American Journal of Data Mining and Knowledge Discovery
2025; 10(1): 1-19
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20251001.11
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

Analysis of Climatic Factors and Utilization of Machine Learning Techniques to Anticipate Humidity Levels in Northern Bangladesh

Most. Rubina Akter, Md. Habibur Rahman*

Department of Statistics and Data Science, Jahangirnagar University, Dhaka, Bangladesh

Email address:
rubinaakter260@gmail.com (Most. Rubina Akter), habib.drj@juniv.edu (Md. Habibur Rahman)
*Corresponding author

To cite this article:
Most. Rubina Akter, Md. Habibur Rahman. (2025). Analysis of Climatic Factors and Utilization of Machine Learning Techniques to Anticipate Humidity Levels in Northern Bangladesh. *American Journal of Data Mining and Knowledge Discovery, 10*(1), 1-19. https://doi.org/10.11648/j.ajdmkd.20251001.11

Received: 30 January 2025; **Accepted:** 3 March 2025; **Published:** 5 March 2025

Abstract: Analyzing meteorological data in the northern region of Bangladesh is crucial for understanding various aspects influenced by humidity. This study employs machine learning algorithms, including k-nearest neighbor, Classification and Regression Trees, C5.0, Naive Bayes, Random Forest, and Support Vector Machine, to forecast the humidity of northern Bangladesh. Data from 1981 to 2020 from two meteorological stations, Rangpur and Dinajpur, were utilized. Results indicate that Rangpur had the highest average daily humidity (80.34%), while Dinajpur had the lowest (77.26%). Cloud amount correlates positively with humidity and inversely with temperature. The k-nearest neighbor, random forest, and support vector machine algorithms generally revealed better prediction performance than other algorithms. All things considered, the Random Forest model demonstrates superior performance on the testing dataset at both stations, achieving 70% accuracy, F_1-score (75.85%), and a kappa value of approximately 53.3% at Rangpur Station, and 74% accuracy, F_1-score (78.4%), and a kappa value of approximately 60% at Dinajpur Station. Subsequently, this study analyzes the best performance and accuracy of the random forest classification algorithms through k-fold cross-validation for predicting humidity. With this piece of information, it is anticipated that the study underscores the importance of random forest in predicting humidity and aiding decision-makers in water demand management, ecological balance, and health quality in the northern region of Bangladesh.

Keywords: Machine Learning, Cross Validation, Classification, Climate, Humidity, Bangladesh

1. Introduction

The climate of Bangladesh is classified as tropical monsoon, which means that it experiences significant changes in precipitation throughout the year, along with high temperatures and high humidity levels [1]. Bangladesh is the most vulnerable country to climate change, which causes river erosion, floods, flash floods, and intrusion of salinity into the land. Climate affects water resource management, crop management, dam operations and hydroelectricity generation, industrial site location, defense planning, tourism and transport, air pollution studies, and nearly all human activities. Humidity of the air is one of the most critical variables in meteorology since it significantly impacts the weather. In the atmosphere, humidity refers to the quantity of water vapor, significantly impacting the climate and the weather. The primary sources of water vapor in the lower atmosphere are evaporation from the Earth's surface and plant transpiration [2]. The hydrological cycle intrinsically relies on water vapor transportation across the atmosphere. Humidity (atmospheric moisture) plays a crucial role in the environment as it impacts the growth of plants, human health, and pollution levels. The growth rate, composition, and morphology of a plant are also affected by it [3]. The investigation of the varied configurations of indigenous flora that have emerged in distinct global regions, alongside the species that undergo growth and maturation during varying periods, elucidates the impact of moisture levels on plant life. There exists a strong correlation between atmospheric moisture levels and specific vegetation patterns [4]. Low air humidity can cause severe plant water loss. Controlling water loss by stomatal closure reduces carbon dioxide (CO_2) diffusion, limiting growth. Conversely, excessive humidity promotes the growth of conditions favorable to the spread of disease [5]. Humidity affects both human beings and animals. Both deficient and high levels of relative humidity can lead to physical discomfort. Epidemiological research suggests that humidity and humidification technology may indirectly influence the prevalence of allergic reactions and infectious respiratory disorders. Humidity also has an impact on electronic equipment, as well as chemical and refinery operations that utilize furnaces [6]. The economy of Bangladesh is primarily based on agriculture. The climate exerts a substantial influence on our nation's agricultural operations. Nonetheless, the process of cultivation is often impeded by anomalous atmospheric phenomena [7]. The humidity level can affect crops through the processes of evaporation, transpiration, and condensation. Accurate forecasting of atmospheric factors is crucial for various applications, including climate monitoring, drought detection, severe weather prediction, agriculture and production, energy and industrial planning, communication, and pollution distribution [1]. The agricultural northwest region of Bangladesh relies heavily on precipitation forecasts, making assumptions-free models like CART crucial for accurate prediction, yielding approximately eighty percent accuracy in precipitation labeling for policymaking and local communities [8]. Rahman et. al examined precipitation patterns in seventeen locations across Bangladesh by utilizing unsupervised machine learning methods, including cluster analysis and multidimensional scaling, and implementing four hierarchical clustering techniques and diverse dissimilarity measures [9]. The study explores the uniformity in Bangladesh regarding atmospheric temperature parameters through unsupervised machine learning techniques, including kmeans clustering and ward linkage clustering, examining four temperature variables across thirtyfour locations [10]. Because of the changing nature of the atmosphere, it is impossible to estimate humidity accurately. It is critical to understand the nature of fluctuations in humidity. The main goal of this paper is to make predictions about the humidity, which is typically used to refer humidity in the northern part of Bangladesh. For this purpose, daily humidity data from two meteorological stations in Bangladesh has been studied based on available data from the past 40 years (1981-2020). The Bangladesh Meteorological Department (BMD) provided all the necessary secondary data. Machine learning (ML) approaches are used to find the best algorithm. Cross-validation is a statistical technique that is used to choose a model that can better predict test errors for predictive models. In this research, the k-fold cross-validation technique is applied to check the accuracy of the recommended algorithm at the selected stations, and the best model is then used to forecast the humidity at the selected stations. Hopefully, this study will help determine the best ML algorithm for forecasting humidity and understanding Bangladesh's future climate. The main objectives of this research work are to reveal the nature of atmospheric data; and to assess the effect of daily total rainfall (RAN), daily mean sea level pressure (SLP), daily average cloud amount (CLA), dry bulb temperature (DBT), daily maximum temperature (MXT), and daily minimum temperature (MNT) on daily humidity (HUM) in the study areas; to fit the machine learning (ML) algorithm for forecasting the humidity of various locations; to evaluate the goodness of the algorithms; to determine which algorithm predicts humidity the most accurately and find the best among the algorithms to predict the humidity; to make a comparative study among the ML algorithms, and to check the accuracy of the fitted algorithms. The research was carried out utilizing secondary data. The information mentioned earlier has been acquired from the BMD. The current investigation collected information on various meteorological parameters, including daily total rainfall (RAN), mean sea level pressure (SLP), average cloud amount (CLA), dry bulb temperature (DBT), maximum temperature (MXT), minimum temperature (MNT), and daily humidity (HUM) over four decades spanning from 1981 to 2020. For this investigation, two stations in the northern region of Bangladesh, namely Rangpur and Dinajpur, were chosen.

2. Methods and Methodology

Machine learning (ML) algorithms are effective when applied to problems whose solutions require knowledge that is difficult to describe. The meteorological parameter 'humidity' characteristics are complex and nonlinear, making it a good candidate for this type of model [11]. This study successfully explored various ML classification algorithms to forecast humidity in the northern region of Bangladesh, including kNN, CART, C5.0, NB, RF, and SVM. The results of this study could improve our understanding of humidity patterns in this region and inform future research in this area. The methodology of this research employed entails a series of steps, including data pre-processing, partitioning the dataset into training and test subsets, implementing ML algorithms on the training subset, assessing the performance of these algorithms on the test subset, and ultimately the most effective algorithm is utilized to forecast humidity based on the entire dataset. The evaluation of performances was based on five distinct performance metrics, namely accuracy, sensitivity, specificity, precision, and F_1-score, which were derived from the confusion matrix. Here, these techniques are succinctly explained.

