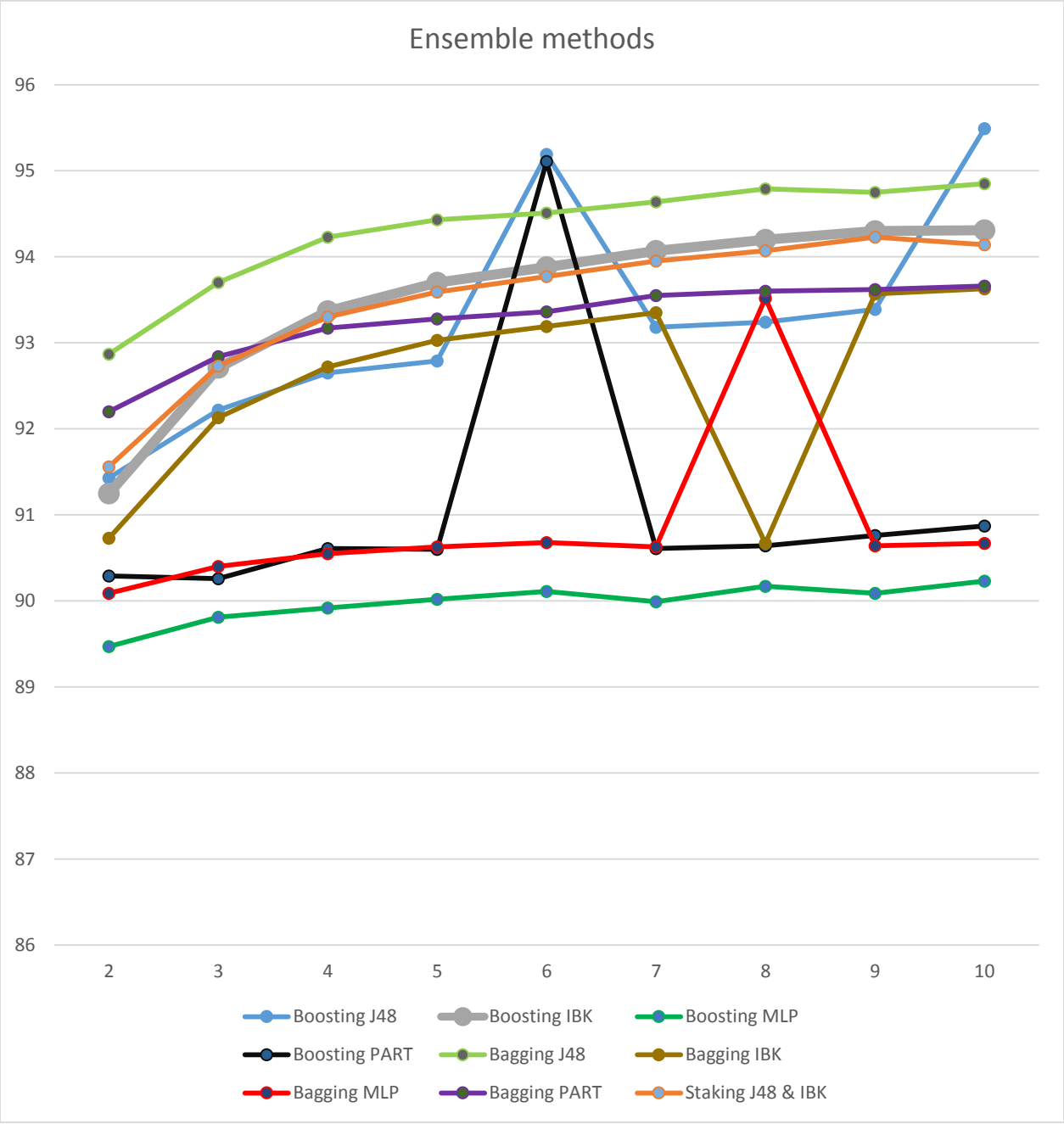


Figure 19: Ensemble learning progress on all attributes and cross validation for a day ahead prediction

Method	Algorithm	Accuracy of models with 75% split
Boosting	J48	92.70
	IBK	93.62
	MLP	90.00
	PART	90.93
Bagging	J48	94.17
	IBK	92.96
	MLP	90.64
	PART	93.26
Staking	J48 & IBK	93.41

Table 16: Ensemble methods with all attributes using 75% split for one day ahead prediction

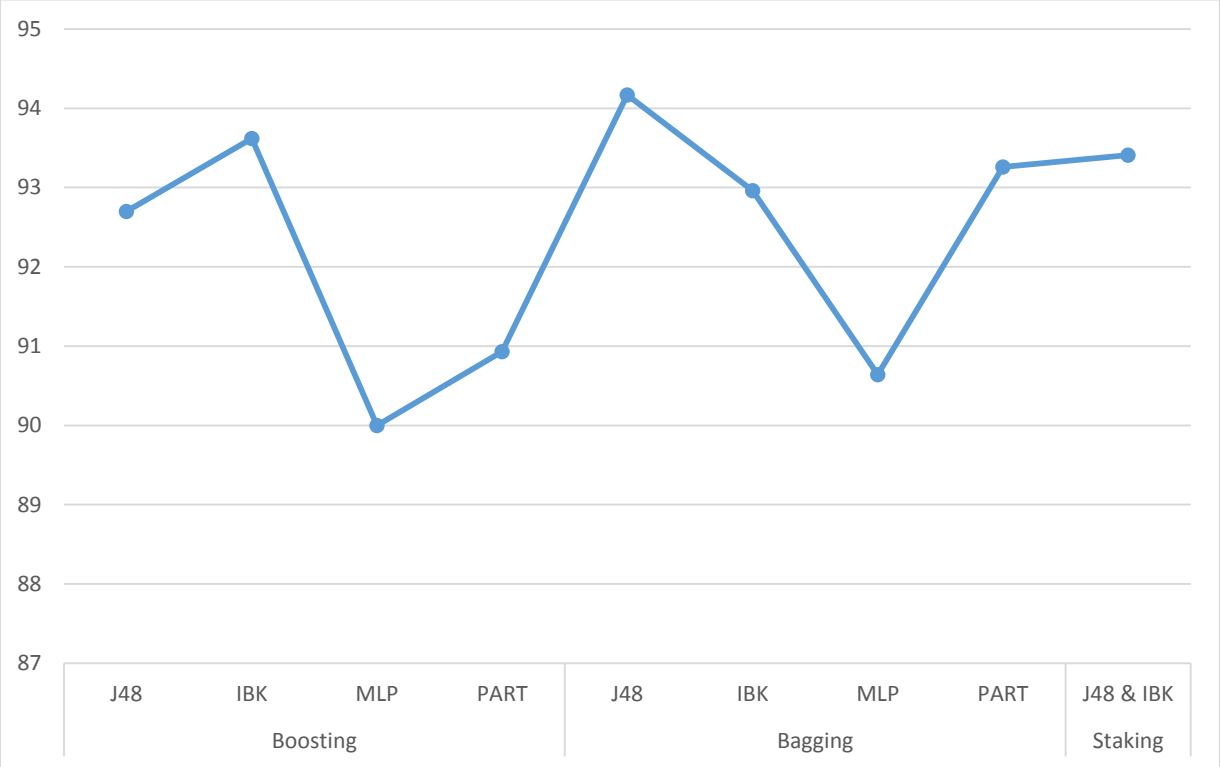


Figure 20: Ensemble learning progress on all attributes and 75% split for a day ahead prediction

4.4.5.4 Model Performance Comparisons using Selected Attributes with Ensemble Method

Method	Algorithm	Accuracy of models with different Number of folds								
		2	3	4	5	6	7	8	9	10
Boosting	J48	92.91	94.09	94.67	94.91	95.13	95.31	95.33	95.34	95.44
	IBK	91.81	93.26	93.85	94.21	94.41	94.56	94.67	94.78	94.83
	MLP	89.35	89.74	89.76	89.79	89.88	89.97	89.98	89.59	89.90
	PART	93.13	94.24	94.81	94.93	95.17	95.41	95.37	95.43	95.46
Bagging	J48	92.65	93.50	94.01	93.99	94.31	94.41	94.50	94.48	94.57
	IBK	91.25	92.64	93.25	93.53	93.76	93.92	94.02	94.12	94.15
	MLP	90.03	90.25	90.41	90.38	90.40	90.42	90.43	90.42	90.47
	PART	91.87	92.39	92.70	92.73	92.76	93.01	92.97	93.07	93.03
Staking	J48 & IBK	91.71	92.93	93.51	93.91	94.15	94.27	94.39	94.50	94.56

Table 17: Ensemble methods with selected attributes using cross validation for one day ahead prediction

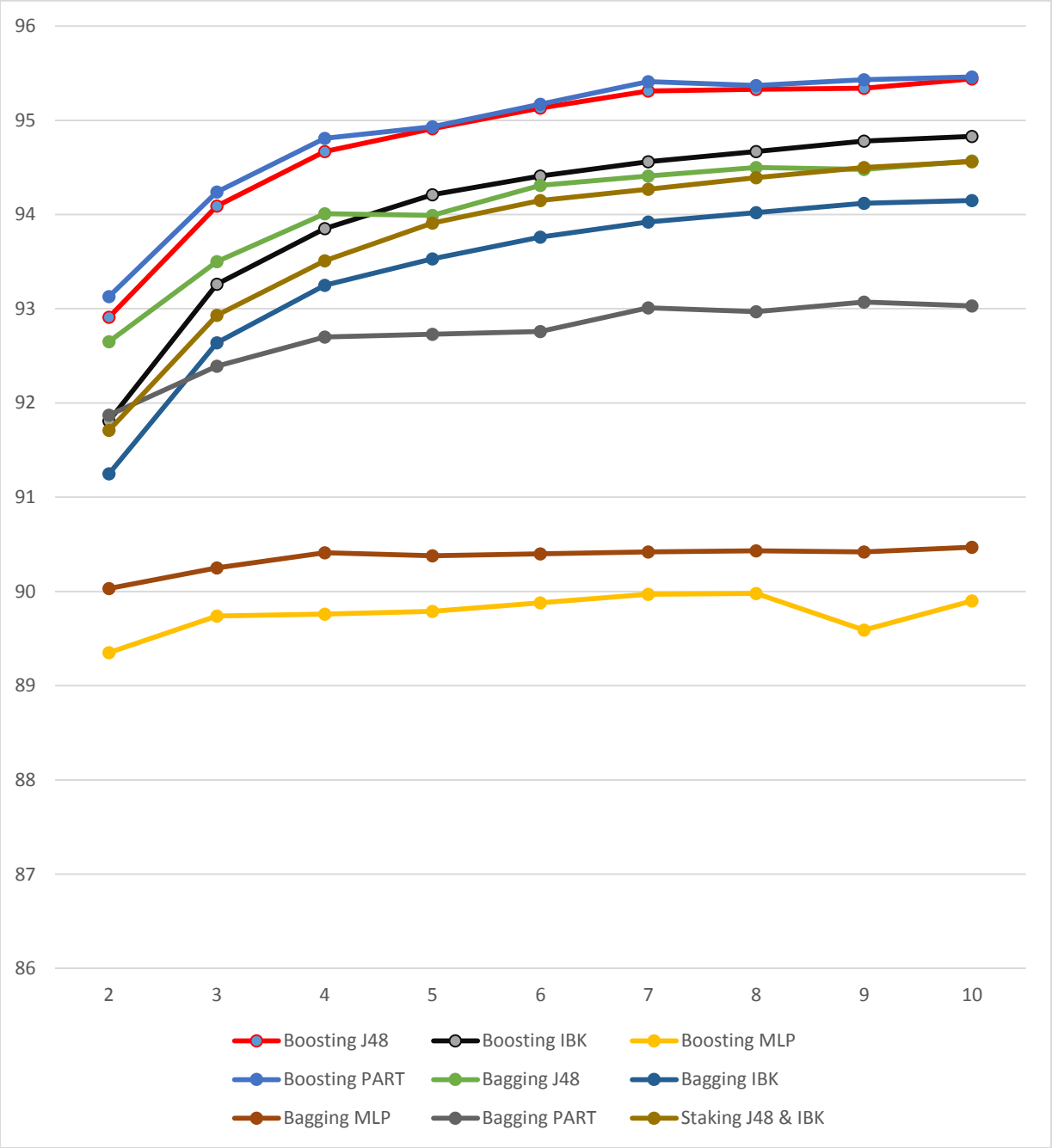


Figure 21: Ensemble learning progress on selected attributes and cross validation for a day ahead prediction

