Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 2 (2014) pp. 309-317

Weather Forecasting in Sudan Using Machine Learning Schemes

Nazim Osman Bushara[1] and Ajith Abraham[2]

[1] Faculty of computer science and information technology, Sudan University for Science and Technology (SUST),
P.O. Box 12094, SUDAPOST, Khartoum, Post Code 111 11, Sudan
nazim_ob@yahoo.com

[2] Machine Intelligence Research Labs (MIR Labs), Scientic Network for Innovation and Research Excellence,
P.O. Box 2259, Auburn, Washington 98071-2259, USA
ajith.abraham@ieee.org

***Abstract*: In this research, effort has been made to examine the relationship of rainfall in Sudan with important parameters such as Station, Wind Direction, Date, Humidity, Min-Temperature, Max-Temperature and Wind Speed. Attention has been made to find out correlation of rainfall with these elements. The goals of this paper are to demonstrate: (1) How feature selection can be used to identify the relationships between rainfall occurrences and other weather conditions and (2) Which classifiers can give the most accurate rainfall estimates? Monthly meteorological data by Central Bureau of Statistics Sudan from 2000 to 2012 for 24 meteorological stations has been used. To perform feature selection and building prediction models, we used group of data mining algorithms. The analysis shows that, Date, Min-T, Humidity and Wind D affect rainfall in Sudan, and we got the best 14 algorithms for building models to predict the rainfall.**

***Keywords*:** Weather forecasting, Rainfall prediction, Data Mining.

I. Introduction

Since ancient times, weather forecasting has been one of the most interesting and challenging area [1]. One of the important fields of weather forecasting is rainfall prediction which is important for food production plan and water resource management. Sudan is an agricultural country and most of its economy depends upon the agriculture. Rainfall plays an important role in agriculture Thus rainfall prediction becomes a significant factor in agricultural countries like Sudan. A wide range of rainfall forecast methods are employed in weather forecasting. Fundamentally, there are two approaches to predict rainfall. They are Empirical method and dynamical methods. The empirical approach is based on analysis of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. The most widely used empirical approaches for climate prediction are regression, artificial neural network, fuzzy logic and group method of data handling.

In dynamical approach, predictions are generated by physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions. The Dynamical approaches are implemented using numerical rainfall forecasting method [2].

The scientists have been tried to forecast the meteorological characteristics using a large set of methods, some of them more accurate than others. Lately, there has been discovered that data mining, a method developed recently, can be successfully applied in this domain. Data mining is about solving problems by analyzing data already present in databases. Data mining is defined as "the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities [3].

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 tasks employed. There are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data and predictive data mining tasks that attempt to do predictions based on inference on available data. This techniques are often more powerful, flexible, and efficient for exploratory analysis than the statistical

Data mining is aggregation of many disciplines which include machine learning, statistics, data base technology, information science and visualization. Techniques like neural networks, support vector machines, fuzzy and rough set theory from other disciplines are often used depending upon the data mining approach used.

Meteorological data mining is a form of Data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Useful knowledge can play important role in understanding the

climate variability and climate prediction. In turn, this understanding can be used to support many important sectors that are affected by climate like agriculture, vegetation, water resources and tourism.

II. Related works

Kannan, et al. [4] used some of the data mining functionalities such as: classification, clustering and regression, they classified what is the reason for rainfall fall in the ground level. They computed values for rainfall fall using five years input data by using Pearson correlation coefficient and predicted for future years rainfall fall in ground level by multiple linear regression. However their predicted values are lie below the computed values. So, it did not show an accurate but show an approximate value.

Poorani and Brindha [5] employed principal component analysis (PCA) and ANN on monthly rainfall data, and the experimental Results showed that PCA has some more benefits over ANN in analyzing climatic time series such as rainfall, particularly with regards to the interpretability of the extracted signals.

Andrew et al. [6] studied the effect of rainfall on local water quantity and quality in a watershed basin at Oxford, Iowa, based on radar reflectivity and tipping bucket (TB) data. They used five data-mining algorithms, neural network, random forest, classification and regression tree, support vector machine, and k-nearest neighbor to build rainfall prediction modes. Three Models are selected for all future time horizons. Model I is the baseline model constructed from radar data covering oxford. Model II predicts rainfall from radar and TB data collected at south Amana (16 km west oxford) and Iowa City (25 km east of oxford). Among 5 algorithms MLP neural network has the best performance in comparison to other algorithms. Their computation results indicated that the three models had a similar performance in predicting rainfall at current time, and model II was more than the other models in predicting rainfall at future time horizons. Different lags like t+15, t+30, t+45, t+60, t+75, t+90, t+105, t+120 were considered. The longest acceptable prediction horizon is 120 min.