2.1. k-Nearest Neighbor (kNN)

The kNN algorithm is a widely recognized non-parametric methodology and a variant of lazy learning or instance-based learning algorithms. This algorithm is characterized by its directness, yet its efficacy, lucidity, user-friendliness, and competitiveness can address classification and regression problems of a relatively modest scale. It is postulated that the complete training dataset, encompassing the intended categorization for each item and the corresponding data within the set, will be utilized to construct the model. Before classification, both the training set items and the novel item must be computed [12]. The training set only considers the k-closest items. The class with the most members from the k nearest objects receives the new item. The majority class is chosen from the k nearest neighbors using the Euclidean distance method 1 for kNN classification. The equation for Euclidean distance is expressed as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (1)$$

In this formula, d represents the Euclidean distance, x and y represent data points with N dimensions, and i represents an index number.

2.2. Classification and Regression Trees

The recursive algorithm known as Classification and Regression Trees (CART) was developed in data mining by [13]. Training data sets allow CART to learn any mapping function. CART can reuse factors across the model to find complex interdependencies between many variables. However, it requires more training data [14]. The CART methodology includes three steps: (i) Maximum tree construction, (ii) Selection of the optimal tree size, and (iii) Data classification or creation using a constructed tree. CART always selects the feature with the minimum Gini information gain in the current data set as the node in the decision tree. Using the Gini index, the sample sets to be classified are split into two sub-sample sets and looped until they are leaf nodes or a precondition for ending the categorization is met [15]. The Gini index of the probability distribution can be defined as: $Gini(p) = \sum_{k=1}^{k} p_k(1-p_k) = 1 - \sum_{k=1}^{k} p_k^2$ where $p_k = \frac{C_k}{D}$. The sample set is D, and C_k is a sample subset of class k. The Gini index formula is then provided as: $Gini(D) = 1 - \sum_{k=1}^{k}\left(\frac{|C_k|}{|D|}\right)^2$ Assuming feature A is the criterion, sample set D is split into the subsets D_1 and D_2, and the Gini index of sample set D is as follows: $Gini(D,A) = \frac{|D_1}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$. This research utilizes the CART algorithm for classification to forecast the humidity of the northern region of Bangladesh.

2.3. C5.0

The C5.0 algorithm is a classification technique in data mining suitable for large datasets and integrates the decision tree approach. The C5.0 algorithm has superseded the ID3 and C4.5 algorithms developed by Ross Quinlan [16]. This algorithm selects characteristics by making use of the most available information. The test attribute for each node in a decision tree is chosen using the information gain size. The following node will inherit the qualities with the most information from its parents. Concerning effectiveness, memory, and speed, it is superior to C4.5 [17].

2.4. Naive Bayes

Bayesian classification is a supervised learning approach that is a statistical classification method. The NB algorithm is founded on the Bayesian theorem. The utilization of it can address issues about both diagnosis and prognostication. The systematic expression of uncertainty about the model is facilitated by calculating outcome probabilities [18]. Given the assumption of independent variables, only the contentions of the variables for each class, not their overall distribution, must be determined. The Naive Bayes classifier has the advantage of requiring less training data to compute the classification-related parameters. It may be used to classify problems into many different binary categories [19].

2.5. Random Forest

Random Forest (RF) is an algorithm created by Dr. [20] that has proven to be a robust general-purpose classification and regression tool. It is a supervised classification method that groups records into categories by constructing multiple classifiers. This algorithm is based on statistical learning theory and uses the Bootstrap randomized re-sampling technique to derive multiple sample sets from the initial training datasets [21]. After constructing a decision tree model for each individual sample set, this algorithm combines all the results obtained from the decision trees to make a prediction regarding the categorization based on the previously determined voting mechanism [22].

2.6. Support Vector Machine

Support Vector Machines (SVM) were initially proposed by Vapnik in 1995 to achieve ”distribution-free learning from data” through statistical learning theory. The support vector machine is a practical and relatively recent technique utilized for learning separate functions in pattern recognition tasks and conducting function estimates in regression issues [23]. SVM is a powerful tool that can accurately classify data instances using a linear separation hyperplane. SVMs are versatile and successfully applied in various applications such as classification, regression, and clustering. SVMs are an excellent choice for multiple applications because they effectively handle overfitting challenges that may arise in high-dimensional spaces through global optimization. The ”kernel technique” enables the transformation of the initial feature space into a higher-dimensional feature space, thereby enhancing the classification abilities of traditional SVMs [24].

2.7. Performance Evaluation Criteria and Measures for Classification

Assessing an ML model’s performance is pivotal in creating an effective ML algorithm. Performance or evaluation metrics are utilized to assess the quality or performance of a model through a range of measurements. All measures are evaluated using a confusion matrix. Table 1 displays a typical confusion matrix for a binary classifier.

Table 1. *A Typical Confusion Matrix for a Binary Classifier to Measure Performance.*

	Class	**Actual Class**		**Measures**
		Positive	**Negative**	
Predicted Class	Positive	True Positive (TP)	False Positive (FP)	PPV
	Negative	False Negative (FN)	True Negative (TN)	NPV
Measures		Sensitivity	Specificity	Accuracy

Several performance measures that are commonly derived from the confusion matrix include: Accuracy is a measurement that determines how frequently the classification makes accurate predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + PN} \tag{2}$$

A model's sensitivity (the true positive rate) measures how many positive cases were correctly identified.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

A model's specificity (also known as the true negative rate) is a measure of how many negative cases were correctly identified.

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

The metric evaluates the accuracy of predicting the positive class.

$$\textit{Precision or Positive Predictive Value (PPV)} = \frac{TP}{TP + FP} \tag{5}$$

The recall is the proportion of correctly identified positive cases. It is also known as sensitivity.

$$\textit{Recall or Negative Predictive Value (NPV)} = \frac{TP}{TP + FN} \tag{6}$$

F_1-Score is the weighted average of Precision and Recall.

$$F_1 - Score = 2 \times \left(\frac{Recall \times Precision}{Recall + Precision}\right) \tag{7}$$

2.8. Validation with k-fold Cross-Validation

The k-fold cross-validation test assesses the process performance of an algorithm by randomly dividing the sample data and grouping it by as much as the value of k [25]. In k-fold cross-validation techniques, all training data sets are utilized for the training and validation processes. This research validates the prediction model using the values 2, 3, 5, 7, 10, and 11.

3. Data Description

The present investigation utilizes secondary data from the BMD of the Rangpur and Dinajpur stations. The BMD data comprises daily measurements of rainfall (RAN), humidity (HUM), mean sea level pressure (SLP), average cloud amount (CLA), dry bulb temperature (DBT), maximum temperature (MXT), and minimum temperature (MNT) for the two specified locations. The meteorological stations in Bangladesh that are used to designate the research regions are depicted with accuracy in Figure 1. Figure 1 accurately presents the meteorological stations in Bangladesh that helped to identify the study areas. This study has chosen the BMD meteorological stations in Rangpur and Dinajpur based on a comprehensive dataset spanning a maximum of 40 years, from 1981 to 2020. In this study, Table 2 presents a comprehensive list of the various climate variables considered for each month of the year, accompanied by a statistical summary.

Table 2. *The list of the corresponding variable names for the Rangpur and Dinajpur stations has been considered in this study.*

Serial No.	**Name**	**Short Name**
01	Daily Humidity for Rangpur Area	RHUM
02	Daily Total Rainfall for Rangpur Area	RRAN
03	Daily Mean Sea Level Pressure for Rangpur Area	RSLP
04	Daily Cloud Average Amount for Rangpur Area	RCLA
05	Daily Dry Bulb Temperature for Rangpur Area	RDBT
06	Daily Maximum Temperature for Rangpur Area	RMXT
07	Daily Minimum Temperature for Rangpur Area	RMNT
08	Daily Humidity for Dinajpur Area	DHUM
09	Daily Total Rainfall for Dinajpur Area	DRAN
10	Daily Mean Sea Level Pressure for Dinajpur Area	DSLP
11	Daily Cloud Average Amount for Dinajpur Area	DCLA
12	Daily Dry Bulb Temperature for Dinajpur Area	DDBT
13	Daily Maximum Temperature for Dinajpur Area	DMXT
14	Daily Minimum Temperature for Dinajpur Area	DMNT

3.1. Data Processing and Partition of Data

To obtain results that are suitable for use, it is imperative to cleanse the data thoroughly. The present dataset comprises a significant number of invalid or absent values, including the value ”null.” In this context, solely the accurate values have been utilized after filtering. The dataset has been partitioned into two subsets, with 75% of the data allocated for training purposes and the remaining 25% reserved for testing. Seven variables are being measured at two different meteorological stations. The target variable among these variables is humidity. To facilitate comprehension, this study has categorized humidity into three different classifications:

1. Low Humidity [H1 = 0-76]; This study considered this range as low humidity, which is favorable for the climate of Bangladesh.
2. Medium Humidity [H2 = 77-84]; Humidity is roughly average. The weather conditions are conducive and adhesive.
3. High Humidity [H3: 85-100]; Lots of moisture in the air, uncomfortable to sustain a healthy lifestyle.