Method	Algorithm	Accuracy of models with 75% split
Boosting	J48	94.84
	IBK	94.10
	MLP	90.00
	PART	95.00
Bagging	J48	93.91
	IBK	93.52
	MLP	90.51
	PART	92.64
Staking	J48 & IBK	94.01

Table 18: Ensemble methods with selected attributes using 75% split for one day ahead prediction

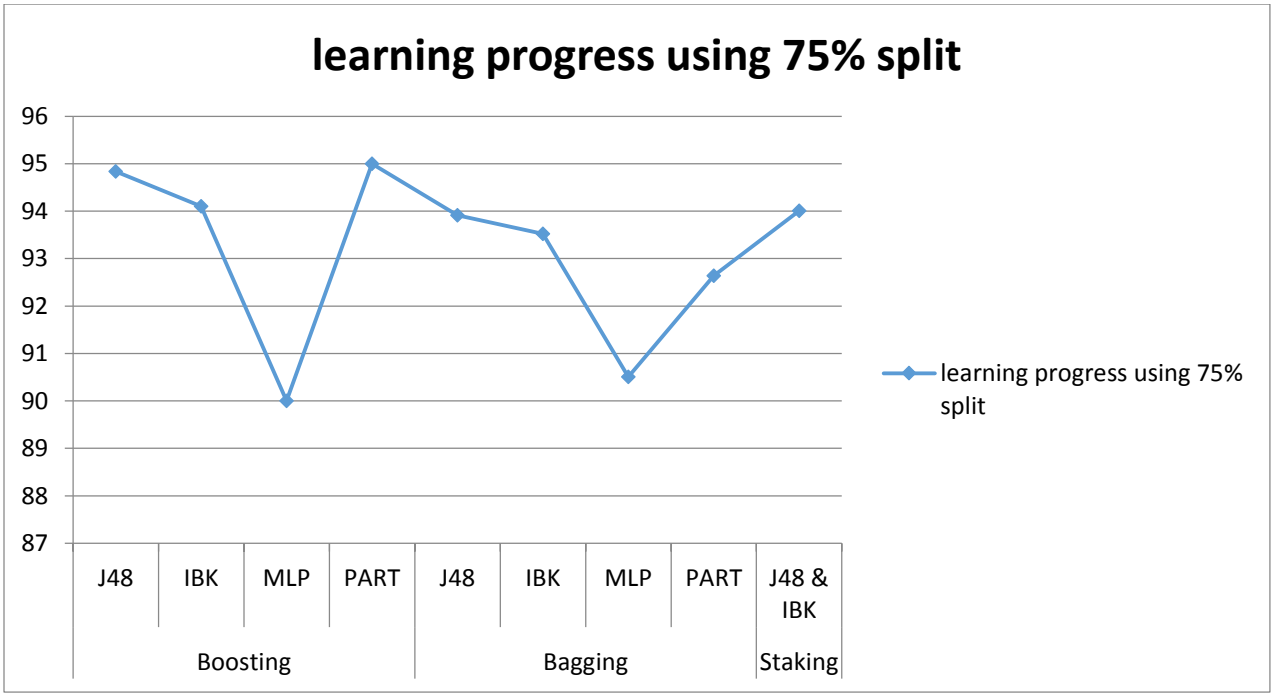


Figure 22: Ensemble method learning progress on selected attributes and 75% split for a day ahead prediction

4.4.6 Model Performance Comparison for one Month ahead Prediction

4.4.6.1 Single Algorithms Performance Comparison Prediction after one Month ahead using all Attributes

Algorithm	Accuracy of models with different Number of folds								
	2	3	4	5	6	7	8	9	10
J48	88.72	89.68	90.93	91.15	91.03	92.57	92.34	91.86	92.63
IBK	87.05	88.24	87.79	88.30	88.91	89.52	88.78	88.82	88.94
MLP	86.86	86.79	86.57	86.73	86.38	86.58	86.83	86.10	86.54
PART	88.17	90.45	90.77	91.12	91.41	91.19	91.22	91.63	92.53

Table 19: Single algorithms with all attributes using cross validation for one month ahead prediction

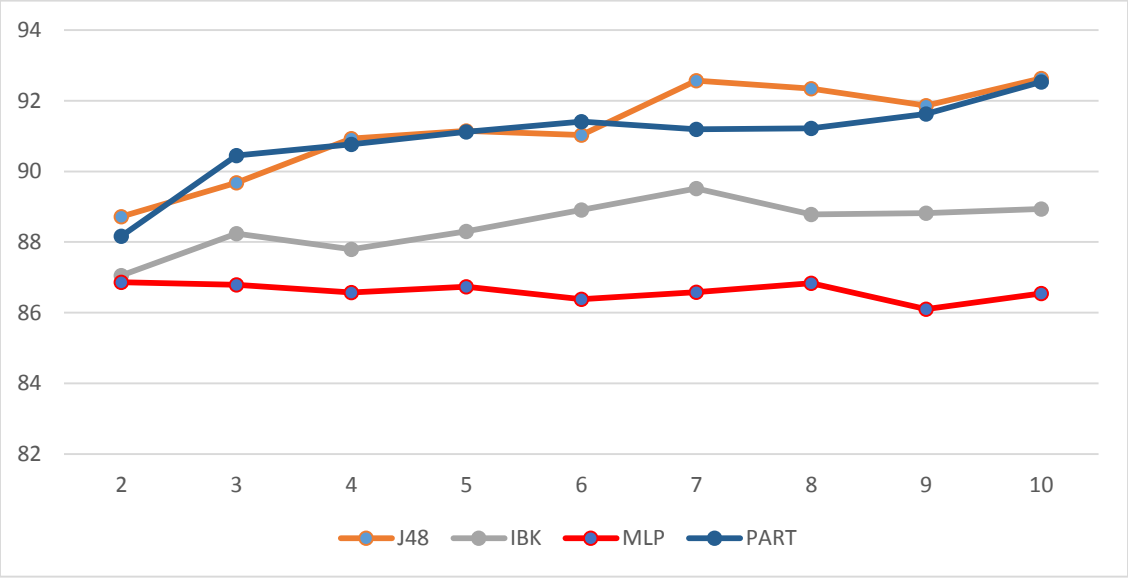


Figure 23: Single algorithms learning on all attributes and cross validation for a month ahead prediction

Algorithm	Accuracy of models with 75% split
J48	92.44
IBK	90.90
MLP	86.67
PART	90.51

Table 20: Single algorithm with all attributes using 75% split for one month ahead prediction

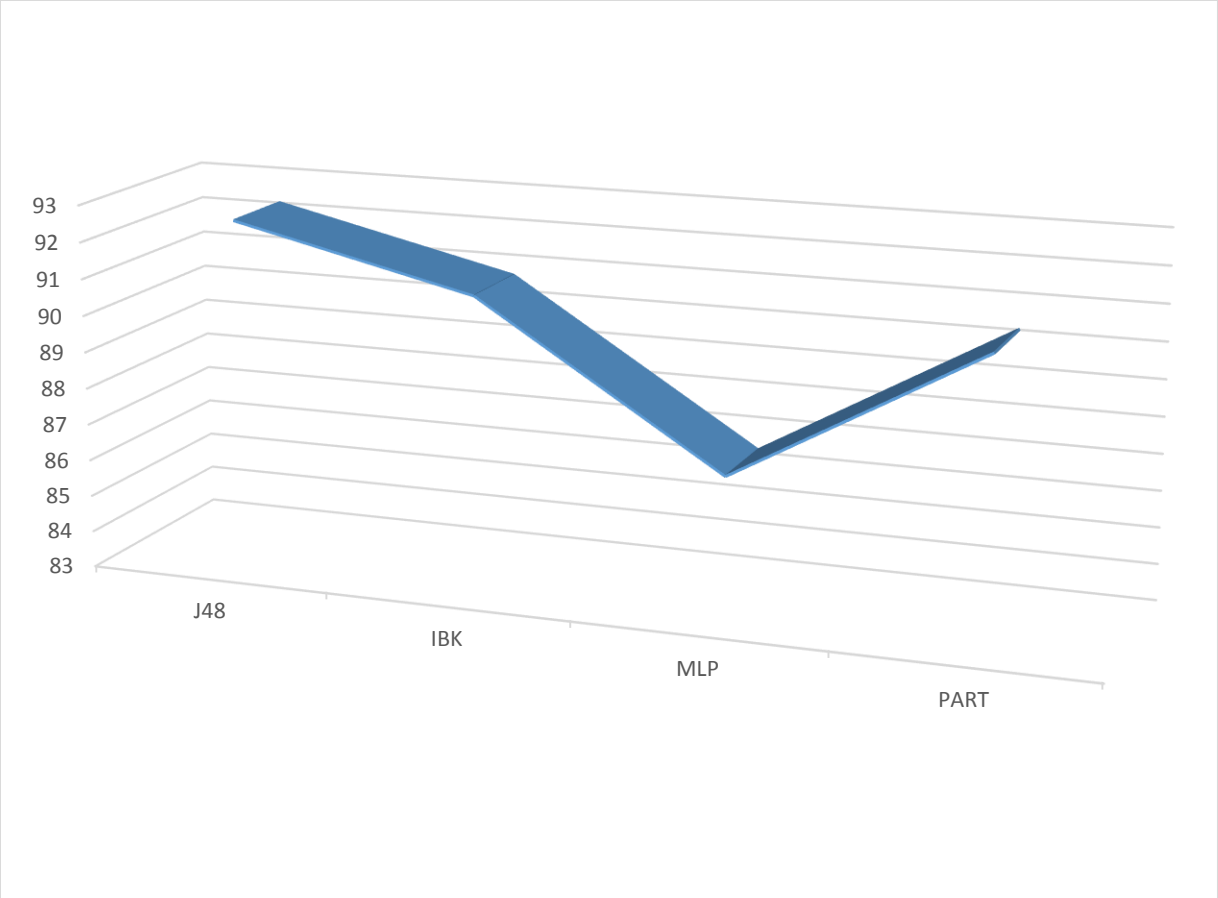


Figure 24: Single algorithms learning progress on all attributes and 75% split for a month ahead prediction

4.4.6.2 Ensemble Algorithms Performance Comparison Prediction after one Month ahead using all Attributes

Method	Algorithm	Accuracy of models with different Number of folds								
		2	3	4	5	6	7	8	9	10
Boosting	J48	91.89	93.46	93.69	93.87	94.39	93.95	94.68	94.39	94.84
	IBK	87.05	88.24	87.79	88.30	88.91	89.52	88.78	88.82	88.94
	MLP	86.12	86.35	85.77	86.67	86.54	85.90	87.08	85.94	86.83
	PART	91.44	93.08	93.56	94.29	94.42	94.88	95.10	94.04	95.35
Bagging	J48	90.35	90.96	91.44	92.44	91.89	92.48	92.72	91.79	92.47
	IBK	86.86	88.30	87.76	88.52	88.94	89.65	88.81	88.75	89.03
	MLP	86.79	86.76	86.54	86.77	86.41	86.38	86.28	86.62	86.41
	PART	90.93	91.54	92.05	92.85	92.63	93.18	92.69	92.98	93.01
Staking	J48 & IBK	88.11	89.81	90.48	91.70	91.70	92.05	91.67	91.73	92.21