Sivaramakrishnan and Meganathan [7] attempted to predict spot rainfall for an interior station Trichirappalli (10º48! N/78º41! E) of south India by using association rule mining, The data is filtered using discretization approach based on the best fit ranges and then association mining is performed on dataset using Predictive Apriori algorithm and then the data need be validated using K* classifier approach. The results showed that the overall classification accuracy for occurrence and non-occurrence of the rainfall on wet and dry days using the data mining technique is satisfactory.

Petre [8] tried to predict the average temperature for a future month and developed Classification and Regression Trees (CART) for weather data collection registered over Hong Kong between 2002 and 2005.

Ingsrisawang et al. [9] proposed machine-learning approaches for short-term rain forecasting system. Decision Tree, Artificial Neural Network (ANN), and Support Vector Machine (SVM) were applied to develop classification and prediction models for rainfall forecasts in the northeastern part of Thailand. They used datasets collected during 2004-2006. There are three main parts in their work. Firstly, a decision tree induction algorithm (C4.5) was used to classify the rain status into either rain or no-rain. The overall accuracy of classification tree achieves 94.41% with the five-fold cross validation. The C4.5 algorithm was also used to classify the rain amount into three classes as no-rain (0-0.1 mm.), few-rain (0.1- 10 mm.), and moderate-rain (>10 mm.) and the overall accuracy of classification tree achieves 62.57%. Secondly, an ANN was applied to predict the rainfall amount and the root mean square error (RMSE) were used to measure the training and testing errors of the ANN. It is found that the ANN yields a lower RMSE at 0.171 for daily rainfall estimates, when compared to next-day and next-2-day estimation. Thirdly, the ANN and SVM techniques were also used to classify the rain amount into three classes as no-rain, few-rain, and moderate-rain as above. The results achieved in 68.15% and 69.10% of overall accuracy of same-day prediction for the ANN and SVM models, respectively. The obtained results illustrated the comparison of the predictive power of different methods for rainfall estimation.

Khandelwal and Davey [10] applied data mining techniques regression analysis on the rainfall dataset of Jaipur city. They specifically used multiple regression analysis to predict rainfall of a year by using different 4 climatic factors temperature, humidity, pressure and sea level. For selecting those factors they applied correlation analysis.

Dutta and Tahbilder [11] used Multiple Linear Regression on six years Meteorological data (2007-2012) for Guwahati, Assam, India. The model considered maximum temperature, minimum temperature, wind speed, Mean sea level as Predictors. Experiments results showed that the prediction model based on multiple linear regression achieved 63% accuracy in variation of rainfall.

Wang and Sheng [12] proposed Generalized Regression neural network model for annual rainfall in Zhengzhou. The results of GRNN have more advantage in fitting and prediction compared with BP neural network and stepwise regression analysis methods. The simulation results of GRNN for annual rainfall are better than that of BP neural network. Accuracy predicted using GRNN is better than BP. The stepwise regression method is inferior to both BP and GRNN in accuracy of simulation and prediction results. GRNN network structure is simple and stable.

Olaiya and deyemo [13] investigated the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed. That was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria. A data model for the meteorological data was developed and this was used to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics, and the algorithm, which gave the best results were used to generate classification rules for the mean weather variables. A predictive Neural Network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods. The results show that given enough case data, Data Mining techniques can be used for weather forecasting and climate change studies.

Sethi and Garg [14] used multiple linear regression (MLR) technique for the early rainfall prediction. The model is implemented with the use 30 years (1973-2002) datasets of the climate data such as rainfall precipitation, vapor pressure, average temperature, and cloud cover over Udaipur City, Rajasthan, India. The model forecasts monthly rainfall amount of July (in mm). The experimental results proved that there is a close agreement between the predicted and actual rainfall amount prediction of rainfall.

Singhratina et al. [15] described the development of a statistical forecasting method for SMR over Thailand using multiple linear regression and local polynomial-based nonparametric approaches. SST, sea level pressure (SLP), wind speed, EiNino Southern Oscillation Index (ENSO), and IOD were chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall was 0.6.

Zaw and Naing [16] presented the MPR technique, an effective way to describe complex nonlinear I/P-O/P relationship for prediction of rainfall and then compared the MPR and MLR technique based on the accuracy.

Kajornrit et al. [17] proposed fuzzy inference system for monthly rainfall prediction in the northeast region of Thailand. The predicted performance of the proposed model was compared to be conventional Box-Jenkins and artificial neural networks model. Accordingly, the experimental results show the modular FIS is good alternative method to predict accurately. The predicted mechanism can be interpreted through fuzzy rules. Auto-regression, Seasonal auto regressive integrated moving average and ANN modular FIS provide better results. The experimental results provide both accurate results and human-understandable prediction mechanism.

