International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 23(2), November, 2023, pp 47-62
DOI: https://doi.org/10.3329/ijss.v23i2.70110

Prediction Ranifall Using Machine Learning Algorithms: Empirical Evidence from Bogura and Rangpur District of Bangladesh

Md. Rafiqul Islam, Md. Mostafizur Rahman*, Afsana Afruz, Md. Abdul Khalek and M. Sayedur Rahman

Data Mining and Environmental Research Group, Department of Statistics,
University of Rajshahi, Rajshahi-6205, Bangladesh.
*Correspondence should be addressed to Md. Mostafizur Rahman
(Email: mostafiz_bd21@yahoo.com)

[Received March 19, 2023; Accepted October 12, 2023]

Abstract

Around the world, forecasting rainfall has been regarded as one of the most difficult task. Exact and timely rainfall forecasting may be extremely helpful. By uncovering novel links between the readily available elements of historical data, data mining algorithms may accurately anticipate the amount of rainfall. Therefore, it remains intriguing to forecast rainfall data with both the highest degree of accuracy by combining and improving various data mining approaches in case of different weather stations. In this study we compare the forecasting performance of different data mining techniques such as Classification and Regression Trees (CART), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Random Forest (RF), and Linear Discriminant Analysis (LDA) in case of Bogura and Rangpur district of Bangladesh. For this analysis, the monthly time series data from January 1964 to December 2017 are taken into account. For empirical investigations, the data mining process, including data collection, data pre-processing, modeling, and assessment, is closely adhered to. The empirical study shows that SVM approach is the best option for predicting rainfall in the case of both Bogura and Rangpur district, Bangladesh, for the next time period. The above study will be useful in providing information to support crop, water, and flood control, which will protect people's lives and property and promote economic in its growth.

Keywords and Phrases: CART, SVM, RF, KNN, Bogura and Rangpur.

AMS Classification: 62P12, 62M10.

1. Introduction

Bangladesh's economy is heavily dependent on the production of agricultural products. One of the most crucial elements in this process is rainfall. It’s a natural occurrence that is the result of interactions between a numbers of the complicated climate system. In various situations, such as monitoring aviation operations, planning the management of water resources, alerting for impending flooding, and restricting transportation and construction activities, precise and timely rainfall prediction is helpful (Wu et. al. 2015). The arrogant Himalayas in the north and funnel-shaped Bay of Bengal in the south have made Bangladesh a meeting place of the life- giving monsoon rains and the catastrophic devastation of floods, cyclones, storm surges, droughts etc

(Paramanik,1991). The climatic information serves not only as guide to the selection of the proper sites for a given crop but also the most desirable period for sowing and harvesting (Amin et. al., 2004). The effects of climate change are evident in the uneven rainfall pattern and temperature increases, solar radiation and may due to increased population and developmental activities (Pingale et. al.2015).

One of the main difficulties in creating rainfall forecasting models due to the high degree of uncertainty. Rainfall occurrence is influenced by elements such as temperature, relative humidity, wind speed, wind direction, cloud cover, etc. A method of manipulating and extracting implicit, previously undiscovered, known, and potentially relevant information from data is called machine learning. It encompasses several classifiers of supervised and unsupervised learning that are used to predict and identify the precise forecasting model for the given data set. To predict the maximum temperature, rainfall, evaporation and wind speed for Nigeria Olaiya and Adeyemo (2012) compared the performance of ANN model and decision tree algorithm and found the performance of ANN model over decision tree algorithm for all of these attributes. In order to predict the rainfall data for the Darjeeling rain gauge station in India, Ramana et al. (2013) compared the performance of the WNN model with the ANN model and found that the performance of WNN model than ANN model. Rahman et.al. (2022) investigated the performance of machine learning algorithm to predict the rainfall data in case of Bogura district for the time period January, 1971 to December, 2015 and found Random Forest algorithm is the most suitable algorithm for predicting rainfall data at this time period.

Zainudin et al. (2016) examined the performance of a number of classifiers, including Naive Bayes, Support Vector Machine, Decision Tree, Neural Network, and Random Forest, in order to estimate rainfall in the case of Malaysia and found that the Random Forest model predict rainfall more accurately than other models. Mishra et al. (2018) created one-month and two-month forecasting models for rainfall prediction using monthly rainfall data spanning 141 years from various meteorological stations in the North of India. Solanki and G.P.B. (2018) presented a Hybrid Intelligent System by integrating Artificial Neural Network and Genetic Algorithm and found the better performance of their hybrid model than other existing model for prediction.

Along with these writers, additional authors Talib et al. (2017), Tharun et al. (2018), and others examined the effectiveness of several machine learning algorithms for forecasting rainfall for certain cities or areas. Besides these several authors examined the performance of different machine learning algorithms for predicting rainfall data for example, Aftab et al (2018), Sivapragasam et al (2001), Monira et al (2010), Kannan et al. (2010), Sethi and Garg (2014) etc.

From the above study we found that machine learning technique is successful to predict rainfall and other climatic variables in case of many countries and there are few studies in case of Bangladesh (Rahman et. al. 2021). So, the aim of this study is to compare the forecasting performance of different machine learning algorithms in case of Bogura and Rangpur districts of Bangladesh. The rest of the paper is organized as follows: section 2 present the geographical description of these studies areas, section 3 present the methodology, section 4 present the result and discussion and finally section 5 present the conclusion and recommendation.