Table 21: Ensemble method with all attributes using cross validation for one month ahead prediction

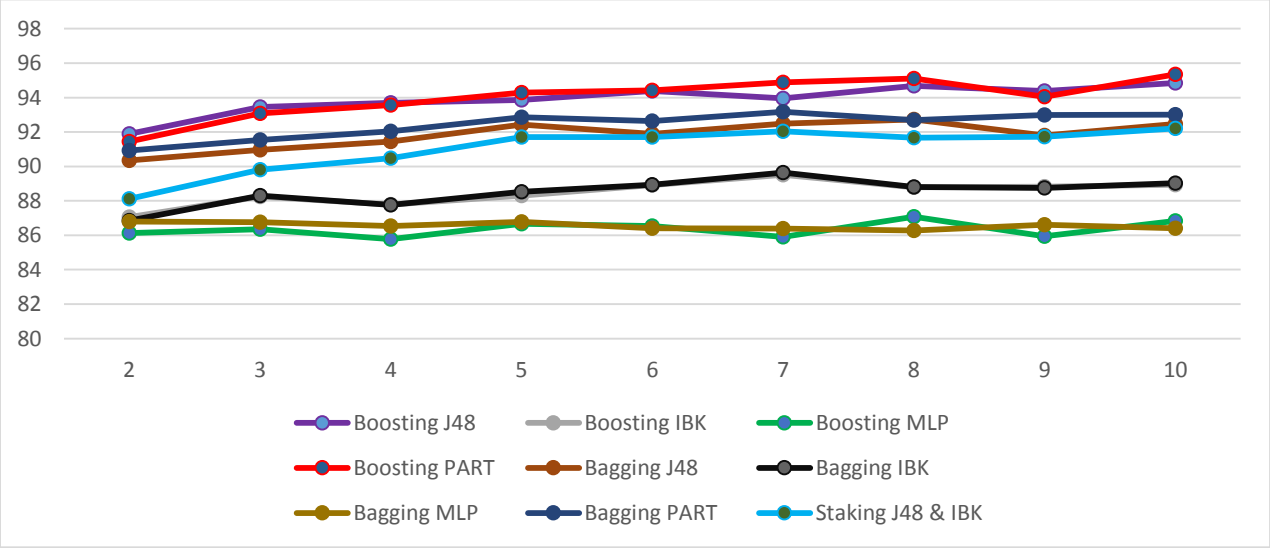


Figure 25: Ensemble learning progress on all attributes and cross validation for a month ahead prediction

Method	Algorithm	Accuracy of models with 75% split
Boosting	J48	94.23
	IBK	90.90
	MLP	86.03
	PART	94.49
Bagging	J48	92.06
	IBK	90.39
	MLP	87.06
	PART	93.59
Staking	J48 & IBK	91.02

Table 22: Ensemble method with all attributes using 75% split for one month ahead prediction

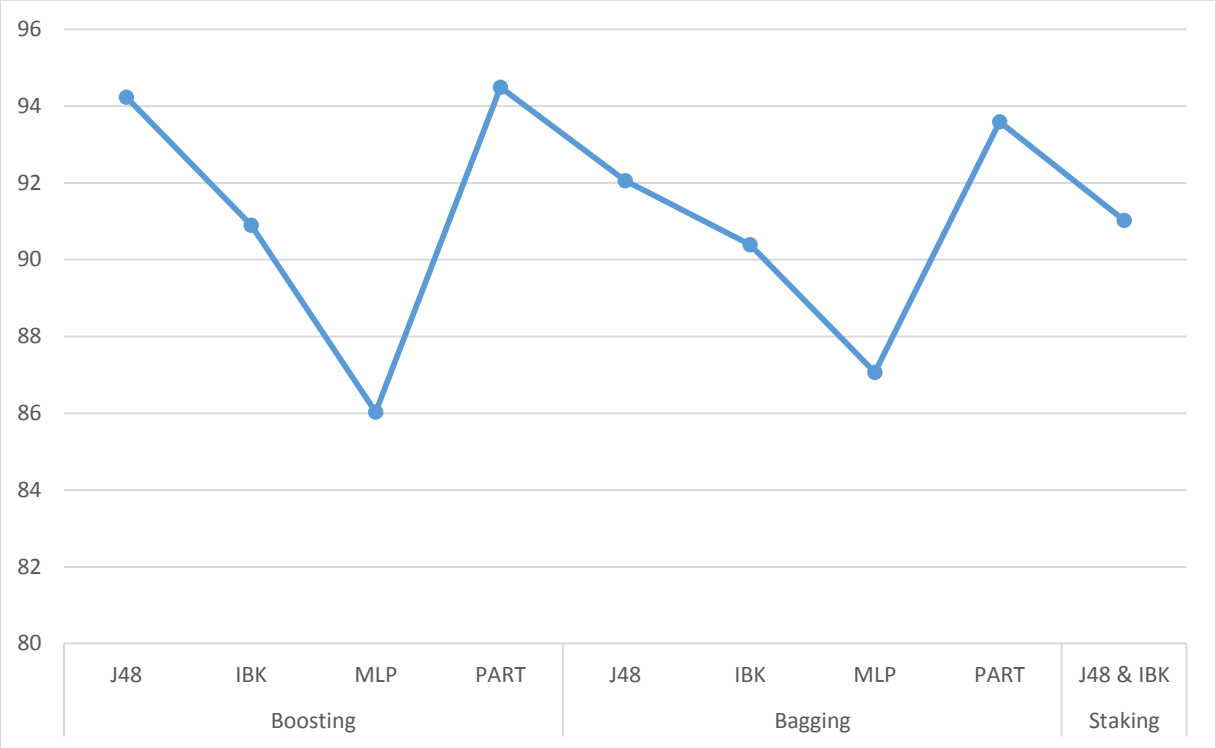


Figure 26: Ensemble learning progress on all attributes and 75% split for a month ahead prediction

4.4.6.3 Single Algorithms Performance Comparison Prediction after one Month ahead using Selected Attributes

Algorithm	Accuracy of models with different Number of folds								
	2	3	4	5	6	7	8	9	10
J48	88.97	91.38	91.38	91.34	91.70	92.86	92.63	92.21	92.82
IBK	90.54		92.63	93.01	93.56	94.24	93.72	93.23	93.94
MLP	86.86	86.79	86.54	86.86	86.41	86.61	86.86	86.14	86.83
PART	88.56	91.09	91.09	91.38	91.41	91.41	91.47	91.57	92.82

Table 23: Single algorithms with selected attributes using cross validation for one month ahead prediction

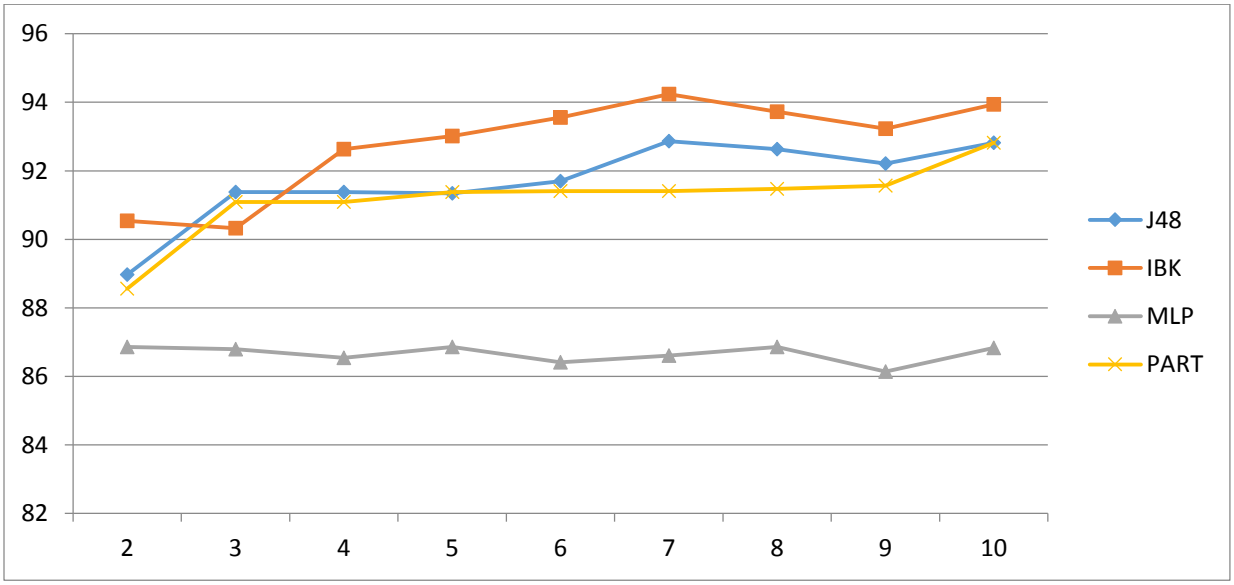


Figure 27: Single algorithms learning on selected attributes and cross validation for a month ahead prediction

Single algorism comparison using 75 percentage split

Algorithm	Accuracy of models with 75% split
J48	92.43
IBK	94.24
MLP	86.67
PART	90.12

Table 24: Single algorithms with selected attributes using 75% split for one month ahead prediction

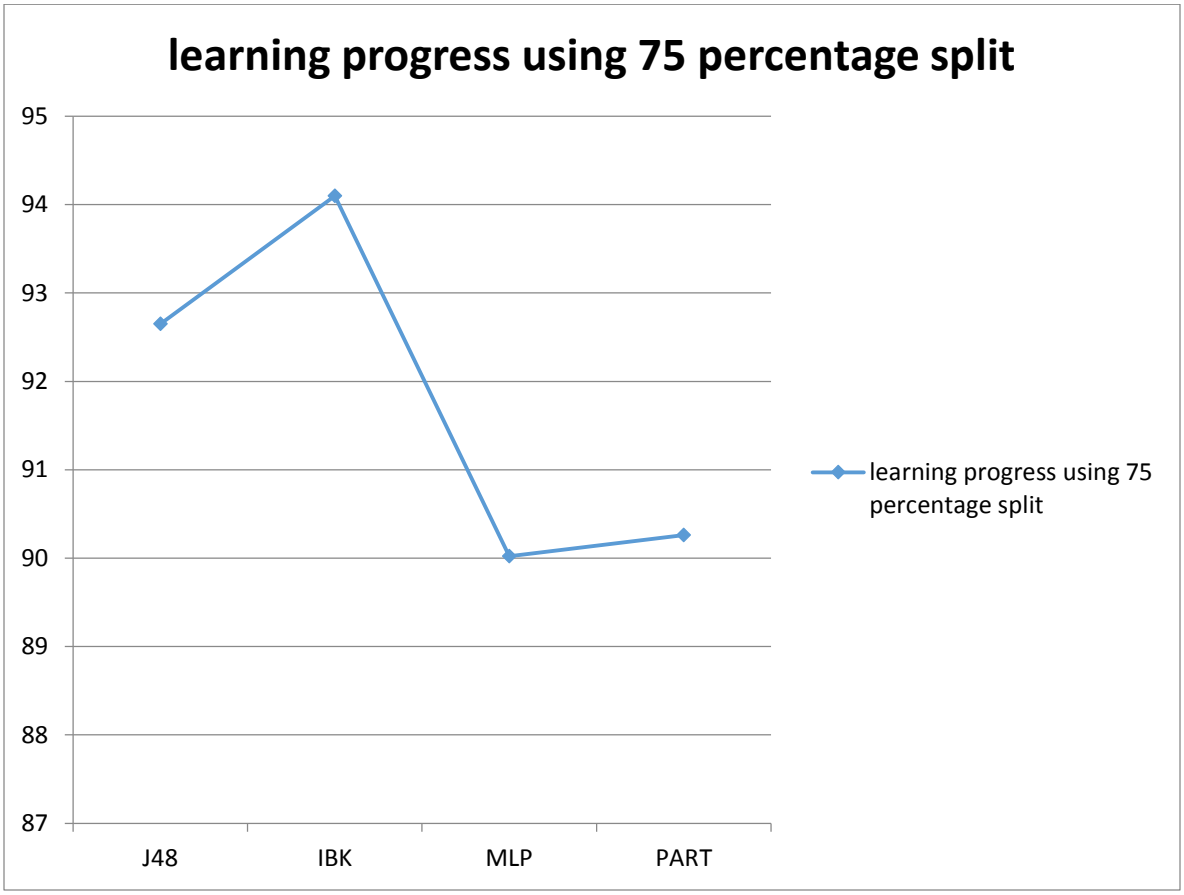


Figure 28: Single algorithms learning on selected attributes and 75% split for a month ahead prediction