Nikam and Meshram [18] proposed Bayesian model for rainfall prediction. Since Bayesian prediction model can easily learn new classes. The accuracy also grows with the increase of learning data. Bayesian model issue is that if the predictor category is not present in the training data, the model assumes that a new record with that category has zero probability. According to this paper, Bayesian model for rainfall prediction provides good accuracy. The features used station level pressure, mean sea level pressure, temperature, relative humidity, vapor pressure, wind speed and rainfall. Some of features is being ignored which are less relevant features in the dataset for model computation.

Ji et al. [19] proposed CART and C4.5 to predict rainfall. To correctly perform rainfall prediction, the chance of rain is first determined. Then, hourly rainfall prediction is performed only if there is any chance of rain. 13 variables are considered, they are wind direction, wind speed, wind gust, outdoor humidity, outdoor temperature, evaporation, solar radiation, wind chill, dew point, pressure altitude, cloud base, air density, vapor pressure. The proposed model would be useful for predicting the chance of rain and estimating hourly rainfall in any geographical regions time-efficiently. CART predicted accurately 99.2% and C4.5 predicted accurately 99.3%. And the average prediction accuracy of estimating hourly rainfall with CART and C4.5 are 92.8% and 93.4% correspondingly. CART and C4.5 both have high accuracy and are efficient algorithm.

In this research different predictors (Station, Wind D, Date, Humidity, Min-T, Max-T and Wind S) have been adopted for rainfall forecasting.

III. Intelligent Data Analysis: Methodologies Used

We use the following methods to create the prediction models:

A. Gaussian Processes

GP is based on the assumption that observations follow a normally distributed stochastic process. This leads to the conclusion, that new observations do not change the probability distribution of earlier ones. Based on this simple property Gaussian process regression allows predictions for unknown values [20]. A Gaussian process is stochastic process, any linear functional applied to the sample function Xt will give a normally distributed result. We can write:

$$f \sim GP(m, K) \quad (1)$$

That mean the random function f is distributed as a GP with mean function m and covariance function K.

B. Linear Regression

Linear Regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models [21]. In the case of prediction or forecasting, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y [22]. If we have a data set $\{y_i, x_{i1}, \cdots, x_{ip}\}\, i = 1$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressors x_i is linear. This relationship is modeled through a disturbance term or error variable ε_i — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form:

$$y = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = x_i^T \beta + \varepsilon_i \quad (2)$$

Where: $i = 1 \cdots n$,T denotes the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β. often these n equations are stacked together and written in vector form as

$$Y = X\beta + \varepsilon \quad (3)$$

Where:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11} \cdots x_{1p} \\ x_{21} \cdots x_{2p} \\ \vdots \quad \ddots \quad \vdots \\ x_{n1} \quad \cdots \quad x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix},$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

C. Multilayer Perceptron

Multilayer Perceptron is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable. It has the ability to cope with the nonlinearities, the speed of computation, the learning capacity and the accuracy made them valuable tools for Time series prediction [23].

D. IBk

IBk is a k-nearest-neighbour classifier that uses the same distance metric. The number of nearest neighbours 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 neighbours. A linear search is the default but further options include KD-trees, ball trees, and so-called "cover trees" [24].

E. KStar

KStar algorithm can be defined as a method of cluster analysis, which mainly aims at the partition of "n" observation into "k" clusters in which each observation belongs to the cluster with the nearest mean. We can describe K* algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. K* is a simple, instance based classifier, similar to KNearest Neighbour (K-NN) [24].

New data instances, x, are assigned to the class that occurs most frequently amongst the k-nearest data points, y_j where j = 1, 2…k. Entropic distance is then used to retrieve the most similar instances from the data set. By means of entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values. The K* function can be calculated as:

$$K^*(y_i, x) = -\ln P^*(y_i, x) \quad (4)$$

Where P* is the probability of all transformational paths from instance x to y.

F. Additive Regression

Is a kind of algorithm for numeric prediction that can build standard regression model (eg. tree) and gather residuals, learn model predicting residuals (eg. tree), and repeat. To predict, it simply sum up individual predictions from all models and also it minimizes squared error of ensemble if base learner minimizes squared error

G. Bagging

Is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach [25].

H. Random SubSpace

Random SubSpace is an ensemble classifier that consists of several classifiers and outputs the class based on the outputs of these individual classifiers. Random subspace method is a generalization of the random forest algorithm. Whereas random forests are composed of decision trees, a random subspace classifier can be composed from any underlying classifiers. Random subspace method has been used for linear classifiers, support vector machines, nearest neighbours and other types of classifiers. This method is also applicable to one-class classifiers [26].