Regarding the climate conditions prevalent in Bangladesh, it is recommended that the humidity levels for forecasting purposes be maintained at a level lower than 85%. This range may be suitable for reaching the highest level of comfort and well-being in Bangladesh, which has a tropical climate.

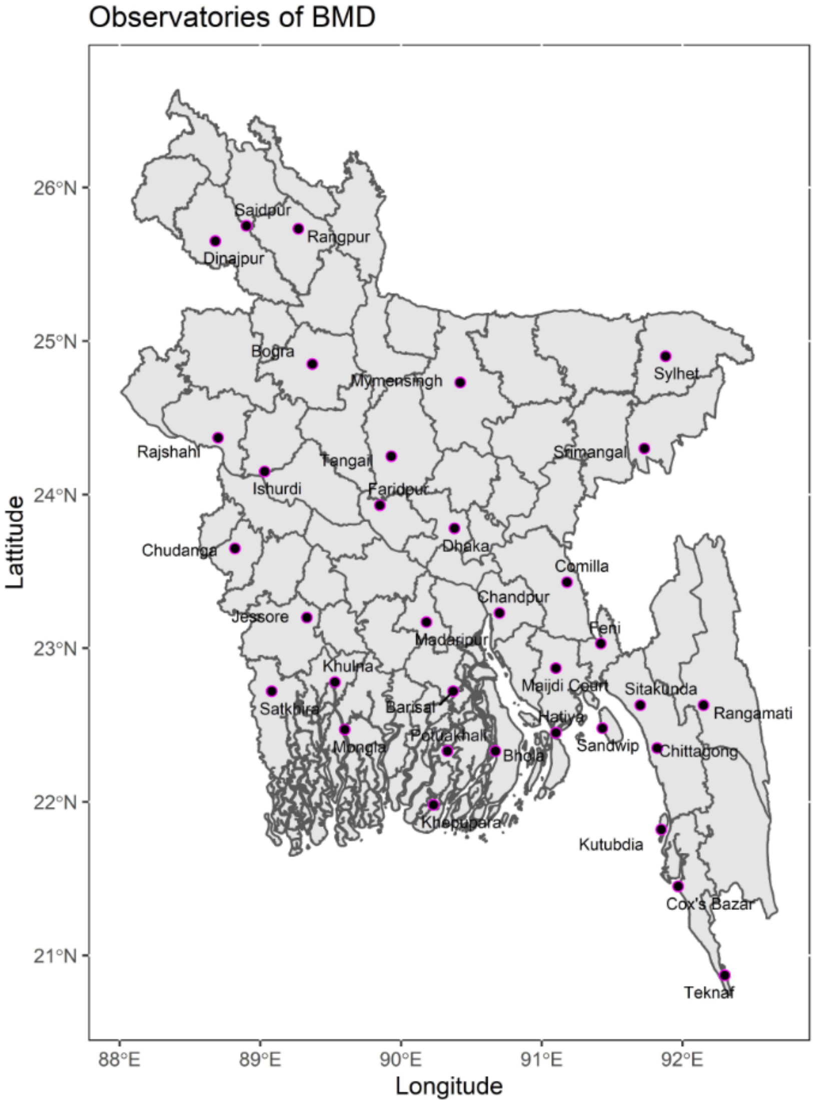

Figure 1. The schematic plot for the meteorological stations in Bangladesh.

3.2. Descriptive Analysis

Descriptive statistics serve as the fundamental basis for the concise representation of data. Summary statistics provide details about a given dataset and sample data summary. The concepts mentioned earlier encompassed: mean, median, mode, standard error, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, count, and confidence level. Table 3 presents the recorded humidity and other meteorological variables data for the Rangpur and Dinajpur stations from 1981 to 2020.

Table 3. *Summary Statistics for different Meteorological Variables from January to December for (a) Rangpur and (b) Dinajpur.*

(a) Rangpur

Name	**RHUM**	**RRAN**	**RSLP**	**RCLA**	**RDBT**	**RMXT**	**MNT**
Mean	80.342	6.179	1008.307	3.627	24.539	29.659	20.235
Median	81	0	1008.3	3.6	25.8	30.4	21.6
Mode	82	0	1012	0	28.3	31	25
Std Error	0.072	0.159	0.048	0.024	0.039	0.033	0.048
Std Deviation	8.562	18.955	5.744	2.812	4.681	3.888	5.681
Variance	73.304	359.277	32.997	7.907	21.916	15.118	32.277
Kurtosis	2.321	49.827	-1.056	-1.462	-0.662	0.462	-1.025
Skewness	-0.977	5.902	-0.091	0.085	-0.615	-0.718	-0.486
Range	89	294	28.9	8	25.4	28.9	27
Minimum	10	0	992.8	0	7.9	10.9	3.5
Maximum	99	294	1021.7	8	33.3	39.8	30.5
Count	14299	14299	14299	14299	14299	14299	14299
Correlation	RHUM	RRAN	RSLP	RCLA	RDBT	RMXT	RMNT
RHUM	1.00	0.35	-0.20	0.56	0.06	-0.23	0.28
RRAN	0.35	1.00	-0.26	0.39	0.14	0.00	0.24
RSLP	-0.20	-0.26	1.00	-0.66	-0.82	-0.67	-0.84
RCLA	0.56	0.39	-0.66	1.00	0.53	0.24	0.68
RDBT	0.06	0.14	-0.82	0.53	1.00	0.91	0.95
RMXT	-0.23	0.00	-0.67	0.24	0.91	1.00	0.77
RMNT	0.28	0.24	-0.84	0.68	0.95	0.77	1.00

(b) Dinajpur

Name	**DHUM**	**DRAN**	**DSLP**	**DCLA**	**DDBT**	**DMXT**	**DMNT**
Mean	77.265	5.375	149.773	3.257	24.88	30.056	19.986
Median	79	0	0	3	26.3	30.8	21.5
Mode	81	0	0	0	28.8	32	26
Std Error	0.092	0.151	2.985	0.023	0.041	0.034	0.05
Std Deviation	11.081	18.154	358.373	2.72	4.883	4.13	5.948
Variance	122.796	329.579	128431.415	7.397	23.844	17.057	35.379
Kurtosis	2.606	109.368	1.902	-1.368	-0.554	0.649	-1.099
Skewness	-1.285	7.959	1.975	0.244	-0.668	-0.725	-0.466
Range	84	508	1020.4	8	26.8	30.4	26.6
Minimum	16	0	0	0	8.5	11	3.2
Maximum	100	508	1020.4	8	35.3	41.4	29.8
Count	14411	14411	14411	14411	14411	14411	14411
Correlation	DHUM	DRAN	DSLP	DCLA	DDBT	DMXT	DMNT
DHUM	1.00	0.29	-0.27	0.55	0.01	-0.23	0.30
DRAN	0.29	1.00	0.00	0.39	0.12	0.00	0.23
DSLP	-0.27	0.00	1.00	0.00	0.06	-0.03	-0.06
DCLA	0.55	0.39	0.00	1.00	0.49	0.20	0.66
DDBT	0.01	0.12	0.06	0.49	1.00	0.91	0.93
DMXT	-0.23	0.00	-0.03	0.20	0.91	1.00	0.76
DMNT	0.30	0.23	-0.06	0.66	0.93	0.76	1.00

Here, there are 14,299 values for Rangpur station and 14,411 values for Dinajpur station. According to the data in the table, the average relative humidity at the Rangpur station is 80.342, the average rainfall is 6.179, the average mean sea level pressure is 1008.307, the average cloud amount is 3.627, the average dry bulb temperature is 24.539, the average minimum temperature is 20.235, and the average maximum temperature is 29.659. The means of RRAN, RSLP, and RCLA are higher than the median. So the distribution of these variables is positively skewed, while the distribution of other variables is negatively skewed. For kurtosis, the values of RHUM, RRAN, and RMXT are all greater than 1, which means that the distribution of these variables is too skewed. The average relative humidity at the Dinajpur station is 77.265, the average rainfall is 5.375, the average mean sea level pressure is 149.773, the average cloud amount is 3.257, the average dry bulb temperature is 24.88, the average minimum temperature is 19.986, and the average maximum temperature is 30.056.