2. Description of the Study Area

2.1 Bogura District

Bogura belong to the northern part of Bangladesh. It is a part of the Rajshahi Division. The well-known Karatoya River divides the area's primary waterway. The Bogura District, which spans

2899 square kilometers, is located at 24.780N and 89.350E in the Rajshahi division. Its northern and southern boundaries are formed by the districts of Joypurhat and Gaibandha, while its eastern and western boundaries are formed by Jamalpur and Nagaon. Alluvial soil from the Karatoya River basin makes up the majority of the material (84%), with dirt from Barind making up the rest. The region's temperature fluctuates from 30.6 degrees Celsius to 11.7 degrees Celsius, and there is an average of 1762 mm of rainfall every year (Banglapedia, Bagura)

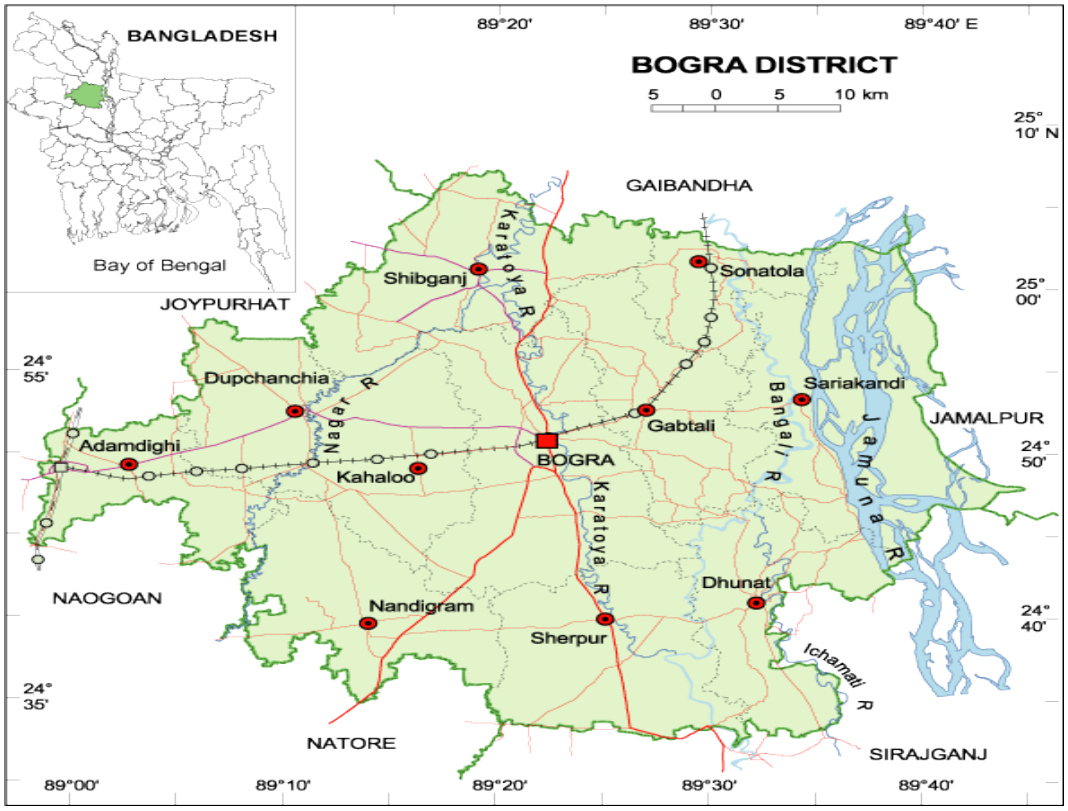


Figure 4.1: Map of Bogura District

2.2 Rangpur District

Rangpur is another district of northern part of Bangladesh and this district belong to Rangpur division. Rangpur District (Rangpur division), with a total area of 2370.45 sq km, is situated between latitudes 25°18' and 25°57' north and 88°56' and 89°32' east. It is bordered on the north by the districts of Nilphamari and Lalmonirhat, on the south by Gaibandha, on the east by Kurigram, and on the west by the districts of Dinajpur. Eighty percent of the soil is alluvial soil from the Teesta River basin, with the remaining twenty percent being Barind soil. The average annual rainfall is 2931 mm, and the temperature ranges from 32 degrees Celsius to 11 degrees Celsius (Banglapedia, Rangpur).

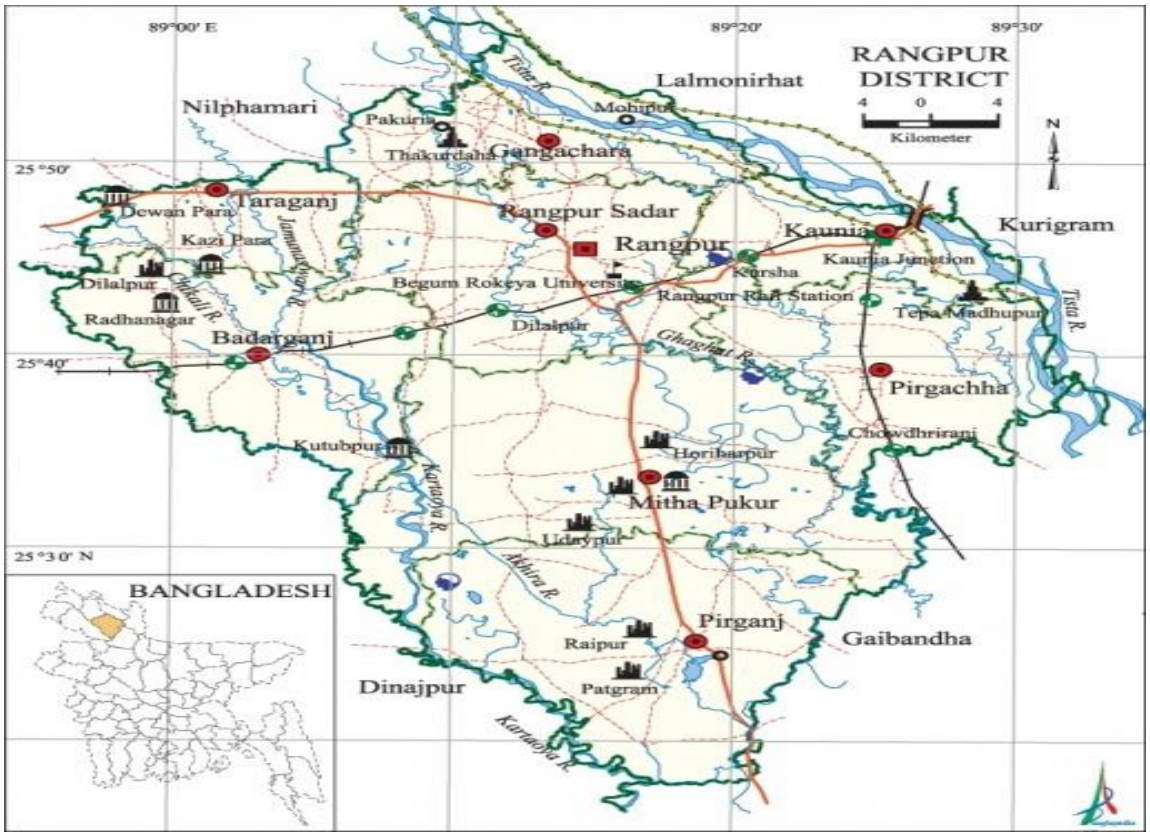
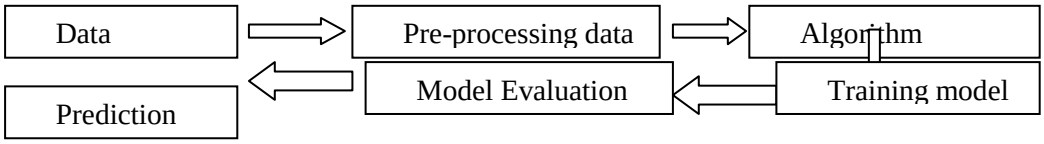