I. Regression by Discretization

Regression by Discretization based on Random Forest (RD-RF). This is a regression scheme that employs a classifier (random forest, in this case) on a copy of the data which have the property/activity value discretized with equal width. The predicted value is the expected value of the mean class value for each discretized interval (based on the predicted probabilities for each interval) [27].

J. Decision Table

Decision Table is precise yet compact way to model complicated logic. Decision tables, like flowcharts and if-then-else and switch-case statements, associate conditions with actions to perform. Each decision corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. Each action is a procedure or operation to perform, and the entries specify whether (or in what order) the action is to be performed for the set of condition alternatives the entry corresponds to. Many decision tables include in their condition alternatives the don't care symbol, a hyphen. Using don't cares can simplify decision tables, especially when a given condition has little influence on the actions to be performed. In some cases, entire conditions thought to be important initially are found to be irrelevant when none of the conditions influence which actions are performed.

K. M5Rules

It generates a decision list for regression problems using separate-and-conquer. In each iteration, it builds a model tree using M5 and makes the "best" leaf into a rule.

L. M5P

Is a model tree is generated in two stages. The first builds an ordinary decision tree, using as splitting criterion the maximization of the intra-subset variation of the target value. The second prunes this tree back by replacing sub-trees with

linear regression functions wherever this seems appropriate. M5rules algorithm produces propositional regression rules in IF-THEN rule format using routines for generating a decision list from M5´Model trees [28]. This model tree is used for numeric prediction and at each leaf it stores a linear regression model that predicts the class value of instances that reach the leaf. In determining which attribute is the best to split the portion T of the training data that reaches a particular node the splitting criterion is used. The standard deviation of the class in T is treated as a measure of the error at that node and each attribute at that node is tested by calculating the expected reduction in error. The attribute that is chosen for splitting maximizes the expected error reduction at that node. The standard deviation reduction (SDR), which is calculated by (5) is the expected error reduction.

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} \times sd(T_i) \quad (5)$$

Where Ti corresponds to T1, T2, T3 ... sets that result from splitting the node according to the chosen attribute. The linear regression models at the leaves predict continuous numeric attributes. They are similar to piecewise linear functions and when finally they are combined a non-linear function is formed [29]. The aim is to construct a model that relates a target value of the training cases to the values of their input attributes. The quality of the model will generally be measured by the accuracy with which it predicts the target values of the unseen cases. The splitting process terminates when the standard deviation is only a small fraction less than the standard deviation of the original instance set or when a few instances remain.

M. REPTree

Is a fast decision tree learner. Builds a decision/regression tree using entropy as impurity measure and prunes it. Only sorts values for numeric attributes once [30]. With the help of this method, complexity of decision tree model is decreased by "reduced error pruning method" and the error arising from variance is reduced [3].

Let Y and X be the discrete variables that have the values {y1, …, yn} and {x1, …, xn}. In this case, entropy and conditional entropy of Y are calculated as shown in equation (6) and (7). After that, information gain of X is calculated as shown in equation (8).

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log P(Y = y_i) \quad (6)$$

$$H(Y \mid X) = -\sum_{i=1}^{l} P(X = x_i) H(Y \mid X = x_i) \quad (7)$$

$$IG(Y; X) = H(Y) - H(Y \mid X) \quad (8)$$

In decision trees, pruning is done in two ways. These are pre-pruning and post-pruning. If the number of instances that reach a node is lower than the percentage of the training set, that node is not divided. It is considered that variance of the model, which is generated by the training with a small number of instances, and accordingly the generalization error will increase. For this reason, if the expansion of the tree is stopped when building the tree, then this is called pre-pruning. Another way of building simple tress is post-pruning. Generally, post-pruning gives better results than pre-pruning in practice [31]. Since the tree does not take steps backward and continues to expand steadily while it is being built, the variance increases. Post-pruning is a way to avoid this situation. In order to do this, firstly, unnecessary sub-trees should be found and pruned.

In post-pruning, the tree is expanded until all the leaves are pure and there is no error in training set. After that, we find the sub-trees that lead to memorizing and prune them. In order to this, we firstly use a major part of training set as growing set and the remaining part as pruning set. Later, we replace each sub-tree with a leaf that is trained by the instances which are covered by the training set of that sub-tree and then we compare these two options on pruning set. If the leaf does not lead to more errors on pruning set, we prune the sub-tree and use the leaf; otherwise we keep the sub-tree [32, 33]. When we compare and contrast pre-pruning and post-pruning, we see that pre-pruning produces faster trees, on the other hand, post-pruning produces more successful trees [31].