4.4.6.4 Ensemble Algorithms Performance Comparison Prediction after one Month ahead using Selected Attributes

Method	Algorithm	Accuracy of models with different Number of folds								
		2	3	4	5	6	7	8	9	10
Boosting	J48	91.92	93.59	93.91	94.45	94.36	94.81	94.78	94.65	95.18
	IBK	90.54	91.41	92.63	93.01	93.56	94.24	93.72	93.23	93.94
	MLP	86.51	86.41	86.12	86.67	86.60	86.57	87.31	86.20	86.92
	PART	91.41	94.01	93.88	94.32	94.78	94.81	94.42	94.64	95.03
Bagging	J48	90.99	92.05	92.02	92.43	92.63	92.63	92.15	92.91	93.14
	IBK	89.10	90.35	91.67	91.22	91.83	91.83	91.99	91.94	92.56
	MLP	86.79	86.47	86.83	86.16	86.60	86.60	86.44	86.74	86.38
	PART	90.51	92.31	92.88	92.94	92.82	92.82	93.21	93.62	93.49
Staking	J48 & IBK	90.26	91.41	92.50	92.40	93.62	93.95	93.72	92.91	93.69

Table 25: Ensemble methods with selected attributes & cross validation for one month ahead prediction

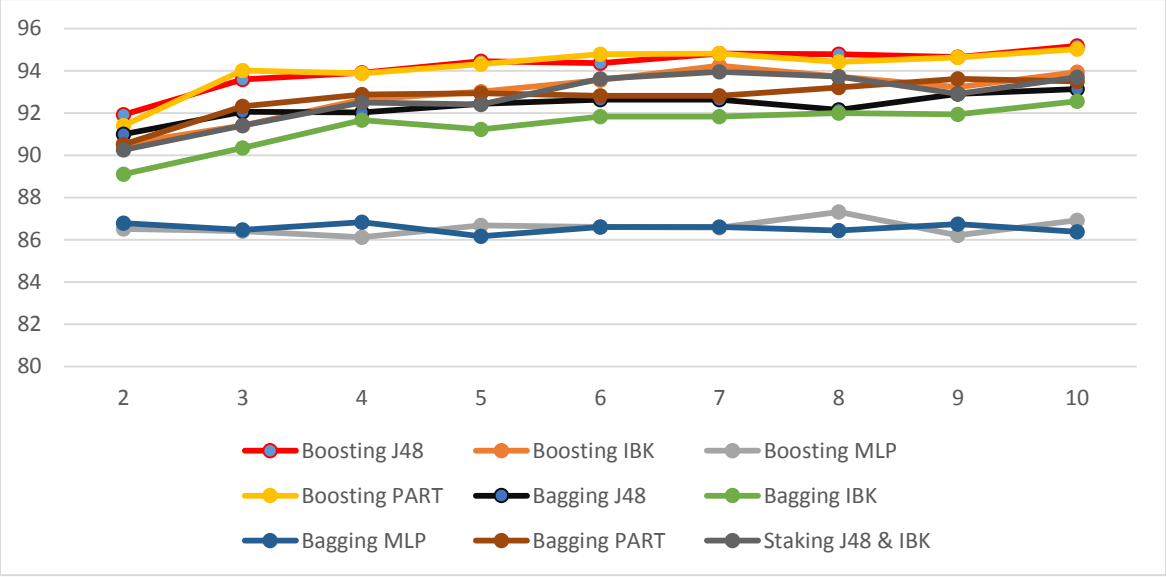


Figure 29: Ensemble learning progress on selected attributes and cross validation for a month ahead prediction

Method	Algorithm	Accuracy of models with 75% split
Boosting	J48	94.23
	IBK	94.24
	MLP	86.29
	PART	94.62
Bagging	J48	92.44
	IBK	93.34
	MLP	86.67
	PART	93.21
Staking	J48 & IBK	93.85

Table 26: Ensemble methods with selected attributes using 75% split for one month ahead prediction

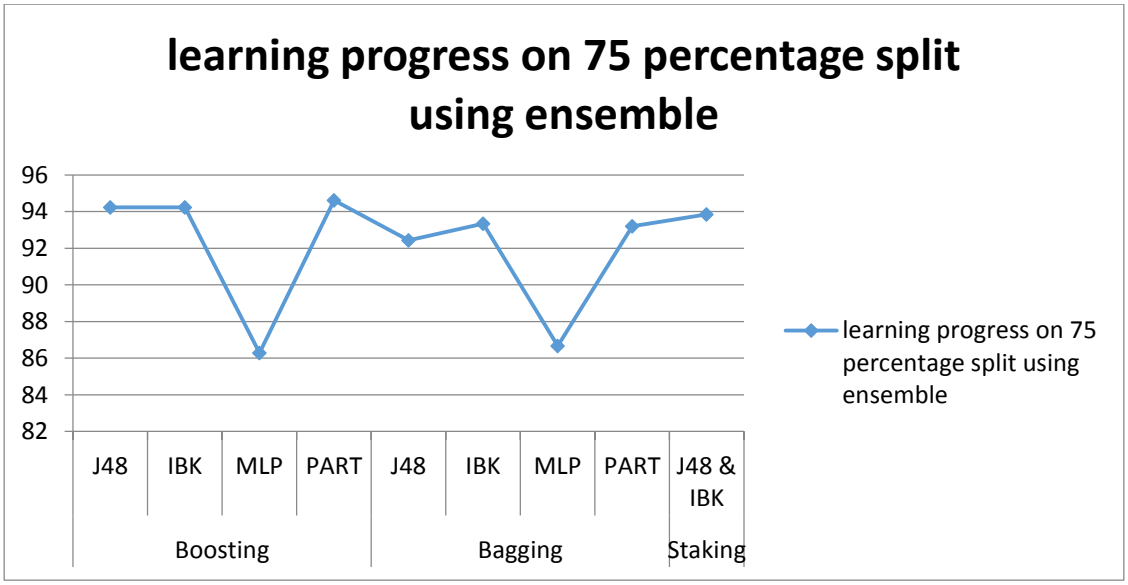


Figure 30: Ensemble learning progress on selected attributes and 75% split for a month ahead prediction

4.5 Selected Model Performance Evaluation

Evaluation of classification algorithms is one of the key points in any process of data mining. The most commonly tools used in analyzing the results of classification algorithms applied are confusion matrix, learning curves and receiver operating curves (ROC) Oprea, [3, 56]. In this study, confusion matrix, ROC curve, MAE and RMSE performance metrics are used.

In this study, the full training set containing a total of 9,543 instances were used with 8 selected and 10 all attributes. PART algorithm produced best performance rather than J48, IBK and MLP algorithms with selected attribute for after a day's advance prediction, and with all attributes for one month's advance prediction. in terms of, it's performance and number of leaves, relatively PART algorithm is more understandable and less complex to human than other models generated. Therefore, the performance of PART algorithm with selected and all attributes for a day's advance and a month's advance prediction respectively, given a valuable information in predicting rainfall as compared to other models.

4.5.1 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 on single Algorithm & Ensemble Method for one Month ahead Prediction using all Attributes

Method	Learning Algorithm	MAE	RMSE
Single Algorithm	PART	0.08	0.24
Boosting	PART	0.05	0.18

The above table showed that boosting classifiers has led to considerable decrease of prediction error from (MAE=0.08) to (MAE = 0.05) compared to single algorithm with selected attributes. Also, the prediction error in boosting decrease from (RMSE=0.24) to (RMSE=0.18) 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

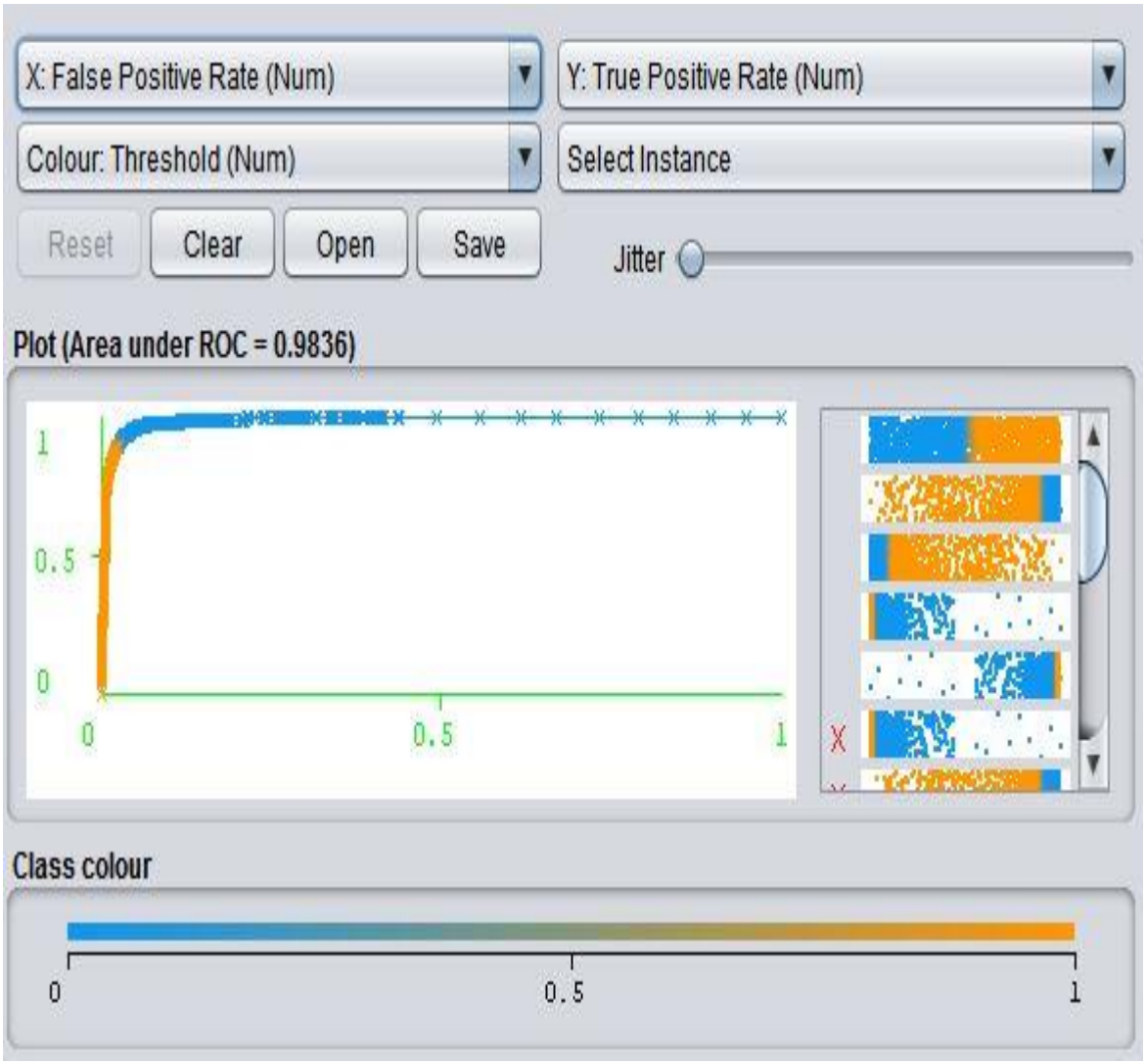
4.5.2 PART Algorithm Model with Selected Attributes for one Day ahead Prediction

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%.