Figure 4.2: Map of Rangpur District

3. Methodology

The aim of this study is to compare the performance of different machine learning algorithms for predicting rainfall data for two districts of Bangladesh. The classical framework, data collection procedure, data pre-processing methods, modeling phases and evaluation phase describes in this section. Machine learning process follows the following steps:



In this we use different machine learning algorithm such Classification and regression trees(CART), Support vector machine(SVM), k-nearest neighbors(K-NN), Random forest(RF) and Linear discriminant analysis(LDA) for predicting rainfall data of these study areas. Data analysis performed by statistical package SPSS, R program and other related software.

3.4 Predictive Models

Recently machine learning algorithms gained popularity for predicting climatic variables. Different algorithms show better predicting performance for different data sets. But still now there is no unique model or algorithms which predict most of the climatic variables properly. So, we need to search predicting performance of different algorithm for different geographic locations. In this study we use several machine algoritms such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Random Forest (RF), and Classification and Regression Trees (CART) for predicting monthly rainfall data from Bogura nd Rangpur district of Bangladesh. These models are described below:

3.4.1 Classification and Regression Trees (CART)

When the classifiers are absolutely binary, with exactly two branches for each decision node, M. Briman et al. (1984) proposed the CART or Classification and Regression Trees approach. A single input variable (x) and a numerical split point on that variable are represented by each root node. A prediction is made using the output variable (y) that is present in the tree's leaf nodes. The CART which is used in machine learning, shows how the values of the target variable may be predicted based on other factors. Each fork of the decision tree is divided into a predictor variable, and at the conclusion of each node is a prediction for the target variable.

Depending on the target value of an attribute, nodes in the decision tree are divided into sub-nodes. The training set is the root node, which is divided into two by taking the best feature and tolerance value into account. Additionally, the subsets are divided according to the same rationale. This continues until the tree has either produced all of its potential leaves or discovered its last pure sub-set.

The technique described below is how the CART algorithm operates:
- Each input's ideal split point is discovered.
- The new "best" split point is determined using the best split points of each input from Step 1.
- Divide the selected input at the "best" split point. • Splitting should continue until a stopping criteria is met or no more acceptable splitting is possible.

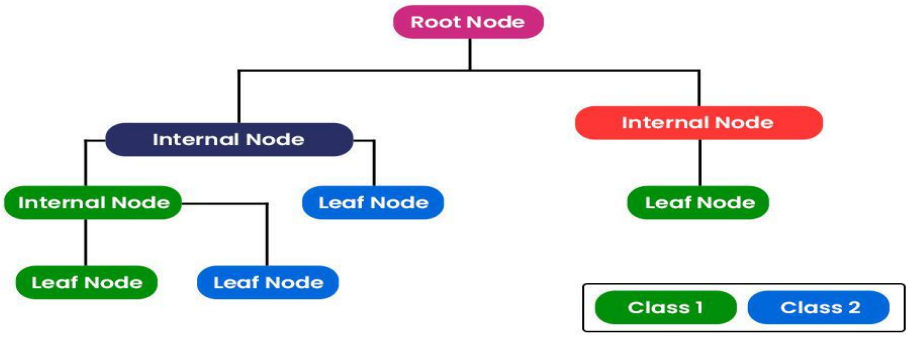


Fig. 3.1: Diagram for decision tree (Source: Brownlee, 2016)

3.4.2 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm which is used for classification and regression analysis. It is proposed by Vapnik with his colleagues (Vapnik, 1995, Boser et al.,

1992). When looking for a high-performing algorithm that requires little adjustment, Support Vector Machines, one of the most well-known machine learning algorithms, were invented in 1990. The support vector machine approach seeks to locate an N-dimensional space hyperplane that clearly categorizes the data points. Figure 3.2 displays a straightforward example of a support vector machine.

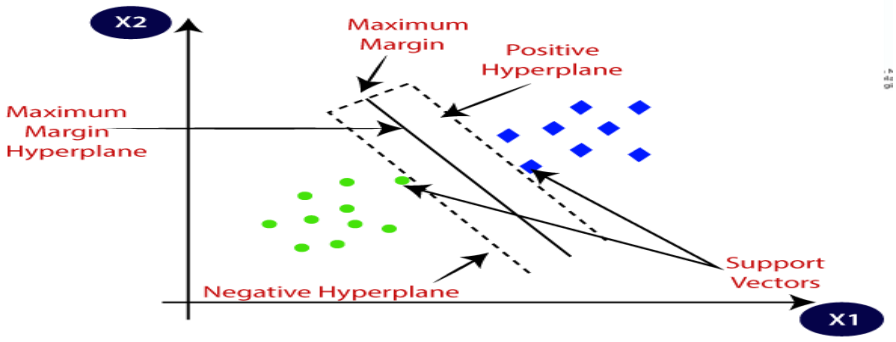


Fig.3.2: Diagram for Support Vector Machine (Source: Gandhi, 2018)

There are a variety of different hyperplanes that might be used to split the two classes of data points. The goal is to identify a plane with the largest margin, or the greatest separation between data points from both classes. To increase the confidence in the classification of next data points, the margin distance should be maximized.