N. UserClassifier

Is special in that it is interactive and lets the user to construct their own decision tree classifier. For the UserClassifier it is best to have numeric attributes because they can be well represented in pixel plots. In the UserClassifier the nodes in the decision tree are not simple tests on attribute values, but are regions the user interactively selects in these plots. So if an instance lies inside the region it follows one branch of the tree, if it lies outside the region it follows the other branch. Therefore each node has only two branches going down from it [34]. We applied the following steps that appear in figure (1) for experiments and analysis.

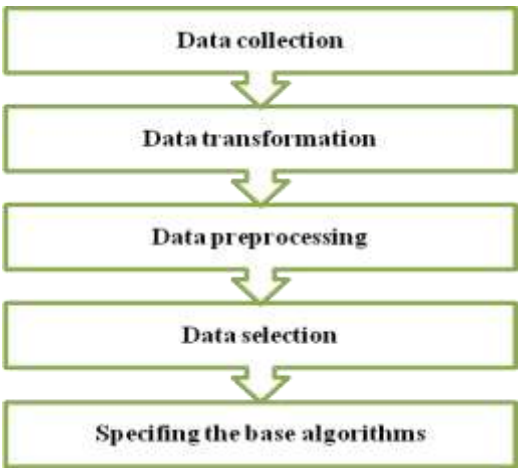

Figure 1. The Work methodology

IV. Data collection and Analysis

The meteorological data that is used in this work has been brought from Central Bureau of Statistics, Sudan for 13 years from 2000 to 2012 for 24 meteorological stations over the country. These stations are: (Khartoum, Dongola, Atbara, Abu Hamad, Karima, Wadi Halfa, Wad Medani, El Deweim, Kassala, Port Sudan, El Gadarif, Elobied, El Nihood, Kadugli, Nyala, Elgeneina, El Fashir, Kosti, El damazen, New Halfa, Babanusa, Rashad, Abu Naam, Sinnar). The dataset had eight (8) attributes containing monthly averages, their type and description is presented in Table 1, while Table 2 shows the analysis of numeric data values.

Attribute	Type	Description
Station	Nominal	Name of meteorological station
Date	Nominal	Month considered
Max-T	Numerical	Monthly Average maximum temperature in centigrade degrees
Min-T	Numerical	Monthly Average minimum temperature in centigrade degrees
Humidity	Numerical	Relative humidity
Wind D	Nominal	Wind direction
Wind S	Numerical	Wind speed (in knot)
Rain	Numerical	Monthly rainfall (in mms)

Table 1. Attributes of Meteorological Dataset.

Attribute	Max	Min	Average	Standard deviation
Station	-	-	-	-
Date	-	-	-	-
Max-T	47.1	20.6	36.43	4.19
Min-T	42.5	7.3	21.63	4.41
Humidity	88	3	36.21	18.11
Wind D	-	-	-	-
Wind S	70	1	5.78	3.40
Rain	295.3	0	26.16	52.4

Table 2. Analysis of numeric data values

A. Data Transformation

It is the stage in which the data is transformed into forms appropriate for data mining. Firstly we converted the hard copy dataset into soft copy. After that the separated tables for 24 stations have been aggregated in one dataset. Also we expressed of rainfall by 0 and 1, zero if rainfall amount is less than or equal 0.1 mms and one if it's greater than 0.1mms. Then the data file was saved in Commas Separated Value (CVS) file format.

B. Data Preprocessing

The data obtained till now is noisy and there are some missing values and some unwanted data. We have to clean the data by filling missing values and removing the irrelevant data. 028135as shown in table 3, we found that all missing values fall in only two attributes Wind-D and Wind-S with total ratio 0. In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data.

Attribute	Missing values
Station	0%
Date	0%
Max-T	0%
Min-T	0%
Humidity	0%
Wind D	0.073955%
Wind S	0.151125%
Rain	0%

Table 3. Analysis of numeric data values.

C. Data Selection

At this stage, after the data preprocessing we have to select the data, which are relevant to our analysis and left all other data. All enable Attribute Evaluators with different Search methods (BestFirst, EvolutionarySearch, ExhaustiveSearch, GeneticSearch, GreedyStepwise, LinearForwardSelection, PSOSearch, RandomSearch, ScatterSearchV1, SubsetSizeForwardSelection, TabuSearch, and Ranker) of attribute selection have been applied to determine the important attributes. We obtained the 4 flowing choices:

- All 7 attributes (station, Wind D, Date, Humidity, Min-T, Max-T, Wind S).
- 4 attributes (Date, Min-T, Humidity, Wind D).
- 3attributes (Min-T, Humidity, Wind D).
- And 1 attribute (Wind D).