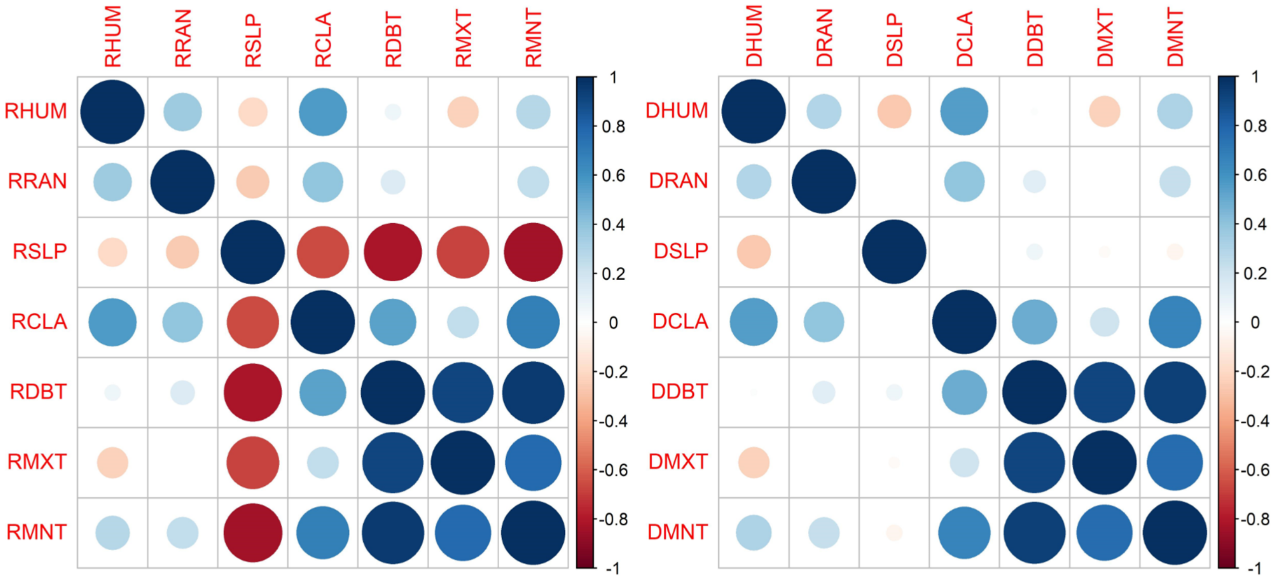

Figure 2. *Schematic plot of the correlation matrix for (a) Rangpur at Left and (b) Dinajpur at Right.*

The means of DRAN, DSLP, and DCLA are higher than the median. So the distribution of these variables is positively skewed, while the distribution of other variables is negatively skewed. For kurtosis, the values of DHUM, DRAN, DSLP, and DMXT are all greater than 1, which means that the distribution of these variables is too skewed. A correlation matrix is a tabular representation of correlation coefficients among a set of variables, which serves to ascertain the presence or absence of any interdependence between the variables. The coefficient serves as an indicator of the magnitude and orientation of the correlation, whether it is positive or negative. The value ranges from -1 to 1 and is computed using the following formula: (i) A value of -1 denotes a totally negative linear correlation between two variables, (ii) 0 means that there is no linear association between two variables, (iii) A correlation coefficient of 1 denotes a positive linear association between two variables. The Table 3 also represents the relationship between two distinct variables. For Rangpur station, the correlation between RHUM and RCLA is fifty-six percent, which indicates that they are moderately positively correlated, and the correlation between RHUM and RSLP is minus twenty, and RHUM and RMXT is minus twenty-three, which indicates that they are weakly negatively and negligibly correlated (Table 3). For Dinajpur station, the correlation between DHUM and DCLA is fifty-five percent, which indicates that they are moderately positively correlated, and the correlation between DHUM and DSLP is minus twenty-seven, and DHUM and DMXT is minus twenty-three, which indicates that they are weakly negatively and negligibly correlated (Table 3). The plot of the correlation matrix between these variables is depicted below in Figure 2, and Table 3.

4. Results and Discussions

The objective of this research is to utilize various ML classification algorithms to accurately forecast the humidity levels in the Rangpur and Dinajpur regions. The machine learning approaches are used to appraise the climate parameters of cloud and rainfall in different studies [26, 27]. There are 14299 records in the Rangpur dataset and 14411 records in the Dinajpur dataset. For all the algorithms, 75% of the data is used as the training set and 25% as the test or validation set, respectively.

4.1. Fitted Models to Forecast the Humidity of the Rangpur Area

4.1.1. kNN Algorithm to Forecast the Humidity of the Rangpur

The performance of the kNN algorithm is evaluated with different values of the number of neighbors for the parameter k. The present study selected 35 as the optimal value of k due to their minimal error rate for the Rangpur station. Figure 3 displays the results of the suggested approaches to the dataset with k=35 values. The outcomes of the various algorithms were demonstrated through the utilization of confusion matrices for classification purposes. The dataset was classified into three distinct categories based on the level of humidity, namely: H1 for low humidity, H2 for medium humidity, and H3 for high humidity. Table 4, and Table 5 present the confusion matrices that display the outcomes of all ML classification algorithms. On the training dataset, the kNN algorithm provided 80% of sensitivity for class H1, 60.02% of sensitivity for class H2, and 74.62% of sensitivity for class H3. The specificity was 77.89% for class H1, 62.98% for class H2, and 76.24% for class H3. On the test dataset, the sensitivity was 82.87% for class H1, 79.49% for class H2, and 87.38% for class H3. The specificity was 85.38% for class H1, 78.96% for class H2, and 87.99% for class H3. Table 5 presents that the kappa coefficient values for the training and test datasets are 0.4858 and 0.5196, respectively, indicating moderate agreement. It achieves an accuracy of 67% for the training dataset and 69% for the test dataset. The test dataset exhibits the highest F_1-Score of 74.83% for the highest category of humidity (H3), indicating its superior performance. As a result, humidity forecasting accuracy on the test dataset is quite good.