3.4.3 K-Nearest Neighbors (K-NN)

k-Nearest Neighbors (k-NN) algorithm was first developed by Fix and Joseph (1951) and later expanded it by Thomas and Peter (1967). It is supervised learning algorithms. K-NN directly uses the training dataset to create predictions. By exploring the complete training set for the K most similar examples (the neighbors) and summarizing the output variable for those K instances, predictions are created for a new instance (x). This might be the mean output variable in a regression; in a classification, this could be the modal (or most prevalent) class value. A distance metric is applied to identify which of the K examples in the training dataset is closest to a new input. The most common distance metric for input variables with real values is Euclidean distance. Algorithm tweaking can be used to determine K's value. It is typically seen as an odd number. K-computational NN's complexity grows as the amount of the training dataset does. K-NN may be made stochastic for very large training sets by selecting a sample from the training dataset and using that sample to determine the K-most similar occurrences. The following algorithm may be used to describe how the K-NN works:

- **Step-1:** Decide on the neighbors' K-numbers.
- **Step-2:** Calculate the Euclidean distance between K neighbors.
- **Step-3:** Based on the determined Euclidean distance, select the K closest neighbors.
- **Step-4:** Count the number of data points in each category among these k neighbors.
- **Step-5:** Assign the fresh data points to the category where the neighbor count is highest.
- **Step-6:** Our model is complete.
- Let's say we need to classify a new data point in order to use it. Think on the photo below:

3.4.4 Random Forest

Random forest algorithm was proposed by Ho in 1995 (Ho, 1995, 1998). Later the stochastic discrimination approach to classification proposed by Kleinberg (1990, 1996, 2000). An extension of the algorithm was proposed by Breiman (2001). Given the wide range of boosting approaches available for classification tasks, the random forest algorithm is more frequently used in these situations. The random forest approach employs averaging to increase accuracy and avoid overfitting when fitting several randomly generated decision trees to subsets of data for regression. The feature set is divided into subsets, and then each tree is produced. Up until the answer variable is equal to a subset at a certain node, the process is repeated. The least-squares boosting method is used. For each fit, the mean-squared error is minimized. The prediction made using a feature set is just the average of guesses made across all students.

The Random Forest method is illustrated in the picture below:

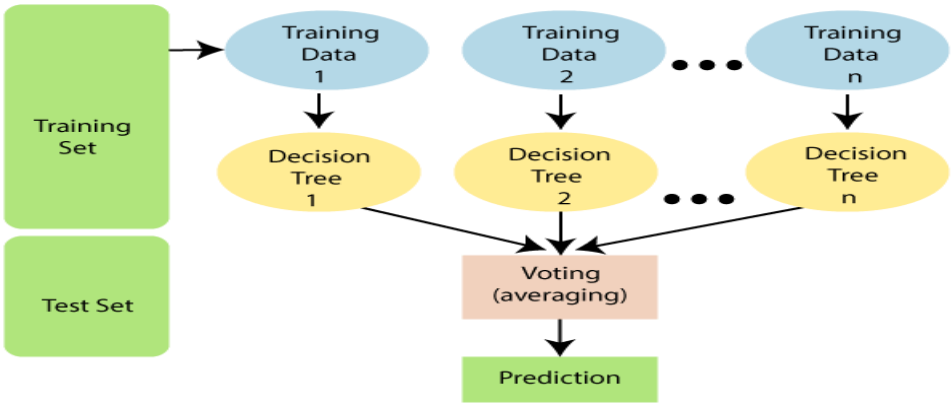


Fig: 3.5 Random forest model

3.4.5 Linear Discriminant Analysis (LDA)

In 1936 R. A. Fisher (1936) developed the dichotomous discriminant analysis. Discriminant function analysis is useful in determining whether a set of variables is effective in predicting category membership (Cohen et al. 2003). The linear Discriminant Analysis is a straightforward model to prepare and use. This is the mean and variance of the variable for each class for a single input variable (x). The means and the covariance matrix, which are derived over the multivariate Gaussian for many variables, have the same characteristics. In order to produce predictions, these statistical characteristics are calculated from the data and entered into the LDA equation. LDA makes a few simplifications.

The data is Gaussian, which means that when plotted, each variable has a bell-shaped form. The variance is the same for all attributes. The LDA model calculates the mean and variance from the data for each class under these presumptions. By calculating the likelihood that a fresh batch of data belongs to each class, LDA creates predictions. A forecast is produced for the output class that has the highest likelihood.

We can quickly convert a 2-D and 3-D graph into a 1-dimensional plane using the dimensionality reduction approach in machine learning known as linear discriminant analysis.

Let's look at an example where we need to efficiently categorize two classes in a 2-D plane with an X-Y axis. LDA allows us to construct a straight line that may entirely divide the two classes of data points, as we have previously seen in the example above. Here, LDA splits an X-Y axis into two separate axes with a straight line, then projects data into the new axis.

As a result, we may shrink the 2-D plane to 1-D and optimize the gap between these classes.

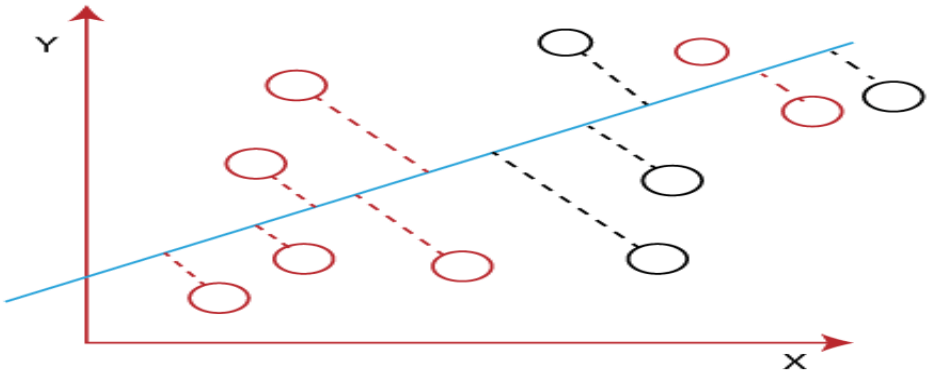


Fig 3.6: LDA plot

Linear Discriminant Analysis use the following standards to develop a new axis:

- It reduces the variation within the particular class.
- It maximizes the distance between the means of two classes.

In order to maximize the distance between the means of the two classes and reduce variance within each, LDA constructs a new axis using the aforementioned two requirements.