Evaluation using ROC area and Confusion Matrix for one Month ahead Prediction using all Attributes

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

Instances: 312

Attributes: 10

Correctly Classified Instances 303 97.1154 %

Incorrectly Classified Instances 9 2.8846 %

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.854	0.011	0.921	0.854	0.886	0.870	0.963	0.931	no
	0.989	0.146	0.978	0.989	0.983	0.870	0.967	0.991	yes
Weighted Avg.	0.971	0.129	0.971	0.971	0.971	0.870	0.966	0.983	

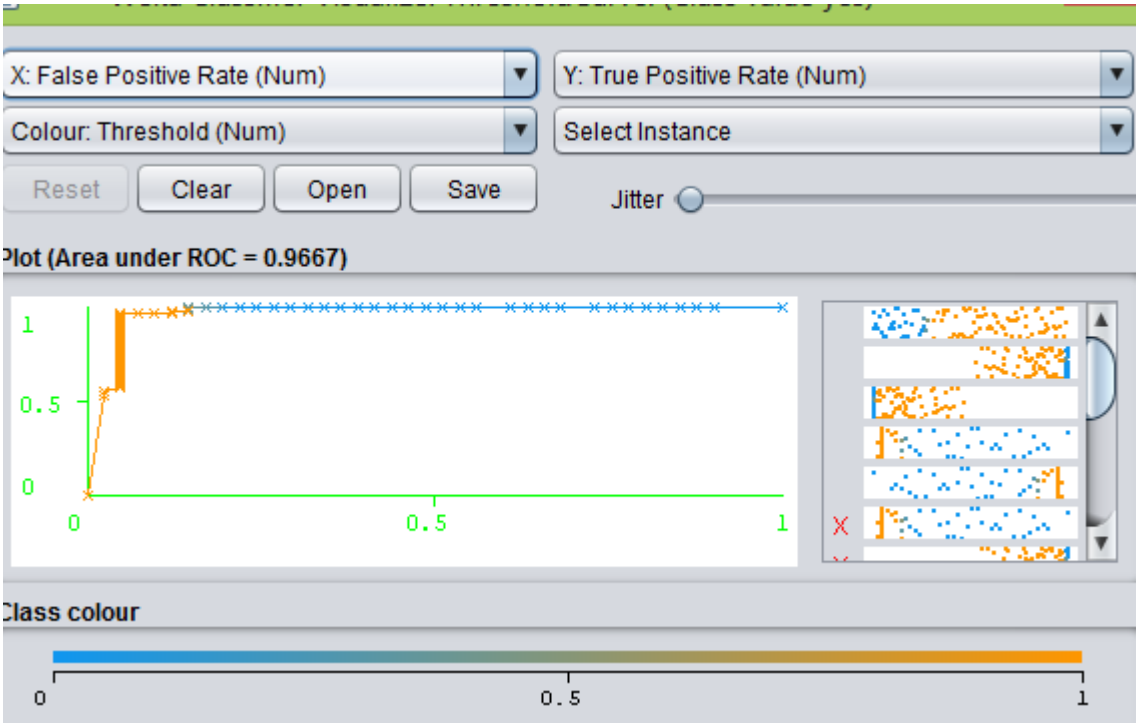
=== Confusion Matrix ===

a	b	<-- classified as
35	6	a = no
3	268	b = yes

4.5.3 PART Algorithm Model with all Attributes for one Month ahead Prediction

In PART algorithm predictive model when the model is built on WEKA 3.8 Explorer, the predictive performance of the model is 97.126% i.e. **303** instances were classified correctly while 9 (2.89%) instances were wrongly misclassified to other class. The model classified **268** instances as 'yes' out of **271** instances; in fact they are yes tested on the test data or which are classified correctly in the class of 'yes'. The remaining **3** instances were misclassified to another class as "no" actually they are "yes". The model classified 35 instances as " no" out of 41 instances that in fact " no" and wrongly classified 6 instances to other class as "yes" while actually they are "no".

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 all attributes.



In the above ROC shows that the model is better in detecting true positive rather than false positive. The area under the model PART algorithm model is 0.9667 which is closer to 1 showing that the class value 'yes' gives ROC accuracy of 96.7%.

4.5.4 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 95.46% and 95.66% accurate prediction, on WEKA experimenter and explorer respectively for one day's advance prediction.

As discussed in the first chapter Ethiopian metrological service agency uses a number of bulletins available information's from a variety of source abroad, such as US and UK, to prepare seasonal in Ethiopia. The methods of forecasting which are applied in preparing seasonal forecast in Ethiopian national metrology agency are based on traditional forecasting method like analogue, trend analysis and statistical assessment. As different scholars explored that support metrological agency by different analysis tools and models could enhance prediction accuracy.

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. Also, the prediction error of single PART and boosted PART algorithm there is decrement of prediction error from (RMSE=0.24) to (RMSE=0.18). Finally, boosted PART algorithm given excellent performance with selected and all attributes for a day's and month's advance prediction respectively.

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.

4.5.5 Rule Extraction by using PART Algorithm for one Day advance Prediction

To extract rules PART algorithm is more human-readable and easy with each path. If the condition is satisfied then the conclusion follows.

Rule 1: rhumidity > 92.0 AND windspeed <= 0.5 AND sunshine > 3.5: yes (65.53/0.0)
If relative humidity is greater than 92.0 and wind speed less than or equal to 0.5 and sunshine is greater than 3.5 then the likelihood of rainfall will be 100%.
The first rule selected from the rules generated by part algorithm model, gives 65.5 correct result out of 65.5 instances covered. From this, the likelihood of predictability of the individual to 'yes' by the above predictors is 100%.

Rule 2: rhumidity <= 34.0 AND windspeed > 1.4 AND rhumidity <= 31.0 AND rhumidity > 25.5: no(28)

If relative humidity is less than or equal to 34.0 and wind speed greater than 1.4 and relative humidity less than or equal to 31.0 and relative humidity is greater than 25.5 then the likelihood of no rainfall will be 100%.

The above rule states that the likelihood of rainfall is no 100% with the concomitant attribute in rule 2.

Rule 3: maxtemp <= 24.0 AND sunshine <= 7.0 AND mintemp > 9.8 AND wind speed <= 0.9 AND: yes (38.5/0.09).

Rule 3: If maximum temperature is less than or equal to 24.0 and sunshine less than or equal to 7.0 and minimum temperature greater than 9.8 4 then the likelihood of rainfall will be 99.8%.

Rule 3: states that maximum temperature and wind speed matters the rainfall 'yes' with incorporating sunshine and minimum temperature. As it has seen in the above rules, the probability of rain is increases as temperature, sunshine and wind speed attainment. Results from PART algorithm data mining algorithm reveals that the likely chance of rainfall is 99.8%.

Rule 4: windspeed <= 1.3 AND mintemp > 9.3 AND maxtemp <= 24.5 AND sunshine > 8.9 AND rhumidity <= 70.0: yes (83.85/0.4).

Rule 4: If wind speed less than or equal to 1.3 and minimum temperature greater than 9.3 and maximum temperature less than or equal to 24.5 and sunshine greater than 8.9 and relative humidity is less than or equal to 70.0 then the likelihood of rainfall will be 99.5%.

Rule 4: states that maximum temperature and minimum temperature matters the rainfall to be 'yes' with incorporating wind speed and relative humidity. As we have seen in the above rules, the probability of rain increases as the maximum temperature, minimum temperature, relative humidity and wind speed attainment. Results from PART algorithm data mining algorithm shows that the likely chance of rainfall is 99.5% .

Rule 5: sunshine > 9.5 AND rhumidity > 70.0 AND mintemp <= 12.0 AND windspeed <= 0.9: yes (56.23/0.34)

If sunshine is greater than 9.5 and relative humidity greater than 70.0 and minimum temperature less than or equal to 12.0 and wind speed is less than or equal to 0.9 then the likelihood of rainfall will be 99.4%.

The above rule selected from PART algorithm also shows that if the predictors in the above rule 5, are fulfilled the chance of the rainfall likely to be 'yes' is 99.4%. Relative humidity, wind speed, temperature and sunshine matter the condition of rainfall to be 'yes'. The domain experts accepted this rule by saying high relative humidity is known factors for rain is "yes". Therefore, PART algorithm of data mining technology revealed that, if relative humidity >70 and minimum temperature <=12 and sunshine > 9.5 and wind speed<=0.9 then the probability of experiencing rainfall is 99.4% with the above predictor attributes.

Rule 6: sunshine <= 3.7 AND mintemp <= 12.3 AND sunshine > 3.5: yes (89.82/0.52)
If sunshine is less than or equal to 3.7 and minimum temperature less than or equal to 12.3 and sunshine is greater than 3.5 then the likelihood of rainfall will be 99.4%.

Rule 7: sunshine <= 2.9 AND rhumidity > 41.0 AND mintemp <= 9.0: yes (74.44/0.95)
If sunshine is less than or equal to 2.9 and relative humidity greater than 41.0 and minimum temperature is less than or equal to 9.0 then the likelihood of rainfall will be 98.7%.
Rule 6 and 7 states that sunshine, minimum temperature and relative humidity matter the fate of the rainfall 'yes' without incorporating wind speed and maximum temperature. As, we have seen on the above rules, the probability of rainfall is increases as the humidity, minimum temperature and sunshine attainment. Results from PART algorithm reveals that the likely chance of rain to occur based on temperature, relative humidity and sunshine is (99.4% and 98.7%) respectively. The domain experts accepted this rule as temperature; relative humidity and sunshine are the central role for rainfall to be 'yes'.
Rule 8: mintemp > 9.6 AND maxtemp > 20.4 AND windspeed > 0.5: yes (54.67/1.04)
If minimum temperature is greater than 9.6 and maximum temperature greater than 20.4 and wind speed greater than 0.5 then the likelihood of rainfall will be 98.1%.
Rule 8, states that temperature and wind speed matter the state of the rainfall 'yes' without incorporating other variables. Results from PART algorithm data mining algorithm reveals that, the likely chance of rain to occur based on the above predictor attributes, are (98.1%) . The domain experts accepted this rule as temperature and wind speed are the central role for rainfall to be 'yes'.

4.5.6 Rule Extraction by using PART Algorithm for one Month advance Prediction

Rule 1: mintemp > 5.59 AND rhumidity > 36.1 AND sunshine <= 9.7 AND windspeed <= 0.84: yes (104.37)

If minimum temperature is greater than 5.59 and relative humidity greater than 36.1 and sunshine less than or equal to 9.7 and wind speed is less than or equal to 0.84 then the likelihood of rainfall will be 100%.

The above rule states that the likelihood of rainfall will be 100% with the concomitant attributes in rule 1.

Rule 2: maxtemp <= 30.9 AND windspeed <= 1.1 AND sunshine > 8.6: yes (26.31)

If maximum temperature less than or equal to 30.9 and wind speed less than or equal to 1.1 and sunshine is greater than 8.6 then the likelihood of rainfall will be 100%.

The above rule states that the likelihood of rainfall will be 100% with the concomitant attributes in rule 2.

Rule 3: maxtemp <= 32.21 AND rhumidity > 31.84 AND maxtemp > 22.8 AND: yes (42.86/0.27)

If maximum temperature is less than or equal to 32.21 and relative humidity greater than 31.84and maximum temperature is greater than 22.7 then the likelihood of rainfall will be 99.4%.

The first rule selected from the rules generated by PART algorithm model, gives 42.86 correct result out of 43.13 instances covered. From this, the likelihood of predictability of the individual to 'yes' by the above predictors is 99.4%.