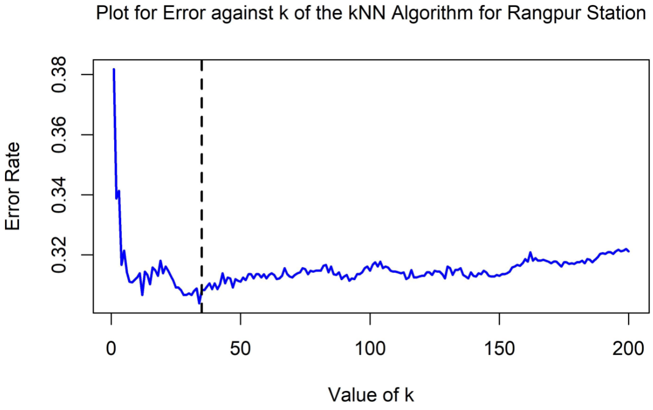

Figure 3. *Optimum value of the kNN Model at Rangpur Station.*

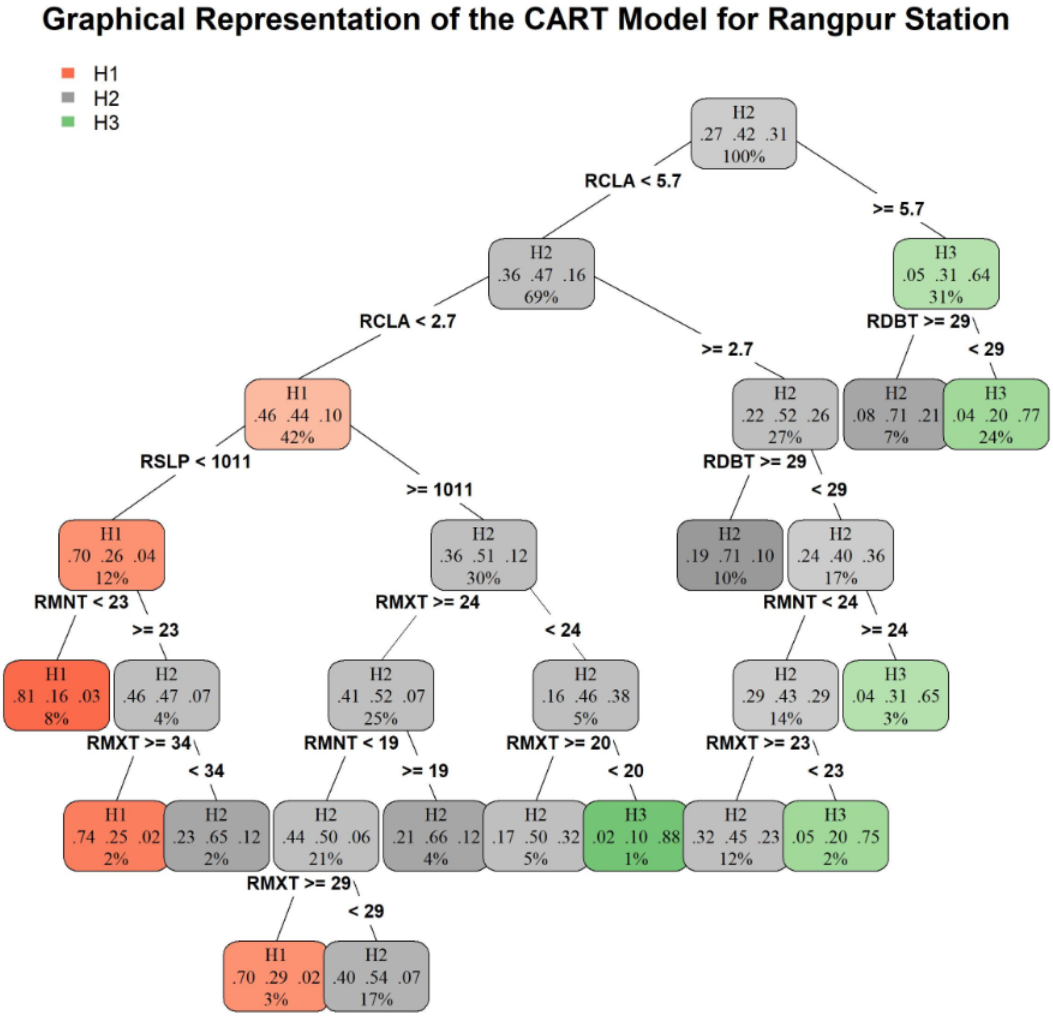

Figure 4. *A graphical representation of the pruned tree for the CART model using the best CP for forecasting the humidity at Rangpur Station.*

4.1.2. CART to Forecast the Humidity of the Rangpur

To conduct CART analyses, prune the data and plot the tree model using the CP value with the lowest error. The best complexity parameter (CP) for each combination is determined through trial and error in order to forecast the humidity using the CART approach. The complexity parameter (CP) is used to control the size of the decision tree and figure out the best size for the tree.

The higher the CP, the smaller the tree. Overfitting is caused by a CP value that is too low, and a tree that is too tiny is caused by a CP value that is too high. Both cases decrease the predictive performance of the model. It is revealed that the best CP is 0.007 for the CART model to predict the humidity of Rangpur, which contains the lowest error. The best CP that describes the majority of the data is used to build the final CART model. The constructed CART model utilizes a set of five variables: RCLA, RDBT, RMNT, RMXT, and RSLP. The resultant CART model is shown in Figure 4.

On the training dataset, the CART model provided 76.96% of sensitivity for class H1, 58.07% of sensitivity for class H2, and 75.50% of sensitivity for class H3. The specificity was 81.0% for class H1, 79.21% for class H2, and 87.22% for class H3. On the test dataset, the sensitivity was 71.07% for class H1, 57.63% for class H2, and 74.04% for class H3. The specificity was 81.81% for class H1, 76.38% for class H2, and 85.92% for class H3. Table 5 demonstrates that the kappa coefficient values for the training and test datasets are 0.45 and 0.43, respectively, indicating moderate agreement. This model attains an accuracy of 66% for the training dataset and 64% for the test dataset. The training dataset exhibits the highest F_1-Score of 73.17% for the highest category of humidity (H3), indicating its superior performance. As a result, humidity forecasting accuracy on the training dataset is quite good.

4.1.3. C5.0 Model to Forecast the Humidity of the Rangpur

In this visual representation of the c5.0 method, the decision tree with its five leaf nodes (classifiers) is presented in Figure 5. First, it splits node 1 into nodes 2 and 7 based on the value of RCLA and node seven into nodes 8 and 9 based on the value of RMXT. Then, it splits node two into node six based on the value of RCLA, and node 3 splits into nodes 4 and 5 based on the value of RSLP. Observations in node 4 are labeled as class H1, observations in node 5 as class H2, observations in node 6 as class H2, observations in node 8 as class H3, and observations in node 9 as class H2. The decision tree uses only three of the five variable measure attributes (RCLA, RSLP, and RMXT).

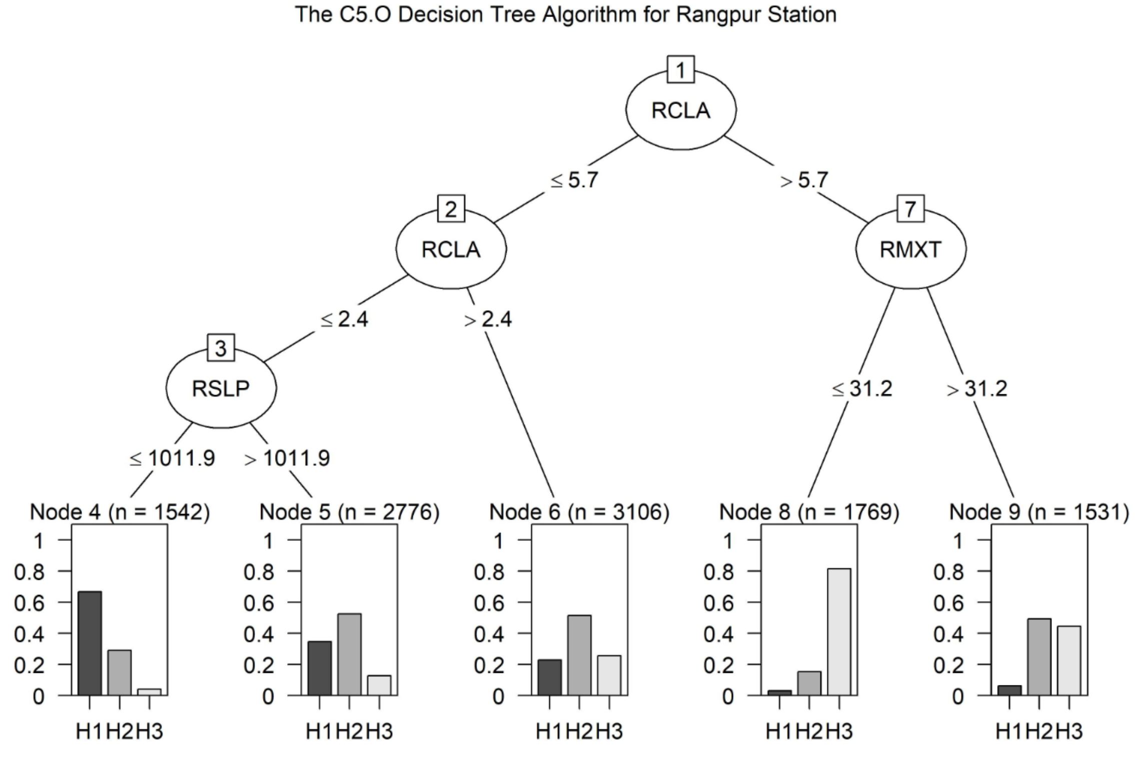

Figure 5. The decision tree of the C5.0 algorithm to forecast humidity for Rangpur station.

On the training dataset, the proposed algorithm provided 66.66% of sensitivity for class H1, 51.42% of sensitivity for class H2, and 81.46% of sensitivity for class H3. The specificity was 80.14% for class H1, 78.22% for class H2, and 78.81% for class H3. On the test dataset, the sensitivity was 62.88% for class H1, 52.73% for class H2, and 81.0% for class H3. The specificity was 81.56% for class H1, 76.97% for class H2, and 78.45% for class H3. Table 5 represents that the kappa coefficient values for the training and test datasets are 0.33 and 0.34, respectively, indicating fair agreement. This model attains an accuracy of 58% for the training dataset and 59% for the test dataset. The test dataset exhibits the highest F_1-Score of 64.41% for the medium category of humidity (H3), indicating its superior performance. As a result, humidity forecasting accuracy on the test dataset is quite good.