3.5 Model Evaluation

The performance of model can be asses by evaluation metrics which is one of the most important tasks for machine learning algorithm. When a model is built then it is necessary to measure how accurately it predicts the expected outcome. We have different evaluation metrics for a different set of machine learning algorithms. To evaluate the forecasting performance of various models, three metrics are typically used: Precision, Recall, and F-Measure. In addition to this, we also take into account overall accuracy, which is the percentage of all forecasts that were accurate.

3.5.1 Precision

In terms of False Positive (FP) entities, the "Precision" compares True Positive (TP) entities. The calculation is as follows: Precision = $\frac{TP}{TP+FP}$, here, TP stands for correctly classified, and FP for incorrectly classified.

3.5.2 Recall

The False Negative (FN) entities are not categorised at all, and the "Recall" analyzes the True Positive (TP) items in relation to the FN entities. This is what it is: Recall = $\frac{TP}{TP+FN}$

3.5.3 F-Measure

The question of which method is superior arises if one has higher accuracy and lesser recall. In this situation, the model cannot be identified by accuracy and recall value alone. F-Measure, which is defined as follows, is used to solve this type of problem.

$$\text{F-measure} = \frac{2(\text{precision})}{\text{precision+recall}}$$

4. Result and Discussion

Two metrological stations Bogura and Rangpur districts are chosen for empirical study. The data used in this thesis was collected from the Bangladesh Meteorological Department (BMD). The daily data covered the period from January 1964 to December 2017. Several atmospheric characteristics, including temperature, humidity, wind speed, sunlight, lowest temperature, and maximum temperature, are included in the input dataset for rainfall prediction. We convert this daily rainfall data in to monthly rainfall data using Microsoft Excel. The attributes temperature, humidity, wind speed, sunshine, minimum temperature and maximum temperature its name, type and measurement unit are given in Table 1.

Table 1: Data and measuring unit

Variables	**Type**	**Measurement**
Rainfall status	Categorical	(yes/no)
Temperature	Continuous	Degrees Celsius
Humidity	Continuous	%
Wind Speed	Continuous	Meters per second
Sunshine	Continuous	Hour
Maximum Temp	Continuous	Degrees Celsius
Minimum Temp	Continuous	Degrees Celsius

4.1 Data Transformation

In our study we consider the monthly rainfall data (rainfall status), temperature, humidity, wind speed, sunshine hour, maximum and minimum temperature. All of these data shows different measuring unit. So, we need to transform these data to make unit free. In this study we use Min-Max Normalization method to transform the data. This transformation method converts all of these numerical variables into 0 to 1 range.

4.2 Data Smoothing

In environmental studies missing data is common. The World Meteorological Organization (WMO) states that although the data collected for each station contains a number of missing values, it is still possible to estimate them successfully because less than 10% of the missing data must be estimated using the Statistical Package for Social Sciences (SPSS) program. These missing numbers were random, and some years also had continuous missing data for one to many months. To get missing data, we used SPSS to average the closest beginning values. We prepare the data for analysis after estimating the missing data using the smoothing SPSS software technique.

4.3 Characteristics of Data

To find out the initial pattern and trend from the data it is necessary to calculate summary statistics which present mean, minimum, maximum, first quartile, third quartile. The estimated summary statistics of these attributes dry bulb, temperature, humidity, wind speed, sunshine hour, maximum and minimum temperature for both districts are given in Table 2.

Table 2: The summary statistics of these variables

		Drybulb	Max.Tem	Min.Tem	Humidity	Sunshine	Windspeeds
Bogura	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1st Qu.	0.3989	0.3859	0.3455	0.5652	0.3561	0.05882
	Median	0.7021	0.5870	0.7277	0.7391	0.5390	0.06765
	Mean	0.5891	0.5324	0.6249	0.6996	0.5044	0.08438
	3rd Qu.	0.7713	0.6630	0.9005	0.8696	0.6598	0.09706
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Rangpur	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.00000
	1st Qu.	0.4192	0.5190	0.3431	0.6552	0.4479	0.05882
	Median	0.7305	0.6571	0.6863	0.7414	0.6068	0.07059
	Mean	0.6450	0.6053	0.6063	0.7115	0.5964	0.08586
	3rd Qu.	0.8802	0.7143	0.8676	0.7931	0.7682	0.09706
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2 shows that mean value of dry bulb, Max.Tem, Min.Tem, humidity, sunshine hour and windspeeds for Bogura district are 0.5891, 0.5324, 0.6249, 0.6996, 0.5044, 0.08438 for respectively and 0.6450, 0.6053, 0.6063, 0.7115, 0.5964, 0.08587 for Rangpur districts which indicates that all of these variables are higher in case of Rangpur district.

4.4 Estimation Result of Predictive models for Bogura

The performance of different data mining techniques, including Support Vector Machine (SVM), Classification and Regression Trees (CART), K-Nearest Neighbors (K-NN), Random Forest (RF), and Linear Discriminant Analysis (LDA), was compared after the pre-processing stage of the process. All of these techniques are regarded as supervised methods. Any supervised machine learning technique's effectiveness may be evaluated by contrasting the results with known classes (pre- classified data). In our study, we use 25% for test set and 75% for training set. The study's data mining algorithms produced accurate findings for all of the accuracy metrics for both the rain-class (yes) and no-rain-class (no) (i.e. Precision, recall and f- measure). Given that it offers the average of recall and precision, the F-measure is a very effective accuracy metric. Besides these we also reported overall accuracy of the predictive models. The model evaluation criteria such as Precision, Recall, F-measure and overall accuracy of all models are given in Table 3. In addition to these three model assessment criteria, we also compare the performance of our models in this work using Box and Whisker plots, density plots, dot plots, and parallel plots.