Rule 4: mintemp > 5.59 AND rhumidity > 36.1 AND windspeed <= 0.84: yes (80.34/2.17)

If minimum temperature is greater than 5.59 and relative humidity greater than 36.1 and wind speed is less than or equal to 0.84 then the likelihood of rainfall will be 97.4%.

Rule 2: states that minimum temperature, humidity and wind speed matter the state of the rainfall 'yes' without incorporating other variables. Results from PART algorithm data mining algorithm reveals that, the likely chance of rain to occur based on the above predictor attributes, are (97.4%) . The domain experts accepted this rule as minimum temperature, humidity and wind speed are the central role for rainfall to be 'yes'.

Rule 5: windspeed > 1.2 AND rhumidity > 31.84: yes (50.29/2.13)

If wind speed is greater than 1.2 and humidity is greater than 31.84 then the likelihood of rainfall will be 95.49%.

Rule 5, states that wind speed and humidity matter the state of the rainfall 'yes' without incorporating other variables. Results from PART algorithm data mining algorithm reveals that, the likely chance of rain to occur based on the above predictor attributes, are (95.49) . The domain experts accepted this rule as humidity and wind speed are the central role for rainfall to be 'yes'.

Rule 6: mintemp <= 10.1 AND sunshine <= 9.3 AND AND rhumidity <= 68.8 AND: no (36.24/2.09)

If minimum temperature is less than or equal to 10.1 and sunshine less than or equal to 9.3 and humidity is less than or equal to 68.8 then the likelihood of there is no rainfall 94.6%.

The first rule selected from the rules generated by part algorithm model, gives 36.24 correct result out of 38.33 instances covered. From this, the likelihood of predictability of the individual to 'yes' by the above predictors is 94.6%.

4.6 Using Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains Changala et al. [31]. Finally, the discovered knowledge is deployed. Also this step is to put discovered knowledge in practical use either by documenting it and reporting it or by embedding it in a computer system. In first sight, this stage might be regarded as trivial and straightforward, but this is not the case. The conclusions drawn from the KDD process often reveal the complex nature of the problem and its solutions. Hence, the implementation of the new knowledge should often be done gradually, while continuously monitoring the result achieved and the degree to which they fulfill the expectations.

4.6.1 Prototype development for Rainfall forecasting

The Prototype is easily to routine and necessary for the end user to use the application of the proposed method easily. The interface in this study developed based on the major determinant attributes, that are used for rainfall prediction as the rules are selected based on the prediction expert suggestion and accuracy. In this thesis, as described in chapter three hybrid methodology is used. One of the advantage of hybrid methodology is the last step 'using the discovered knowledge' which allows to develop the prototype for the potentially useful knowledge.

As a result of the above useful rules the prototype is developed by using VB.NET 2010, the proposed interface for rainfall prediction is developed as shown below the figure.

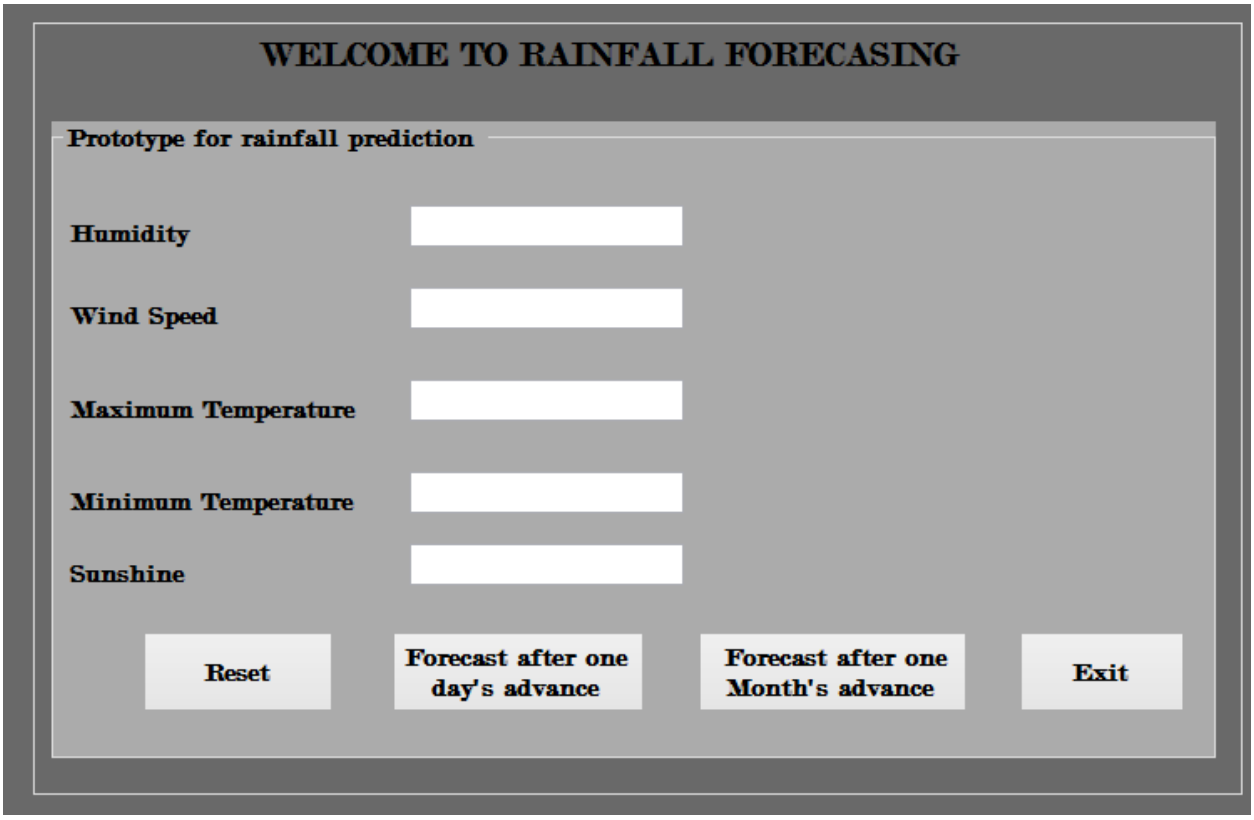


Figure 31: Prototype for rainfall prediction

CHAPTER FIVE

5. Conclusion and Recommendation

5.1 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 data mining 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, data mining, evaluation and use of the discovered knowledge.

In this study, we compare the effectiveness and performances of several data mining 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.

On the other hand, the experiments after one month's ahead prediction showed, that single algorithm IBK produced better results rather than J48, MLP and PART algorithm. IBK produces 93.94 % prediction accuracy, followed by J48 and PART algorithm with 92.82%, prediction accuracy using selected attributes. Also, Boosted PART algorithm produced better result on 10 fold cross validation and 75 percentage split, using all attributes rather than J48, IBK and MLP i.e. 95.35% and 94.49 respectively. In boosting ensemble method PART algorithm and J48 decision tree produced better prediction accuracy which is 95.03% and 95.18%, prediction accuracy respectively, on WEKA experimenter. However, PART algorithm produced 95.35% and 97.12% prediction accuracy; on WEKA experimenter and explorer respectively using all attributes for one month'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. Generally, PART algorithm with 10 fold cross validation and all attributes produced best result for one month's advance prediction. 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.

5.2 Recommendation

According to the presented conclusions, this report intends to provide some guidelines for improving current practice and initiate further research. This will enhance applicability of data mining technology in rainfall forecasting and activities in in line with advocacy efforts of early warning system to reduce drought.

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 data mining 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.

- In the present study in order to develop the rainfall forecasting model only historical data have been used and the model is developed after one day's ahead and one month's ahead. It's possible to extend the development of the model to longer forecast period such as ten days ahead and one year's ahead etc.
- Most likely association rules require additional investigation and contemporaneously more work in order to produce a reasonable number of logical and interpretable rules. The further study of the implementation of association rule and predict another atmospheric variable as well as using another variable like humidity and wind seed is a possible development of this research.
- 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. Y. Tafesse, "Artificial neural network for weather forecasting", Msc. Thesis, Dept. Information Science, Addis Ababa University., Addis Ababa, 2004.
5. 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.
6. The guardian, "Global Development about Food Security", December, 2015,http://www.theguardian.com
7. 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.
8. T. Ephrem, "Application of Data Mining for Weather Forecasting", Msc. Thesis, Dept. Information Science, Addis Ababa University, Addis Ababa, 2015.
9. Dhanya& Kumar, "Data mining for evolution of association rules for droughts and floods in India using climate inputs", Journalof geophysical research, Vol. 114,pp.1-15, January 2009.
10. V. Vaithiyanathan, "comparison of different classification techniques using different datasets",International Journal of Advances in Engineering & Technology, vol. 6, pp.764-768, May 2013.
11. V. Mistry and V.Patel, "Weather Condition Prediction Using Semi-Supervised Data Mining Technique",International Journal of Engineering Trends and Technology (IJETT), vol. 20, no. 4, pp. 179-183, Feb. 2015.
12. P. Atole et.al," Statistical Modeling for Rainfall Prediction using Data Mining Technique", An International Journal of Engineering & Technology (AIJET), vol.2,no. 1,pp.35-39, January, 2015.
13. W. Zaw and T. Naing, "Empirical Statistical Modeling of Rainfall Prediction over Myanmar", International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol.2, no. 10, 2008.
14. N.Sethi and K.Garg, "A Review of Application of Data Mining Technique for Rainfall Prediction",International Journal for Scientific Research & Development, vol. 2, pp.1095-1098, 2014.
15. A. R. Chaudhari et.al, "Data Mining with Meteorological Data", International Journal of Advanced Computer Research, vol. 3, no. 3, pp. 25-29, September 2013.

16. A. Williamson et.al, "The Socio-Economic Value of Improved Weather and Climate Information", George Washington University, Washington, DC, 20052, 2002.
17. T.Frei, "Economic & social benefits of meteorology and climatology", Switzerland meteorological application, 2009.
18. M. Wolfson and David A. Clark, "Advanced Aviation Weather Forecasts",Lincoln Laboratory Journal,Vol. 16, no. 1, 2006.
19. A.Firat, "Technological Forecasting A Review", Massachusetts Institute of Technology, Cambridge, MA 02142, 2008.
20. D.Wolters, "An exploratory study of the possibilities of analog post processing", Wageningen University, Netherland, intern rep. 2,2009.
21. P. Block and L. Goddard," Statistical and dynamical climate predictions to guide water resources in Ethiopia,International Research Institute for Climate and Society,Columbia University.
22. I.Maqsood et.al. (2004 may 20). "An ensemble of neural networks for weather forecasting". Springer-Verlag London. [online]: 10.1007/s00521-004-0413-4.
23. Smitha and V. Sundaram, "Knowledge Discovery from Real Time Database using Data Mining Technique", International Journal of Scientific and Research Publications, vol. 2, pp. 1-3, April 2012.
24. Kaur&Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications,Vol. 98,No.22, pp.13-17,July 2014.
25. S.Vijayarani and M.Divya, "An Efficient Algorithm for Classification Rule Hiding", International Journal of Computer Applications vol. 33, no.3, pp. 39-45, November 2011.
26. L. Rokach. (2009, Nov. 19). "Ensemble-based classifiers". Springer Science and Business Media. [Online]. Avilable: 10.1007/s10462-009-9124-7.
27. A.Tiwari and A.Prakash, "Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA", International Journal of Engineering and Technical Research (IJETR), vol. 2, pp. 207-209, Sept. 2014.
28. Galar et.al. "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transactions on systems, man, and cybernetics—part C: Applications and reviews, pp. 1-22, 2011.
29. U. Shafique and H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)" International Journal of Innovation and Scientific Research, vol. 12, no. 1, pp. 217-222, Nov. 2014.
30. O. Marban et.al, "Data Mining and Knowledge Discovery in Real Life Applications", "A Data Mining & Knowledge Discovery Process Model", J. Ponce and A. Karahoca, Ed. Spain,Madrid:2009.[Online].Availablefrom:http://www.intechopen.com
31. Changala et.al, "Knowledge Discovery Process: The Next Step for Knowledge Search",International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, pp. 4277-4283, May 2015.