4.1.4. Naive Bayes to Forecast the Humidity of the Rangpur

The NB algorithm attained the maximum sensitivity, specificity, and F_1 - measure, which were 79% of H1 for the training dataset, 52.67% of H1 for the testing dataset, and 58% of H3 for the testing dataset, respectively. In the test dataset, the NB algorithm performed at a maximum accuracy of 53.4%.

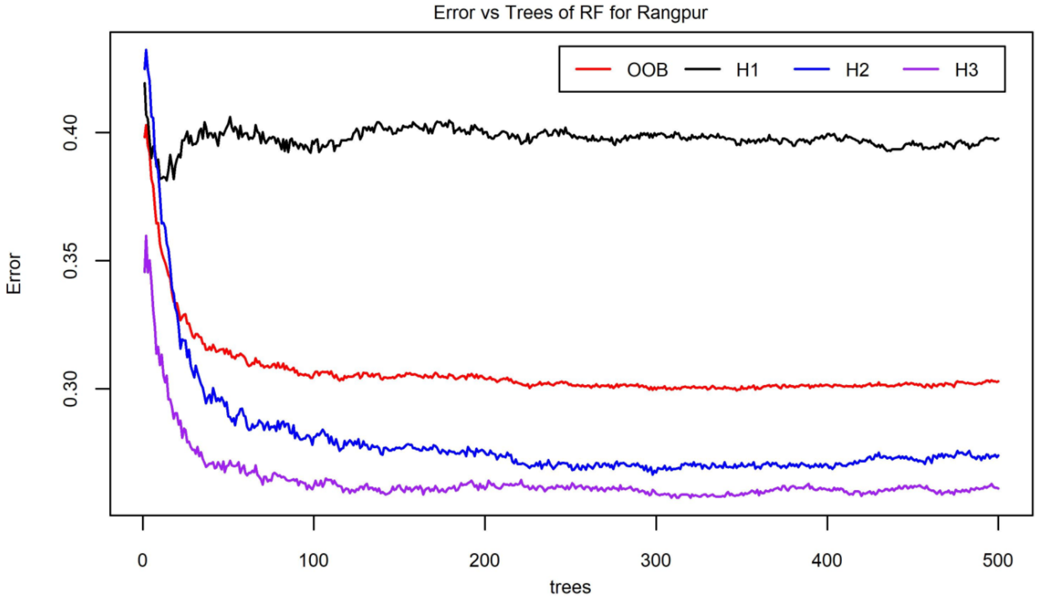

Figure 6. Error rate of RF algorithm to forecast humidity for Rangpur station.

Table 4. The confusion matrix of various ML algorithms for training data and test data at Rangpur station using meteorological data, 1981-2020.

Name	Predicted						
	Training Dataset				Test Dataset		
	Category	H1	H2	H3	H1	H2	H3
kNN	H1	1292	1440	120	472	406	28
	H2	295	3545	693	126	1186	230
	H3	28	921	2390	8	291	828
CART	H1	1082	1656	114	344	532	30
	H2	290	3588	655	120	1182	240
	H3	34	935	2370	20	337	770
C5.0	H1	1028	1768	56	349	539	18
	H2	449	3812	272	170	1276	96
	H3	65	1833	1441	36	605	486
NB	H1	2183	622	47	710	180	16
	H2	2150	2021	362	747	674	121
	H3	642	1194	1503	218	383	526
RF	H1	1718	1062	72	540	341	25
	H2	612	3291	630	202	1140	200
	H3	41	831	2467	20	281	826
SVM	H1	1533	1225	94	469	405	32
	H2	483	3446	604	176	1172	194
	H3	18	975	2346	6	347	774

4.1.5. Random Forest to Forecast the Humidity of the Rangpur

It is instructive to understand how the error in the random forest test varies with the number of trees before moving on to the primary results on forecasting humidity. Figure 6 shows the error rate and the number of trees. The model comprises of 500 trees, and two variables are evaluated at every split. Accordingly, the classification error of H1 is 39.76%, while that of H2 and H3 are 27.39% and 26.11%, respectively. Moreover, the out-of-bag error (OOB) of the RF algorithm is estimated to be 30.29%. The RMXT variable is the most significant variable, followed by the RMNT variable, then RCLA, RDBT, RSLP, and RRAN. The Gini reduction values are as follows: RRAN has a value of 489.3729, RSLP is 1070.9395, RCLA is 1308.2639, RDBT is 1219.0828, RMXT is 1451.9536, and RMNT is 1318.4302.

On the training dataset, the RF provided 72.46% of sensitivity for class H1, 63.48% of sensitivity for class H2, and 77.85% of sensitivity for class H3. The specificity was 70.87% for class H1, 64.70% for class H2, and 78.59% for class H3. On the test dataset, the sensitivity was 86.42% for class H1, 77.58% for class H2, and 88.46% for class H3. The specificity was 86.99% for class H1, 77.83% for class H2, and 88.07% for class H3. Table 5 demonstrates that the kappa coefficient values for the training and test datasets are 0.5303 and 0.5332, respectively, indicating moderate agreement. The proposed model achieves a classification accuracy of 69% for the training dataset and 71% for the test dataset. The test dataset exhibits the highest F_1-Score of 75.85% for the highest category of humidity (H3), indicating its superior performance. RF achieves the maximum classification accuracy for forecasting humidity at the Rangpur station and performs well for both datasets.

Table 5. *The performance of the different statistics for forecasting humidity at Rangpur Station.*

Name	Statistics	Training Dataset			Test Dataset		
		H1	H2	H3	H1	H2	H3
kNN	Sensitivity	0.8000	0.6002	0.7462	0.7789	0.6298	0.7624
	Specificity	0.8287	0.7949	0.8738	0.8538	0.7896	0.8799
	Precision	0.4530	0.7820	0.7158	0.5210	0.7691	0.7347
	F_1-Score	0.5785	0.6792	0.7307	0.6243	0.6926	0.7483
	Accuracy	0.6739			0.6954		
	Kappa	0.4858			0.5196		
CART	Sensitivity	0.7696	0.5807	0.7550	0.71074	0.5763	0.7404
	Specificity	0.8100	0.7921	0.8722	0.81818	0.7638	0.8592
	Precision	0.3794	0.7915	0.7098	0.37969	0.7665	0.6832
	F_1-Score	0.5082	0.6699	0.7317	0.49496	0.6579	0.7107
	Accuracy	0.6565			0.6422		
	Kappa	0.4551			0.4290		
C5.0	Sensitivity	0.66667	0.5142	0.8146	0.62883	0.5273	0.8100
	Specificity	0.66667	0.5142	0.8146	0.81556	0.7697	0.7845
	Precision	0.36045	0.8409	0.4316	0.38521	0.8275	0.4312
	F_1-Score	0.46791	0.6382	0.5642	0.47775	0.6441	0.5628
	Accuracy	0.5857			0.5905		
	Kappa	0.3298			0.3350		
NB	Sensitivity	0.4388	0.5267	0.7861	0.4239	0.5449	0.7934
	Specificity	0.8836	0.6353	0.7916	0.8968	0.6287	0.7936
	Precision	0.7654	0.4458	0.4501	0.7837	0.4371	0.4667
	F_1-Score	0.5578	0.4829	0.5725	0.5502	0.4851	0.5877
	Accuracy	0.5322			0.5343		
	Kappa	0.3016			0.3085		
RF	Sensitivity	0.7246	0.6348	0.7785	0.7087	0.6470	0.7859
	Specificity	0.8642	0.7758	0.8846	0.8699	0.7783	0.8807
	Precision	0.6024	0.7260	0.7388	0.5960	0.7393	0.7329
	F_1-Score	0.6579	0.6774	0.7581	0.6475	0.6901	0.7585
	Accuracy	0.6971			0.7010		
	Kappa	0.5303			0.5332		
SVM	Sensitivity	0.7537	0.6103	0.7707	0.7204	0.6091	0.7740
	Specificity	0.8482	0.7859	0.8707	0.8505	0.7759	0.8629
	Precision	0.5375	0.7602	0.7026	0.5177	0.7601	0.6868
	F_1-Score	0.6275	0.6771	0.7351	0.6024	0.6763	0.7278
	Accuracy	0.6831			0.6755		
	Kappa	0.5265			0.4878		

4.1.6. SVM to Forecast the Humidity of Rangpur

Table 6 summarizes the performance metrics, including sensitivity, specificity, precision, and F_1-Score. The SVM algorithm yields a maximum sensitivity score of 77% for H3 in both the training and test datasets. The specificity values that are at their highest for H3 are 87 and 88 percent, respectively. For the H3 classification, the maximum precision value (true positive rate) can be found in both datasets, which is 76%. Moreover, class H3 displays the highest F_1-Score of 73.51%, implying that it possesses the greatest ability to accurately classify observations. Table 6 shows that the highest kappa value is 53%, indicating that the agreement is considered moderate for the training dataset. In addition, SVM achieves a classification accuracy of 68.31% for the training dataset and 67.55% for the test dataset.