Table 3: Model evaluation criteria for Bogura district

Model	Class	Precision	Recall	F-measure	Overall accuracy
SVM	Yes	0.8889	0.9175	0.9136	0.8830
	No	0.8167	0.8525	0.9010	
K-NN	Yes	0.7934	0.8258	0.8183	0.8006
	No	0.6435	0.8094	0.8030	
CART	Yes	0.8301	0.8478	0.8501	0.8144
	No	0.8085	0.8338	0.8129	
LDA	Yes	0.8490	0.8308	0.8483	0.8266
	No	0.8071	0.8194	0.8330	
RF	Yes	0.8112	0.8937	0.8989	0.8158
	No	0.8067	0.8129	0.8387	

The estimated result from Table 3 indicate that the Precision value of SVM model shows higher both yes and no class and the other two statistics Recall and F measure also present the higher value in case of SVM model where as K-NN model shows the worst performance based on these statistics. The overall accuracy indicates that the predictive performance in case of SVM model is higher compare to all of the other models. So, from the Table 3 we conclude that SVM model is the most successful model to predict the monthly rainfall data of Bogura district. The predictive performance of these machine learning models also view by graphical plots. The box plot for predictive models is given in Figure 1.

The Figure indicates that the boxes are arranged in ascending order of mean accuracy. In comparison to other models, we discovered that the SVM technique had greater overall accuracy. The density plot shows that the SVM method's density map shows the largest peak. These are helpful charts since they provide the 95% confidence interval (i.e., the range in which 95% of observed scores fall) as well as the mean estimated accuracy. Comparing the means and estimating the overlap of the spreads amongst methods proved to be helpful. Finally, parallel graphs support the SVM method's superior performance to other approaches. For more confirmation we also change the training and test data as different combinations and calculate the summary statistics of the accuracy measurement. The summary statistics of accuracy measurement for all of these models are given in Table 4.

Table 4: Summary for accuracy measurement for Bogura district

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SVM	0.8163265	0.8367347	0.8470833	0.8461173	0.8541667	0.8750000
K-NN	0.7142857	0.7959184	0.8144133	0.8115000	0.8489583	0.8775510
CART	0.7600000	0.7881250	0.8248299	0.8195187	0.8358844	0.8979592
LDA	0.7291667	0.7857143	0.8645833	0.8312823	0.8775510	0.8800000
RF	0.7500000	0.7937500	0.8266667	0.8234796	0.8563988	0.8958333

Table 4 provides the accuracy summaries for the various techniques. When forecasting the rainfall data for Bogura, Bangladesh, the mean statistics for the SVM approach performed better than other models, however the K-NN models performed the poorest.

Therefore, from the above Table and Figure we conclude that SVM algorithm is the most successful algorithms to predict the rainfall status in case of Bogura district during this study period.

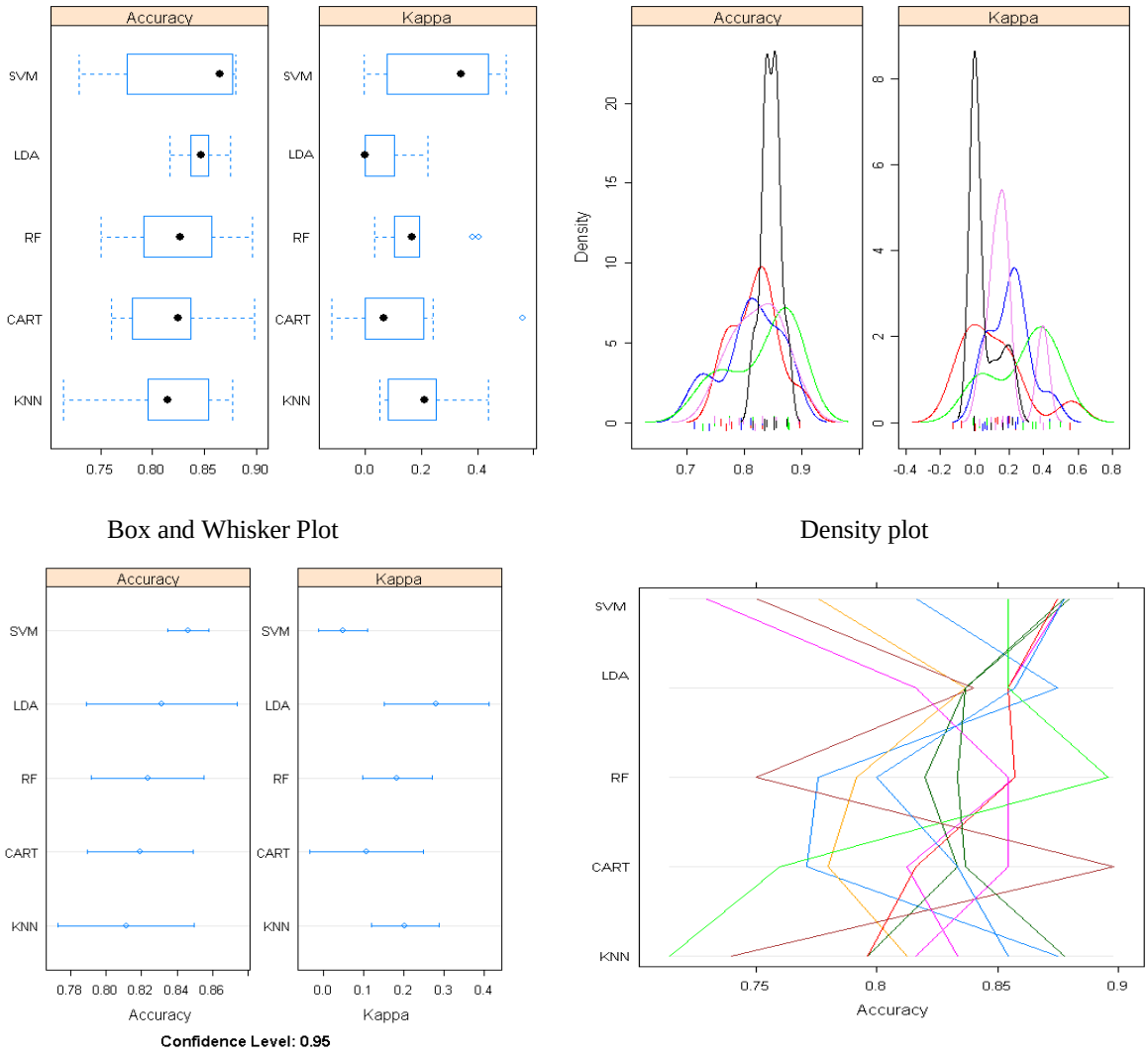


Figure 1: Box and Whisker, density, dot and parallel plot in case of Bogura district

4.5 Estimation Result of Predictive models for Rangpur

The model evaluation criteria such as Precision, Recall, F-measure and overall accuracy of all models in case of Rangpur district is given in Table 5. In addition to these three model assessment criteria, we also compare the performance of our models in this work using Box and Whisker plots, density plots, dot plots, and parallel plots.