D. Specifying the base classifiers

The prediction algorithms should use one of the 4 choices of dataset. The testing methods adopted for this research were the four test options: Use training set, cross validation fold, percentage split and Supplied test set (70% for training and 30%for testing"). Selection of classifier for use in prediction is a challenge. To select the best classifier comparisons can be made on various aspects of the classifiers. The key objective of this paper was to compare performance of all used classifiers. Finally for selecting the appropriate algorithms that produce best models for the rainfall forecasting, the following performance metrics were used:

- Correlation Coefficient: This measures the statistical correlation between the predicted and actual values. This method is unique in that it does not change with a scale in values for the test cases. A higher number means a better model, with a 1 meaning a perfect statistical correlation and a 0 meaning there is no correlation at all. Karl Pearson's [33] correlation coefficient formula is used and it is shown in equation (9).

$$R_{x,y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\ \left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\ \sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \qquad (9)$$

- Mean Absolute Error: Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value (predicted) and its corresponding correct value (actual). MSE calculations are shown in equation (10).

$$MAE = \frac{|a_1 - c_1| + \cdots + |a_n - c_n|}{n} \qquad (10)$$

Assuming that the actual output is a, expected output is c.

- The Root mean-squared Error is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values. Error rate of an estimator arises just because of an arbitrary estimation or lack of information that may provide an accurate estimation [35]. RMSE is shown in equation (11).

$$RMSE = \sqrt{\frac{(a_1 - c_1)^2 + \cdots + (a_n - c_n)^2}{n}} \quad (11)$$

If the values of MAE and RMSE rates are closer to zero, the error rates will be lower. In addition, acceptable error values for MSE and RMSE are different for each learning problem.

V. Experimental Results

After attributes selection process and creating rain-forecasting models, our experimental results show that, the results of 4 and 7 attributes are more accurate than others, and they are close to each other. Table 4 summarizes the performance metrics for the four choices of meteorological dataset.

No of attributes	Average correlation coefficient	Average MAE	Average RMSE
1	0.508215	0.312182	0.401687
3	0.560234	0.277623	0.373868
4	0.590222	0.252397	0.355354
7	0.572118	0.260825	0.35856

Table 4. The performance of the 4 choices of attributes.

The 4 attributes dataset choice has been selected as the rainfall predictors, because it has the maximum average of correlation coefficient, and minimum average of both Mean absolute error and Root mean squared error. Thus the others choices has been excluded. Beyond that we adopted the prediction models, which used the algorithms that applied with (Supplied test set) as the most accurate models, because they have the maximum Average of correlation coefficient as shown in Table 5. Finally we determine the base algorithms according to their correlation coefficient too (the ratio of correlation coefficient must be greater than (0.8) and eliminating the others. We obtained 14 base algorithms; all have adopted Supplied test set as test option. Table 6, shows the best 14 algorithms with their performance metrics.

Test option	Average correlation coefficient	Average MAE	Average RMSE
Cross validation	0.571464	0.253876	0.362064
Percentage split	0.596612	0.245464	0.352136
Supplied test set	0.602664	0.257224	0.363116

Table 5. The performance for the prediction algorithms according to test option.

Algorithm	Average correlation coefficient	Average MAE	Average RMSE
Gaussian Processes	0.8656	0.1638	0.2512
Linear Regression	0.8642	0.1643	0.2527
Multilayer Perceptron	0.8594	0.1327	0.2654
IBk	0.8192	0.0905	0.3005
KStar	0.8901	0.1091	0.2285
Additive Regression	0.8052	0.196	0.2965
Bagging	0.8529	0.1237	0.2614
Random SubSpace	0.8651	0.1636	0.2563
Regression by Discretization	0.8154	0.1397	0.2914
Decision Table	0.8351	0.1219	0.2775
M5Rules	0.8642	0.1113	0.2529
M5P	0.8863	0.1047	0.2322
REPTree	0.8262	0.1286	0.2841
User Classifier	0.8801	0.2352	0.32

Table 6. Performance of the base algorithms

VI. Discussions

According to the results that obtained from the attribute selection process, we applied different prediction algorithms on our rainfall dataset with different number of attributes. In Table 4 we can observe that datasets which contain one and three attributes gave the worse results, minimum averages of correlation coefficient and maximum averages for the both mean absolute error and root mean squared error. The worse results are those that belong to dataset of one attribute. The dataset of seven attributes has improved performance to some extent and gave better results than the two preceding, but the best results belong to dataset of four attributes, it has the maximum average of correlation coefficient 0.590222, minimum average of mean absolute error 0.252397 and minimum average of root mean squared error 0.355354. Thus we can conclude that the most influencing variables that affect on rainfall in Sudan out of the previous seven predictors are: Date, Minimum Temperature, Humidity, and Wind Direction.