Table 6. *Comparison of the performance of ML algorithms in percent for forecasting humidity at Rangpur station. Here, NPV means Neg Pred Value, DER means Detection Rate, and BAC means Balanced Accuracy.*

Algorithm	k	Accuracy	Kappa	F_1 Score	Recall	Specificity	Precision	NPV	DER	BAC
kNN	2	0.684	0.506	0.679	0.669	0.830	0.706	0.836	0.228	0.749
	3	0.687	0.511	0.683	0.672	0.832	0.708	0.837	0.229	0.752
	5	0.689	0.514	0.685	0.674	0.833	0.710	0.838	0.230	0.753
	7	0.690	0.516	0.687	0.676	0.833	0.712	0.838	0.230	0.755
	10	0.691	0.517	0.687	0.676	0.834	0.715	0.839	0.230	0.755
	11	0.691	0.517	0.687	0.676	0.834	0.713	0.839	0.230	0.755
RF	2	0.697	0.530	0.696	0.688	0.839	0.710	0.841	0.232	0.763
	3	0.700	0.534	0.699	0.689	0.840	0.716	0.843	0.233	0.765
	5	0.702	0.536	0.701	0.691	0.840	0.718	0.844	0.234	0.766
	7	0.700	0.534	0.700	0.691	0.840	0.715	0.843	0.233	0.766
	10	0.701	0.535	0.700	0.690	0.840	0.718	0.844	0.234	0.765
	11	0.700	0.534	0.699	0.690	0.840	0.718	0.843	0.233	0.765

4.1.7. k-fold Cross-validation Results of the Best-fitted Two Algorithms: kNN and RF to Forecast the Humidity of Rangpur

Based on the findings presented above, this study has determined that the kNN and RF algorithms produce greater efficiency than other algorithms for the Rangpur station. These algorithms attain the maximum accuracy with the maximum kappa coefficient. k-fold cross-validation methods, which are now being applied to the kNN and RF algorithms, are being used to assess the efficacy and efficiency of the best models that fit the data. Table 6 shows the results of the evaluation of the kNN and RF algorithms with the kCV technique. This study utilizes the numbers 2, 3, 5, 7, 10, and 11 for the parameter k to partition the provided data set.

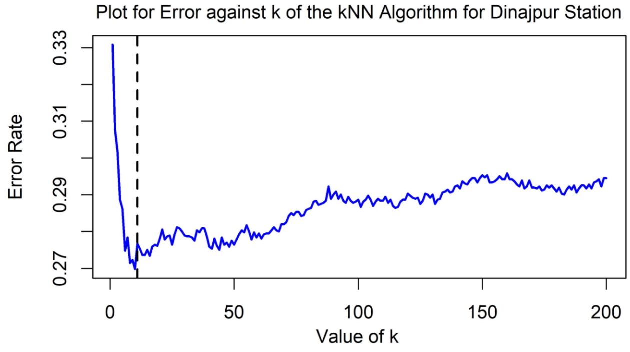

Figure 7. *Optimum value of the kNN Model at Dinajpur Station.*

According to Table 6, the average accuracy of kNN and RF algorithms is approximately 70% after model training and validation completion. Every fold attains a degree of accuracy that does not exhibit a statistically significant difference. The kNN algorithm works effectively with k equal to 10 and has an average accuracy of 69.1 percent. The RF algorithm achieves its best accuracy on fold-5, with an average accuracy of 71%. Results from K-fold cross-validation testing show that the random forest method is more practical and valuable for humidity forecasting at the Rangpur station, where this research was conducted.

4.2. Fitted Models to Forecast the Humidity of the Dinajpur Area

4.2.1. kNN to Forecast the Humidity of Dinajpur

At Dinajpur station, 11 was chosen as the highest possible value of k since it had the lowest mistake rate. Figure 7 displays the results of the suggested approaches to the dataset with $k = 11$ values.

Table 7, and Table 8 present the confusion matrices that display the outcomes of all ML classification algorithms and the performance of the different statistics for the Dinajpur station. On the training dataset, the kNN algorithm provided 78% of sensitivity for class H1, 62.63% of sensitivity for class H2, and 74.15% of sensitivity for class H3. The specificity was 83.37% for class H1, 79.65% for class H2, and 91.57% for class H3. On the test dataset, the sensitivity was 81.14% for class H1, 65.19% for class H2, and 74.94% for class H3. The specificity was 84.80% for class H1, 80.44% for class H2, and 92.53% for class H3. Table 8 presents that the kappa coefficient values for the training and test datasets are 0.5537 and 0.5856, respectively, indicating moderate agreement. It achieves an accuracy of 71% for the training dataset and 73% for the test dataset. The test dataset exhibits the highest F_1-Score of 76.54% for the lowest category of humidity (H1), indicating its superior performance. As a result, humidity forecasting accuracy on the test dataset is pretty well.

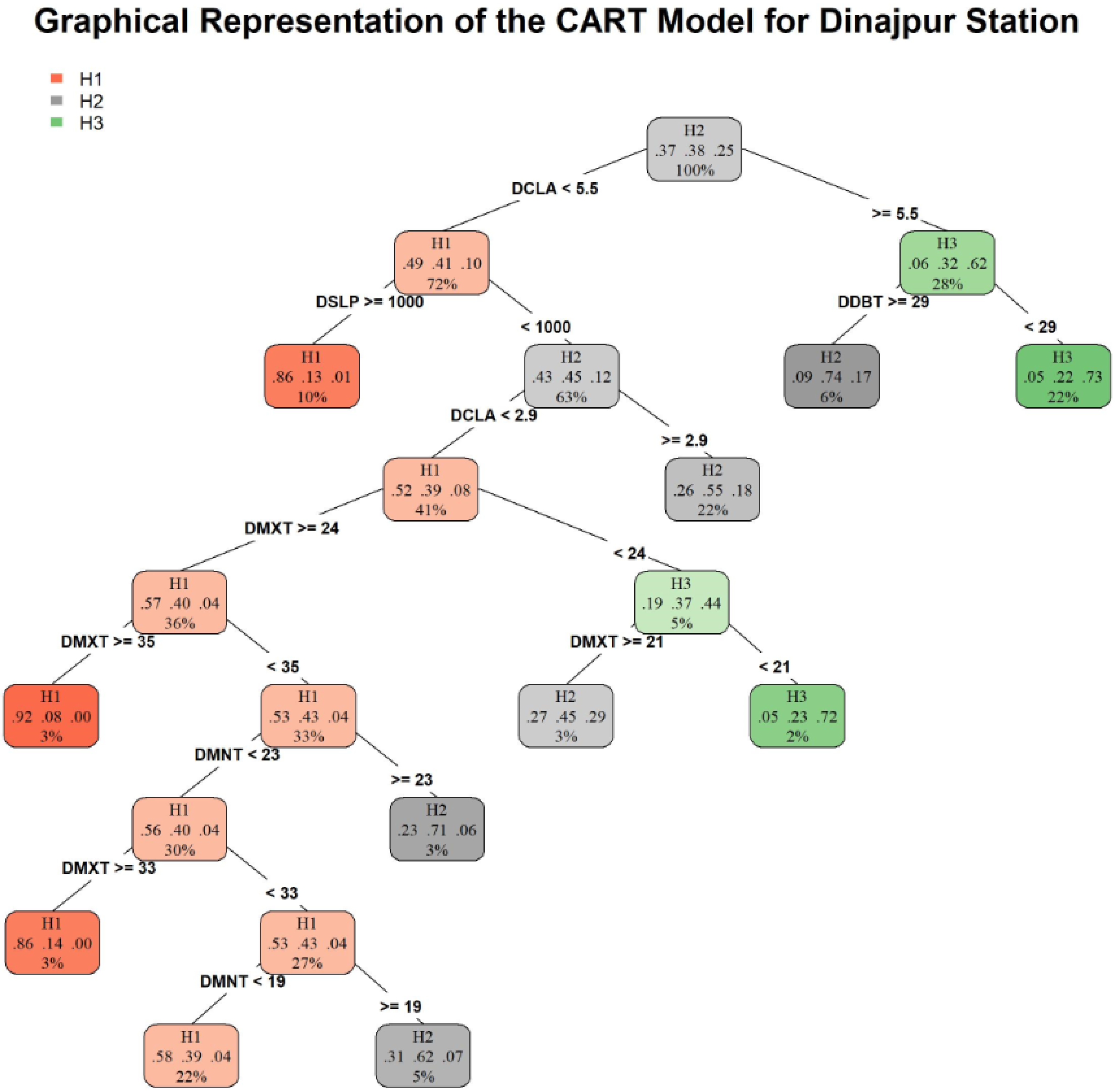

Figure 8. A graphical representation of the pruned tree for the CART model using the best CP for forecasting the humidity at Dinajpur Station.