The estimated result from Table 5 indicate that the Precision value for SVM model give best predictive result for both yes and no class whereas LDA model provide worst predictive

performance. The Recall value shows that SVM model give highest value for both yes and no class and LDA provide lowest Recall value. Similar results obtained from the F-measure. Finally, the overall accuracy of all the models show that the predictive performance in case of SVM model is high and LDA model is low. So, from the Table 5 we find that SVM model is the most successful model to predict the monthly rainfall data of Rangpur district.

Table 5: Model evaluation criteria for Rangpur data

Model	Class	Precision	Recall	F-measure	Overall accuracy
SVM	Yes	0.9200	0.9487	0.9363	0.9104
	No	0.9109	0.8883	0.9001	
K-NN	Yes	0.9091	0.9287	0.9138	0.9006
	No	0.8817	0.8549	0.8703	
CART	Yes	0.8890	0.9285	0.9082	0.9030
	No	0.9167	0.8594	0.8871	
LDA	Yes	0.8821	0.9157	0.8998	0.8578
	No	0.8676	0.8152	0.8378	
RF	Yes	0.9019	0.9227	0.9183	0.9016
	No	0.8871	0.8594	0.8703	

The predictive performance of these machine learning models also view by graphical plots. The box plot for predictive models is given in Figure 2.

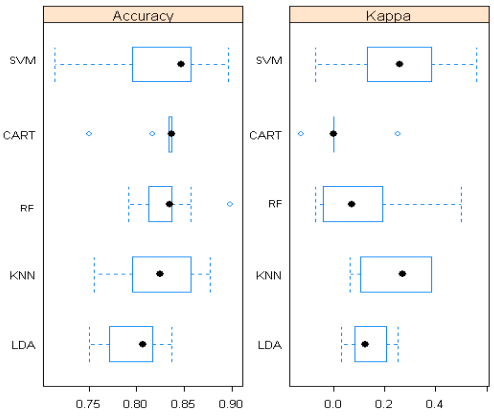
From the Figure 2 we found that the boxes are arranged in ascending order of mean accuracy which shows the SVM technique had better overall accuracy. The density plot of SVM present the largest peak and other two dot plot and parallel plots also shows the better performance of SVM model in case of Rangpur district. The summary statistics of accuracy measurement for all of these models are given in Table 6.

Table 6: Summary for accuracy measurement for Rangpur data

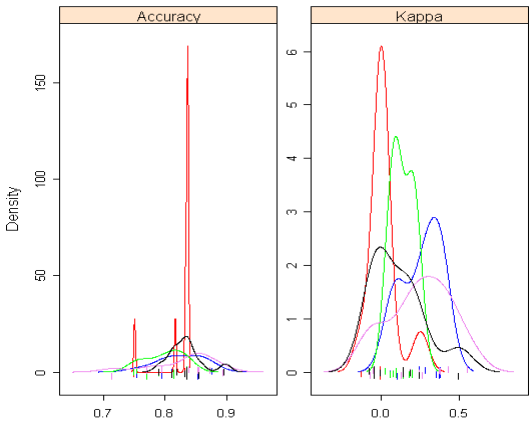
Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SVM	0.7916667	0.8143750	0.8350340	0.8335306	0.8367347	0.8979592
K-NN	0.7551020	0.8000638	0.8248299	0.8255102	0.8563988	0.8775510
CART	0.7500000	0.8333333	0.8367347	0.8253401	0.8367347	0.8367347
LDA	0.7500000	0.7760417	0.8061224	0.7985969	0.8163265	0.8367347
RF	0.7142857	0.8061224	0.8470833	0.8297058	0.8571429	0.8958333

Table 6 provides the accuracy summaries for the various techniques. When forecasting the rainfall data for Rangpur, Bangladesh, the mean statistics for the SVM approach performed better than other models, however the LDA models performed the poorest.

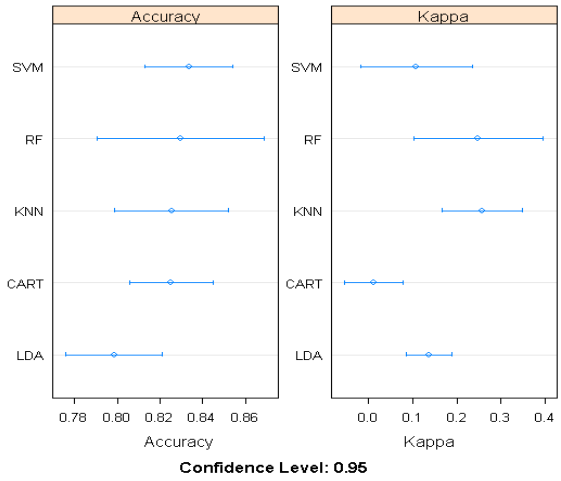
Therefore, from this chapter we found that SVM model is the most successful model to predict monthly data in case of both Bogura and Rangpur districts of Bangladesh during the study period.



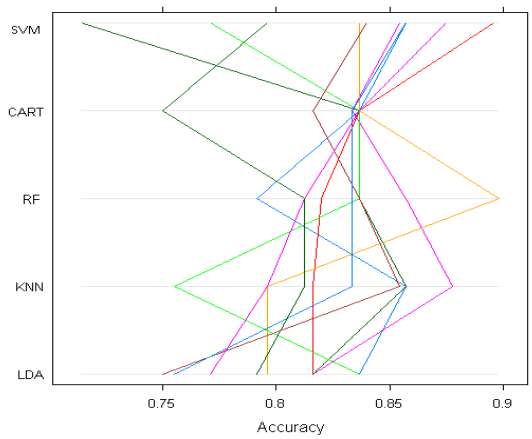
Box and Whisker Plot



Density plot



Dot plot



Parallel plot

Figure 2: Box and Whisker, density, dot and parallel plot in case of Rangpur district

5. Conclusion and Recommendation

Around the world, rainfall has a significant influence on agriculture and the economy. By obtaining and using the hidden information from historical meteorological data, machine learning algorithms accurately anticipate the rainfall data. The performance of various data mining techniques, including Classification and Regression Trees (CART), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Random Forest (RF), and Linear Discriminant Analysis (LDA), for predicting rainfall in the Bogura and Rangpur District of Bangladesh from January 1964 to December 2017 on a monthly basis, is therefore compared in this paper. Pre-processing techniques, which include cleaning and normalization procedures, are employed for better

prediction. The effectiveness of well-known data mining approaches was examined using several graphical comparison tools, as well as metrics for precision, recall, and f-measure. The empirical findings imply that the SVM technique is the best appropriate technique for Bogura and Rangpur District of Bangladesh, this prediction for the ensuing timeframes. The result also shows that K-NN model case of Bogura district and LDA model in case of Rangpur district give worst predicting performance.