32. C.Latha et.al, "A Service Oriented Architecture for Weather Forecasting Using Data Mining", Int. J. of Advanced Networking and Applications, vol. 2, pp. 608-613, May 2010.
33. 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.
34. 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.
35. R. Kanth et.al, "analysis of indian weather data sets using data mining techniques"Computer Science & Information Technology (CS &IT),pp. 89–94, 2014.
36. 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.
37. P.Hemalatha, "Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System", International Journal of Computational Engineering Research (ijceronline.com), vol. 3, pp. 198-202, March 2013.
38. R. Kumar, "Decision Tree for the Weather Forecasting", International Journal of Computer Applications, vol. 76, no. 2, pp. 31-34, August 2013.
39. Mandale& B.A Jadhawar, "Weather forecast prediction: a Data Mining application", International Journal of Engineering Research and General Science, vol. 3,pp. 1279-1284, March-April, 2015.
40. S. Meganathan et.al, "Rain prediction using environmental data", RJC, vol. 8, no. 1, pp. 56-58, March 2015.
41. H. El-Shafie et.al, "Artificial neural network technique for rainfall forecasting applied to Alexandria, Egypt" International Journal of the Physical Sciences , Vol. 6, pp. 1306-1316, March 2011.
42. P. Gelder et.al, "Statistical Modeling for Rainfall Prediction using Data Mining Technique",An International Journal of Engineering & Technology, vol. 2, no. 1, pp.35-39, Jan. 2015.
43. Khan et.al, "Hourly Based Climate Prediction Using Data Mining Techniquesby Comprising Entity Demean Algorithm", Middle-East Journal of Scientific Research,vo. 23,no, 1, pp.35-40,2015.
44. P. Gaikwad and V. B. Nikam, "Different Rainfall Prediction Models and General Data Mining Rainfall Prediction Model", International Journal of Engineering Research & Technology (IJERT), vol. 2, pp. 115-123, July 2013.
45. Mahajan&Mazumdar," Flood A Fast Learning Algorithm for Rainfall Prediction", International Journal of Computer Applications, vol. 85,no. 15, January 2014.
46. Lusia and Suhartono, "Ensemble Method Based for Rainfall Forecasting in Indonesia "International Journal of Science and Research (IJSR), vol. 2, February 2013.
47. Bushara and Abraham, "Novel Ensemble Method for Long Term Rainfall Prediction "International Journal of Computer Information Systems and Industrial Management Applications, vol. 7,pp. 116-130, 2015.

48. Coelho, “Forecast Calibration and Combination: Bayesian Assimilation of Seasonal Climate Predictions”, PhD. Dissertation, dept.Meteorology, Reading university, South America, 2005.
49. N. Bhargava, “Selection Criteria for Data Mining Software: A Study”, International Journal of Computer Science Issues, vol. 10, no. 2, pp. 308-312, May 2013.
50. N.Rajasekhar and R.Kanth, “Weather analysis of guntur district of andhra region using hybrid SVM data mining techniques”,International Journal of Engineering and Advanced Technology (IJEAT), vol. 3, pp. 133-136, April 2014.
51. 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.
52. C. Onwubolu et.al, “Self-organizing data mining for weather forecasting”, IADIS, European Conference on Data Ming, 2007.
53. J.Kaiser, “Dealing with Missing Values in Data”, Journal of Systems Integration, vol. 1, pp. 42-51, 2014.
54. 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.
55. 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.
56. C. Opera, “Performance Evaluation of the Data mining Classification Methods” Annals of the Constantin Brancuşi, University of Targujiu, Economy Series, pp.249-253,2014.

Annex 1: Outlier detection

Extreme Values

			Case Number	Value
mintemp	Highest	1	9105	24.5000
		2	9094	23.4000
		3	2117	19.0000
		4	8206	19.0000
		5	8920	17.5000
	Lowest	1	1488	.4000
		2	2953	.7000
		3	396	.7000
		4	3302	.8000
		5	8783	1.0000
rhumidity	Highest	1	8	1.0000
		2	2970	1.0000
		3	2982	1.0000
		4	3082	1.0000
		5	3140	1.0000
	Lowest	1	5524	.0000
		2	5523	.0000
		3	5522	.0000
		4	5521	.0000
		5	5520	.0000

Extreme Values

			Case Number	Value
sunshine	Highest	1	157	11.8000
		2	3267	11.8000
		3	4212	11.8000
		4	4579	11.8000
		5	8363	1.1800
	Lowest	1	9064	.0000
		2	9024	.0000
		3	8518	.0000
		4	8360	.0000
		5	8312	.0000
windspeed	Highest	1	7372	11.3000
		2	7584	11.3000
		3	8131	11.3000
		4	4816	11.2000
		5	4820	1.1200
	Lowest	1	8335	.0000
		2	8319	.0000
		3	8304	.0000
		4	8300	.0000
		5	8298	.0000

Annex 2: Experimental Set up

Experiential setup on WEKA Experimenter with J48 algorithm

Parameters	Descriptions	Parameter type
Binary Split	When to use binary split on nominal attributes when building the trees.	Boolean
Confidence Factor	The confidence factor used for pruning (smaller values incur more pruning).	Numeric
Debug	If set to true, classifier may out put additional Info. to the console.	Boolean
minNumObj	The minimum number of instances per leaf.	Numeric
NumFolds	Determines the amount of data used for Reduced-error pruning. One fold is used for Pruning, the rest for growing the tree.	Numeric
reducedEroorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning.	Boolean
saveInstanceData	Whether to save the training data for visualization.	Boolean
Seed	The seed used for randomizing the data when reduced-error pruning is used.	Numeric
subtreeRaising	Whether to consider the sub tree raising operation when pruning.	Boolean
Unpruned	Whether pruning is performed.	Boolean
usedLaplace	Whether counts at leaves are smoothed based on Laplace.	Boolean

Experimental setup for IBK algorithm

Parameters	Descriptions	Parameter type
KNN	K-nearest neighbor on the data set.	numeric
Cross validate	Appropriate value of K based on cross validation.	Boolean
debug	If set to true, classifier may out put additional Info. to the console.	Boolean
Distance weighting	Gets the distance weighting method used.	Boolean
Mean squared	Weather the mean square error is used rather than mean absolute error during cross validation for regression problems.	Boolean
numdecimalplaces	The number of decimal places to be used for the output of numbers in the model.	numeric
Window size	Gets the maximum number of instances allowed in training model.	numeric

Experimental setup for PART algorithm

Parameters	Descriptions	Parameter type
Binary Split	When to use binary split on nominal attributes when building the trees.	Boolean
Confidence Factor	The confidence factor used for pruning (smaller values incur more pruning).	Numeric
Debug	If set to true, classifier may out put additional Info. to the console.	Boolean
minNumObj	The minimum number of instances per leaf.	Numeric
NumFolds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	Numeric
reducedEroorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning.	Boolean
saveInstanceData	Whether to save the training data for visualization.	Boolean
Seed	The seed used for randomizing the data when reduced-error pruning is used.	Numeric
subtreeRaising	Whether to consider the sub tree raising operation when pruning.	Boolean
Unpruned	Whether pruning is performed.	Boolean
usedLaplace	Whether counts at leaves are smoothed based on Laplace.	Boolean

Experimental setup on Bagging

Parameters	Descriptions	Parameter type
Bag size percent	Size of each bag, as a percentage of the training set size.	numeric
Bach size	The preferred number to pre process, if bach prediction is performed.	numeric
calcoutofbag	Weather out of bag error is calculated.	Boolean
Do not check capabilities	If set classifier capabilities are not checked before classifier is built.	Boolean
Num decimal places	The number of decimal places used for the output number in the model	Numeric
Num execution slot	The number of execution slot to construct ensemble.	Numeric
Num iteration	The number of iterations to be performed.	Numeric
Output out of bag complexity statistics	Weather to output complexity based statistics when out of bag calculation is performed.	Boolean
Represent copier using weights	Weather to represent copies of instances using weights rather than explicitly.	Boolean
seed	Random number of seed to be used.	numeric
Store out of bag predictions	Weather to store out of bag prediction.	Boolean

Experimental setup on Staking

Parameters	Descriptions	Parameter type
Bach size	Size of each bag, as a percentage of the training set size.	numeric
classifier	Base classifiers to be used	numeric
debug	If set to true, classifier may out put additional Info. to the console.	Boolean
Do not check capabilities	If set classifier capabilities are not checked before classifier is built.	Boolean
Num decimal places	The number of decimal places used for the output number in the model.	Numeric
Num execution slots	The number of execution slot to construct ensemble.	numeric
Num folds	Number of folds used for cross validation.	numeric
seed	Random number of seed to be used.	numeric

Annex 3: Sample Experimentation with Selected Attributes

Experimental result using 10- fold cross-validation on boosting Ensemble for one day ahead prediction

Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 2 -stddev-width 2 -sig-width 1 -count-width 5 -print-col-names -print-row-names -enum-col-names"

Analysing: Percent_correct

Datasets: 1

Resultsets: 4

Confidence: 0.05 (two tailed)

Sorted by: -

Date: 7/27/16 6:42 AM

Dataset	(1) meta.Ada	(2) meta.	(3) meta.	(4) meta.
god	(100) 95.44	94.83 *	89.90 *	95.46
	(v/ /*)	(0/0/1)	(0/0/1)	(0/1/0)

Key:

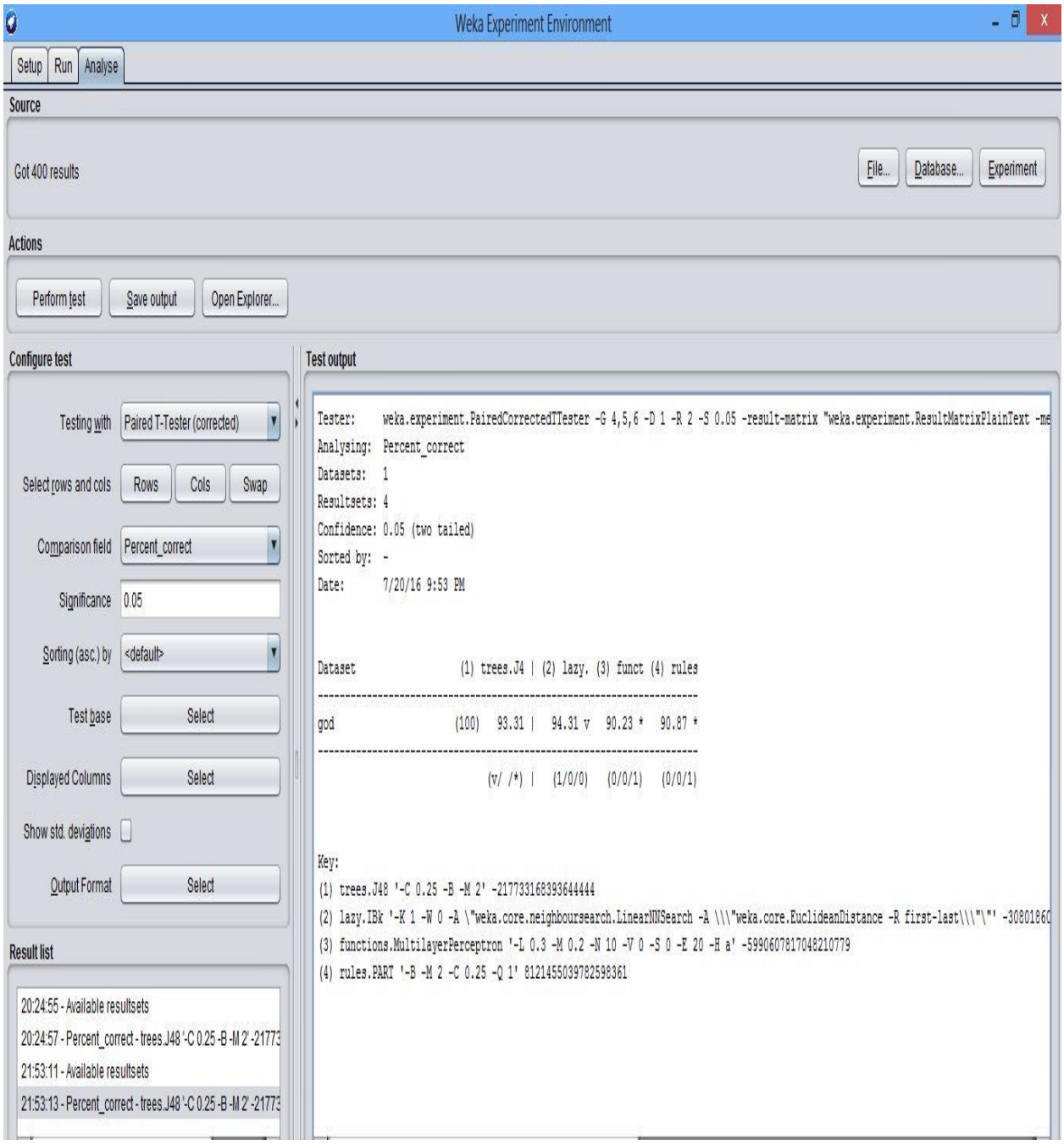
(1) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W trees.J48 -- -C 0.25 -B -M 2' -1178107808933117974