In this study, all the test options have been adopted with the different prediction algorithms, as shown in Table 5, cross validation test option give the minimum average of correlation coefficient 0.571464, but it give less average of both mean absolute error 0.253876 and root mean squared error 0.362064 than Supplied test set option. On the other side Supplied test set option provides the maximum average of correlation coefficient 0.602664, however it provides the worse averages of both mean absolute error and root mean squared error 0.257224, 0.363116 respectively. In the case of Percentage split test option we obtained the minimum averages for the both mean absolute error 0.245464 and root mean squared error 0.352136, but we also obtained average of correlation coefficient 0.596612 that is less than Supplied test set option and greater than cross validation test option.

According to the Experimental results that appear in the Table 6, we find that KStar algorithm has the maximum correlation coefficient 0.8901, the minimum root mean squared error 0.2285 and the third lower mean absolute error 0.1091. M5P algorithm comes in second place after KStar as the second highest correlation coefficient 0.8863; the second

less mean absolute error 0.1047 and second less root mean squared error 0.2322. User Classifier algorithm comes in third place in terms of the standard correlation coefficient 0.8801, but it's the worst on both levels of mean absolute error 0.2352 and root mean squared error 0.32. IBK algorithm has the minimum mean absolute error 0.0905, but at the same time it has the second biggest root mean squared error 0.3005 and unsatisfactory correlation coefficient 0.8192 compared with the other selected algorithms

VII. Conclusions and Future Work

The 14 base algorithms considered Date, Minimum Temperature, Humidity and Wind Direction as predictors for the rainfall, and they have adopted Supplied test set as test option. The Correlation coefficient of all base classifiers is greater than 0.8. In this study we considered only seven predictors for rainfall prediction, if we use some more climate factors such as atmosphere pressure, sea surface temperature, etc, so we may obtain more accurate prediction. Also if Ensemble methods have been applied the results may be improved.

References

[1] N.Bushara and A. Abraham, "Computational Intelligence in Weather Forecasting: A Review", Journal *of Network and Innovative Computing 1*, pp. 320-331, 2013.

[2] Z. Ismail, A. Yahya and A. Shabri, "Forecasting Gold Prices Using Multiple Linear Regression Method", *American Journal of Applied Sciences 6 (8)*, pp. 1509-1514, 2009.

[3] I. H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools and Techniques*, 2nd Edition, Elsevier Inc., 2005.

[4] M. Kannan, S. Prabhakaran and P. Ramachandran, "Rainfall Forecasting Using Data Mining Technique", *International Journal of Engineering and Technology, 2 (6)*, pp. 397-401, 2010.

[5] K. Poorani and K Brindha, "Data Mining Based on Principal Component Analysis for Rainfall Forecasting in India", *International Journal of Advanced Research in Computer Science and Software Engineering 3 (9)*, pp. 1254-1256, 2013.

[6] A. Kusiak, X. Wei, A. P. Verma and E. Roz, "Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach", *IEEE Transactions on Geoscience and Remote Sensing, Vol.51 (4)*, pp. 2337 – 2342, 2013.

[7] T. R. Sivaramakrishnan and S. Meganathan, "Point Rainfall Prediction using Data Mining Technique", Research *Journal of Applied Sciences, Engineering and Technology 4 (13)*, pp.1899-1902, 2012.

[8] Elia Georgiana Petre, "A Decision Tree for Weather Prediction", *Petroleum - Gas University of Ploiesti Bulletin, Mathematics, Vol.61 (1)*, pp. 77-83, 2009.

[9] L. Ingsrisawang, S. Ingsriswang, S. Somchit, P. Aungsuratana, and W. Khantiyanan, "Machine Learning Techniques for Short-Term Rain Forecasting System in the Northeastern Part of Thailand", *World Academy of Science, Engineering and Technology 43*, pp. 248-254, 2008.

[10] 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 2 (5)*, pp. 1695-1697, 2012.

[11] P. S. Dutta and H. Tahbilder, "Prediction of Rainfall Using Data mining Technique over Assam", *Indian Journal of Computer Science and Engineering 5 (2)*, pp. 85-90, 2014.

[12] G.Shoba and G.Shobha, "Rainfall Prediction Using Data Mining techniques: A Survey", *International Journal of Engineering and Computer Science 3 (5)*, pp. 6206-6211, 2014.