The finding of this study will help police maker to take necessary steps to solve the problem of water for continuing sustainable agriculture production of these study areas. This will ensure smooth agriculture production.

References

[1] Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I. and Nawaz, Z. (2018). Rainfall Prediction using Data Mining Techniques: A Systematic Literature Review. International Journal of Advanced Computer Science and Applications, Vol. 9, No. 5, 143-150.

[2] Amin, M. G. M.; Ali, M. H. and Islam, A. K. M. R. (2004). Agro-climatic analysis for crop planning in Bangladesh, Bangladesh J. Agri. Eng., 15(1 &2):1-40.

[3] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees, Chapman & Hall/CRC Press, Boca Raton, FL.

[4] Breiman, L. (2001). Random Forests. Machine Learning. 45 (1): 5–32.

[5] Brownlee, J. (2016). Classification and Regression Trees for Machine Learning. Machine Learning Mastery (https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/)

[6] Boser, B. E., Guyon, I. M., Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. p. 144.

[7] Cohen et. al. (2003). Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences 3rd ed. Taylor & Francis Group.

[8] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. **7** (2): 179–188.

[9] Fix, E. and Joseph, H. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (Report). USAF School of Aviation Medicine, Randolph Field, Texas.

[10] Ho, T. K. (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[11] Kleinberg, E. (1990). Stochastic Discrimination (PDF). Annals of Mathematics and Artificial Intelligence. 1 (1–4): 207-239.

[12] Kleinberg, E. (1996). An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition. Annals of Statistics. 24 (6): 2319-2349.

[13] Kleinberg, E. (2000). On the Algorithmic Implementation of Stochastic Discrimination. IEEE Transactions on PAMI, 22 (5): 473-490.

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

[15] Mishra, N., Soni, H. K., Sharma, S. and Upadhyay, A. K. (2018). Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data, Int. J. Intell. Syst. Appl., vol. 10, no. 1, pp. 16–23.

[16] Monira, S. S., Faisal, Z. M. and Hirose, H. (2010). Comparison of artificially intelligent methods in short term rainfall forecast. Proc. 2010 13th Int. Conf. Comput. Inf. Technol. ICCIT 2010, no. Iccit, pp. 39–44, 2010.

[17] Olaiya, F. and Adeyemo, A. B. (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. I.J. Information Engineering and Electronic Business, 2012, 1, 51-59.

[18] Paramanik, M. A. H. (1991). Natural Diseaster. Article prepared for Bangladesh Space Research and Remote Sensing Organization (SPARRSO), Dhaka, Bangladesh.

[19] Pingale M. S., Khare D., Jat K. M., and Adamowaki J. (2016). Trend analysis of climatic variables in an arid and semi-arid region of the Ajmer District, Rajstan, India. Journal of Water and Land Development, DOI: 10.1515/jwld-2016-0001, No. 28 (I–III): 3–18.

[20] Rahman, M. M., Khalek, M. A. and Rahman, M. S. (2021). Performance of Different data mining methods for predicting rainfall of Rajshahi district, Bangladesh. Book chapter. Data Science and SDGs, Challenges, opportunities and Realities. 67-78.

[21] Rahman, M.M., and Rahman, M.S. (2022). Predicting rainfall based on machine learning algorithm: An evidence from Bogura district, Bangladesh. Int. J. Adv. Res. 10(08), 850-858 .

[22] Ramana, R. V., Krishna, B. Kumar, S. R. and Pandey, N. G. (2013). Monthly Rainfall Prediction Using Wavelet Neural Network Analysis, Water Resour. Manag, vol. 27, no. 10, pp. 3697–3711.

[23] Sivapragasam, C., Liong, S. and Pasha, M.(2001). Rainfall and runoff forecasting with SSA-SVM approach, J. Hydroinformatics, no. April 2016, pp. 141–152.

[24] Sethi, N. and Garg, D. K. (2014). Exploiting Data Mining Technique for Rainfall Prediction. International Journal of Computer Science and Information Technologies, Vol. 5 (3), 3982-3984.

[25] Solanki, N. and G. P. B, (2018). A Novel Machine Learning Based Approach for Rainfall Prediction. Inf. Commun. Technol. Intell. Syst. (ICTIS 2017) - Vol. 1, vol. 83, no. Ictis 2017.

[26] Talib, M. R., Ullah, T., Sarwar, M. U., Hanif, M. K. and Ayub, N. (2017). Application of Data Mining Techniques in Weather Data Analysis. International Journal of Computer Science and Network Security, Vol. 17, No. 6, 22-28.

[27] Tharun, V. P., Prakash, P. and Devi, S. R. (2018). Prediction of Rainfall Using Data Mining Techniques. Proceeding of the 2 nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Eplore Compliant-Part Number: CFP18BAC-ART: ISBN: 978-1-5386-1974-2.

[28] Thomas, M. and Peter, H. E. (1967). Nearest neighbor pattern classification (PDF). IEEE Transactions on Information Theory. 13 (1): 21–27.

[29] Vapnik, V. (1995). Support-vector networks (PDF). Machine Learning. 20 (3): 273–297.

[30] Wu, J., Long, J. and Liu, M. (2015). Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm. Neurocomputing, vol. 148, pp. 136–142.

[31] Zainudin, S., Jasim, D. S. and Bakar, A. A. (2016). Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction, Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1148–1153.