(2) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W lazy.IBk -- -K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -1178107808933117974

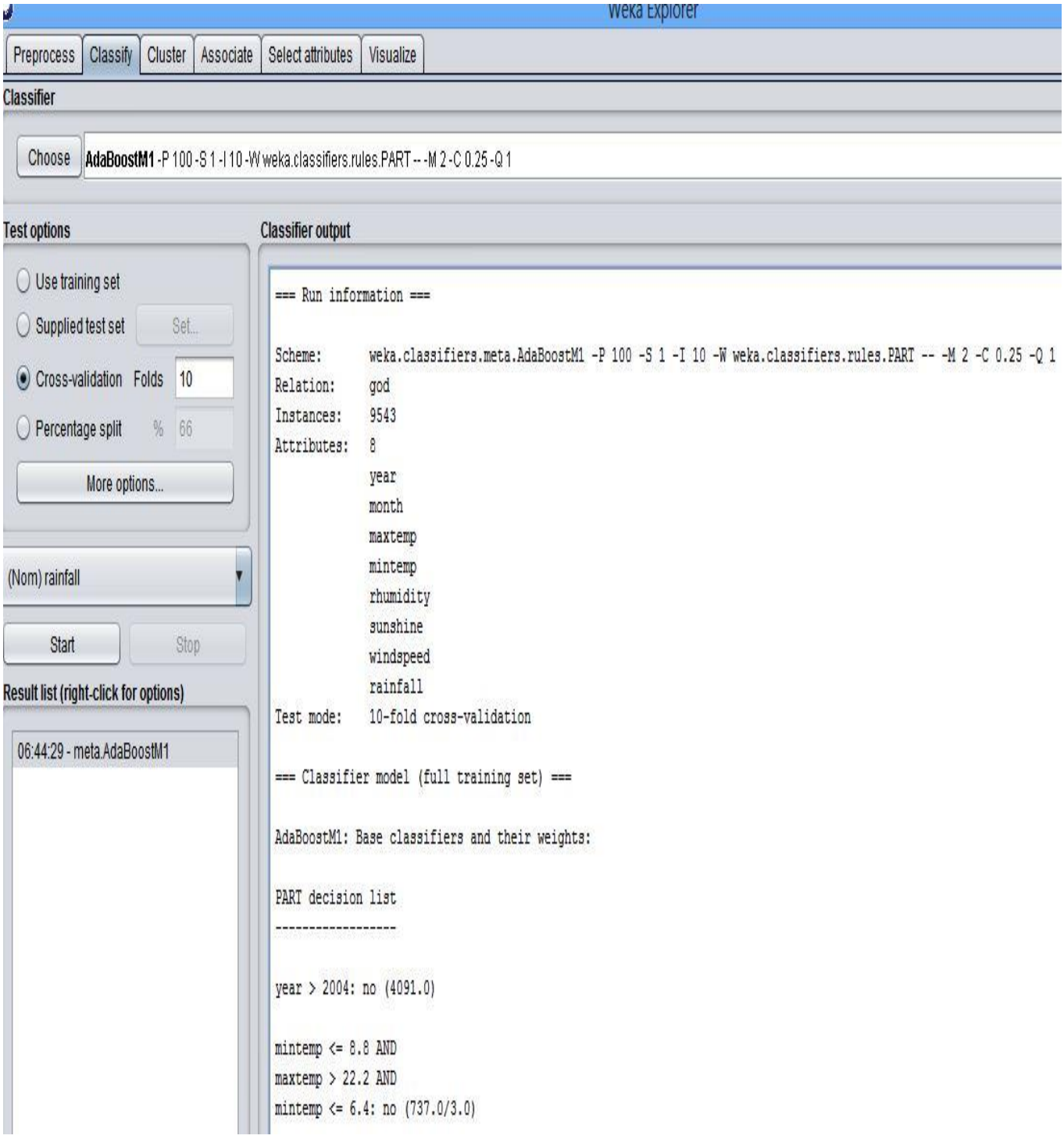
(3) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W functions.MultilayerPerceptron -- -L 0.3 -M 0.2 -N 10 -V 0 -S 0 -E 20 -H a' -1178107808933117974

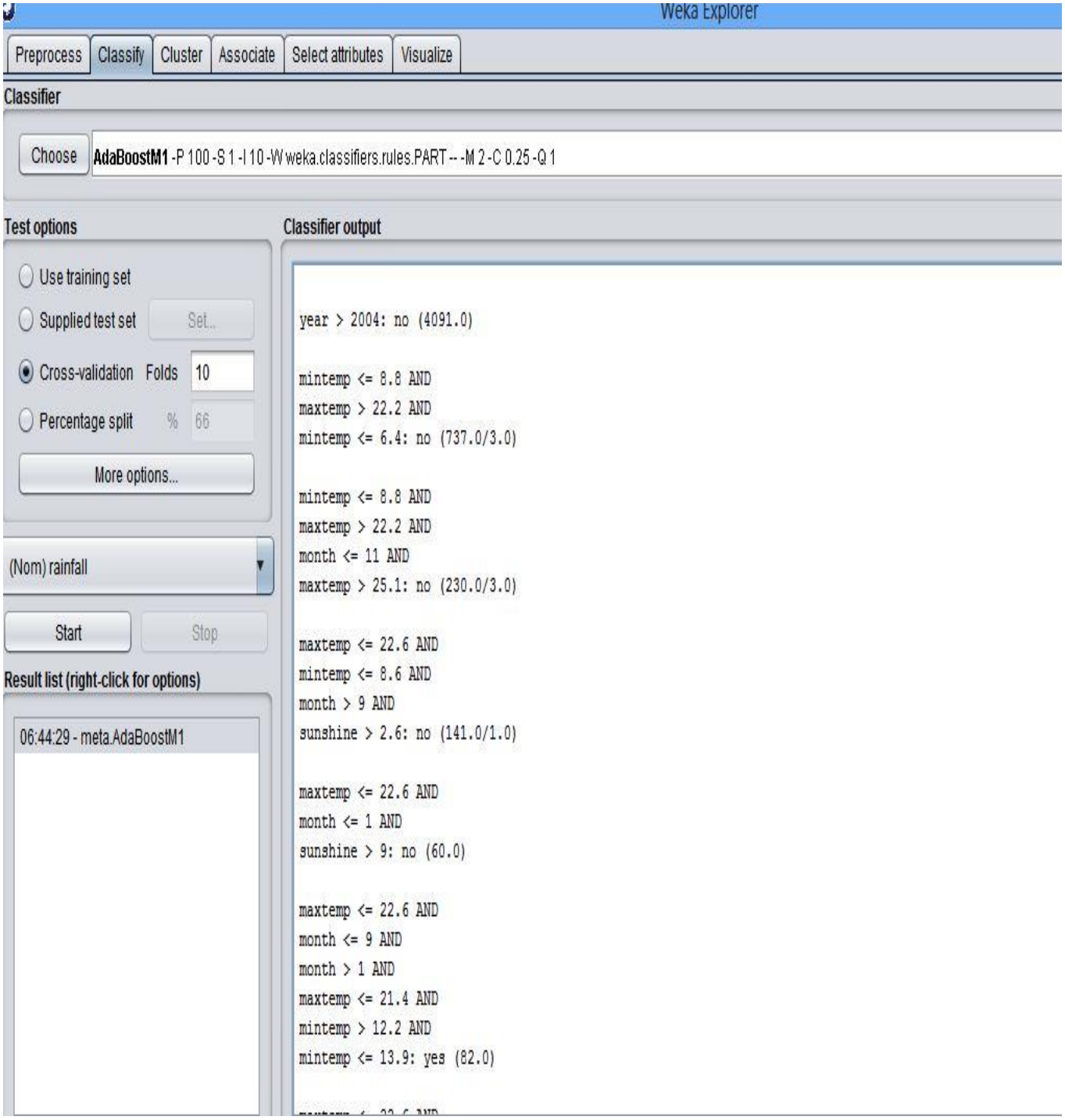
(4) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W rules.PART -- -B -M 2 -C 0.25 -Q 1' -1178107808933117974

Experimental result using 10- fold cross-validation on single algorithms



Annex 4: Partial view of PART algorithm rule for one day ahead prediction




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART -- -M 2 -C 0.25 -Q 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) rainfall
Start
Stop
Result list (right-click for options)
06:44:29 - meta.AdaBoostM1
Classifier output
year > 2004: no (4091.0)
mintemp <= 8.8 AND
maxtemp > 22.2 AND
mintemp <= 6.4: no (737.0/3.0)
mintemp <= 8.8 AND
maxtemp > 22.2 AND
month <= 11 AND
maxtemp > 25.1: no (230.0/3.0)
maxtemp <= 22.6 AND
mintemp <= 8.6 AND
month > 9 AND
sunshine > 2.6: no (141.0/1.0)
maxtemp <= 22.6 AND
month <= 1 AND
sunshine > 9: no (60.0)
maxtemp <= 22.6 AND
month <= 9 AND
month > 1 AND
maxtemp <= 21.4 AND
mintemp > 12.2 AND
mintemp <= 13.9: yes (82.0)

Annex 5: VB.NET Sample Code for Rainfall Forecasting

```
PublicClass Form1

PrivateSub Button2_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles Button2.Click
If humidity.Text = " "Then
            MsgBox("Please Enter humidity")
ElseIf windspeed.Text = " "Then
            MsgBox("Pleae Enter Wind Speed")
ElseIf mintemp.Text = " "Then
            MsgBox("Please Enter Minimum Temperature")
ElseIf maxtemp.Text = " "Then
            MsgBox("Please Enter Maximum Temperature")
ElseIf sunshine.Text = " "Then
            MsgBox("Please Enter Sunshine")
EndIf
If humidity.Text > 92 And windspeed.Text <= 0.5 And sunshine.Text > 3.5 And
mintemp.Text > 9.3 And maxtemp.Text <= 24.5 Then
            MsgBox(" TMORROW WILL BE  RAINY")
Else
            MsgBox(" TOMORROW WILL BE MOSTLY SUNNY")
EndIf
EndSub

PrivateSub Button1_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles Button1.Click
        humidity.Clear()
        windspeed.Clear()
        sunshine.Clear()
        mintemp.Clear()
        maxtemp.Clear()
EndSub
EndClass
```