[13] F. Olaiya and A. B. Adeyemo, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", *International Journal of Information Engineering and Electronic Business4 (1)*, pp. 51-59, 2012.

[14] N. Sethi and K. Garg, "Exploiting Data Mining Technique for Rainfall Prediction", *International Journal of Computer Science and Information Technologies5 (3)*, pp. 3982-3984, 2014.

[15] N. Singhratina, B. Rajagop, M. Clark and K. Kumar "Seasonal Forecasting of Thailand Summer Monsoon Rainfall", *International Journal of Climatology 25 (5)*, pp. 649-664, 2005.

[16] W. Zaw and T. Naing, "Empirical Statistical Modeling of Rainfall Prediction over Myanmar", *World Academy of Science, Engineering and Technology 2*, pp. 492-495, 2008.

[17] J. Kajornrit, K. Wong and C.Fung, "Rainfall Prediction in the Northeast Region of Thailand Using Modular Fuzzy Inference System", In *Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ*, pp. 1-6, 2012.

[18] Valmik. B. Nikam and B.B.Meshram, "Modeling Rainfall Prediction Using Data Mining Method", In *Proceedings of the Fifth International Conference on Computational Intelligence, Modeling and Simulation*, pp132-136, 2013.

[19] S. Ji, S. Sharma , B. Yu and D. Jeong, "Designing a Rule-Based Hourly Rainfall Prediction Model", In *Proceedings of the IEEE 13th International Conference on Information Reuse and Integration (IRI)*, pp303-308, 2012.

[20] C. Rodriguez, J. Pucheta, H. Patino, J. Baumgartner, S. Laboret and V. Sauchelli, "Analysis of a Gaussian process and feed-forward neural networks based filter for forecasting short rainfall time series", *The International Joint Conference on Neural Networks (IJCNN)*, pp. 1 – 6, 2013.

[21] M.Rahman, M.Rafiuddin and M.Alam, "Seasonal forecasting of Bangladesh summer monsoon rainfall using simple multiple regression model", *Journal of Earth System Science 122(2)*, pp. 551–558, 2013.

[22] C. Udomboso and G. Amahia, "Comparative Analysis of Rainfall Prediction Using Statistical Neural Network and Classical Linear Regression Model", *Journal of*

Modern Mathematics and Statistics5 (3), pp. 66-70, 2011.

[23] R. Deshpande, "On The Rainfall Time Series Prediction Using Multilayer Perceptron Artificial Neural Network", *International Journal of Emerging Technology and Advanced Engineering2 (1),* pp. 148-153, 2012.

[24] S. Vijayarani and M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms", International *Journal of Advanced Research in Computer and Communication Engineering, Vol. 2 (8*), pp. 3118-3124, 2013.

[25] D. Jeong and Y. Kim, "Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction", *Hydrological Processes19(19),* pp 3819–3835, 2005.

[26] M. Skurichina and R. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers", *Pattern Analysis & Application 5*, pp. 121-135, 2002.

[27] S. M. Mwagha, M. Muthoni and P. Ochieg, "Comparison of Nearest Neighbor (ibk), Regression by Discretization and Isotonic Regression Classification Algorithms for Precipitation Classes Prediction", *International Journal of Computer Applications96 (21),* pp. 44-48, 2014.

[28] W. Average, S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou and K. Menagias, "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", *International Journal of Computer, Information, Systems and Control Engineering1 (2),* pp. 382-386, 2007.

[29] B. Bhattacharya and D. P. Solomatine, "Neural networks and M5P model trees in modelling water level-discharge relationship", *Neurocomputing 63*, pp. 381-396, 2005.

[30] K. Wisaeng, "A Comparison of Decision Tree Algorithms For UCI Repository Classification", *International Journal of Engineering Trends and Technology 4 (8),* pp. 3393-3397, 2013.

[31] E. Alpaydın, *Introduction to Machine Learning*, The MIT Press, 2004.

[32] J. R. Quinlan, "Simplifying decision trees", *International Journal of Man-Machine Studies 27 (3),* pp. 221 – 234, 1987.

[33] M. Zontul, F. Aydin, G. Dogan, S. Sener and O. Kaynar, "Wind Speed Forecasting Using RepTree and Bagging Methods in Kirklareli-Turkey", *Journal of Theoretical and Applied Information Technology 56 (1),* pp. 17-29, 2013.

[34] S. Peyvandi, H. M.Shirazi and A. Faraahi, "Proposing a Classification Algorithm for User Identification According To User Web Log Analysis", *Australian Journal of Basic and Applied Sciences 5 (9),* pp. 645-652, 2011.

[35] E.L. Lehmann and G. Casella, *Theory of Point Estimation, Second Edition*, Springer: New York, 1998.