4.2.2. CART to Forecast the Humidity of the Dinajpur

Now, prune the data and plot the final tree model for the Dinajpur station using the CP value with the lowest error. It also shows that the best CP is 0.007 for the CART model to predict the humidity of Dinajpur, which contains the lowest error. Use the best CP that best describes most of the data to fit the final best CART model. The resultant CART model is shown in Figure 8.

On the training dataset, the CART model provided 69.95% of sensitivity for class H1, 59.33% of sensitivity for class H2, and 73.35% of sensitivity for class H3. The specificity was 83.15% for class H1, 74.62% for class H2, and 90.56% for class H3. On the test dataset, the sensitivity was 70.33% for class H1, 59.71% for class H2, and 70% for class H3. The specificity was 83.84% for class H1, 74.46% for class H2, and 89.52% for class H3. Table 8 demonstrates that the kappa coefficient values for the training and test datasets are 0.49 and 0.48, respectively, indicating moderate agreement. This model attains an accuracy of 66.69% for the training dataset and 66.31% for the test dataset. The training dataset exhibits the highest F_1-Score of 72% for the highest category of humidity (H3), indicating its superior performance. As a result, humidity forecasting accuracy on the training dataset is quite good.

4.2.3. C5.0 to Forecast the Humidity of the Dinajpur

In this visual representation of the c5.0 method, the decision tree with its four leaf nodes (classifiers) is presented in Figure 9. First, it splits node 1 into nodes 2 and 5 based on the value of DCLA. Then, it splits node two into node three and four and node 5 splits into nodes 6 and 7 based on the value of DCLA. Observations in node 3 are labeled as class H1, observations in node 4 as class H2, observations in node 6 as class H2, and observations in node 7 as class H3.

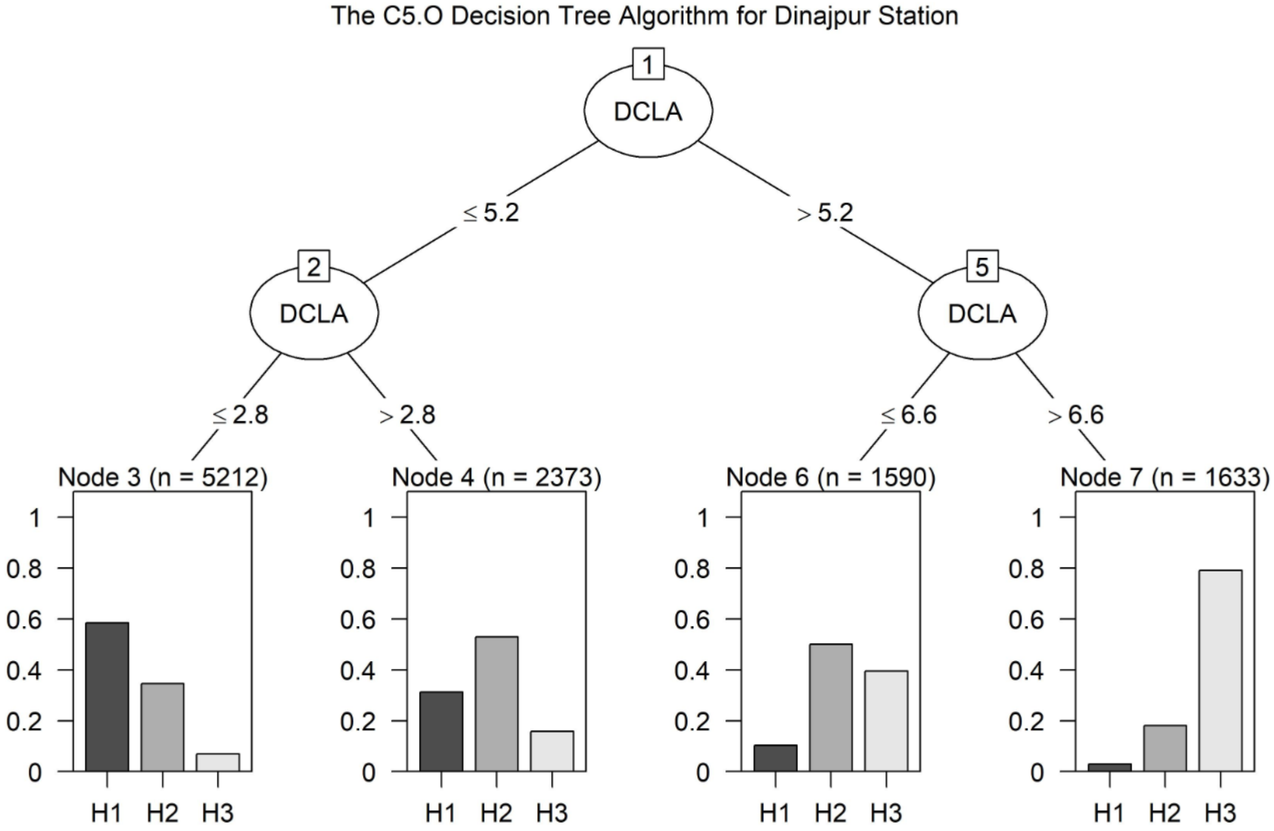

Figure 9. *The decision tree of the C5.0 algorithm to forecast humidity for Dinajpur station.*

The decision tree uses only one of the five variable measure attributes (DCLA). On the training dataset, the proposed algorithm provided 58.46% of sensitivity for class H1, 51.78% of sensitivity for class H2, and 80.15% of sensitivity for class H3. The specificity was 82.93% for class H1, 69.38% for class H2, and 85.09% for class H3. On the test dataset, the sensitivity was 58.38% for class H1, 52.27% for class H2, and 77.74% for class H3. The specificity was 83.01% for class H1, 69.25% for class H2, and 84.74% for class H3. Table 8 shows that the kappa coefficient values for the training and test datasets are 0.3647 and 0.3594, respectively, indicating fair agreement. This model attains an accuracy of 59% for the training dataset and 58% for the test dataset. The training dataset achieves the highest F_1-Score of 66.14% for the lowest category of humidity (H1), indicating its superior performance. As a result, humidity forecasting accuracy on the training dataset is good.

4.2.4. Naive Bayes to Forecast the Humidity of the Dinajpur

The NB algorithm achieved the maximum sensitivity, specificity, and F_1-measure, which were 81% of H1 for the training dataset, 82% of H1 for the testing dataset, and 68% of H1 for the testing dataset, respectively. In the test dataset, NB algorithm performed at a maximum accuracy of 61%.

4.2.5. Random Forest to Forecast the Humidity of the Dinajpur

Figure 10 shows the error rate and the number of trees. The model comprises 500 trees, and two variables are evaluated at every split. Accordingly, the classification error of H1 is 23.34%, while that of H2 and H3 are 31.19% and 24.49%, respectively. Moreover, the RF algorithm's out-of-bag error (OOB) is estimated to be 26.64%. For this station, the DCLA variable is the most important, followed by the DMXT variable, then DMNT, DDBT, DRAN, and DSLP. The Gini reduction values are as follows: DRAN has a mean decrease Gini of 503.9100, DSLP is 377.5477, DCLA is 1374.0417, DDBT is 1187.6672, DMXT is 1290.2070, and DMNT is 1243.8667.

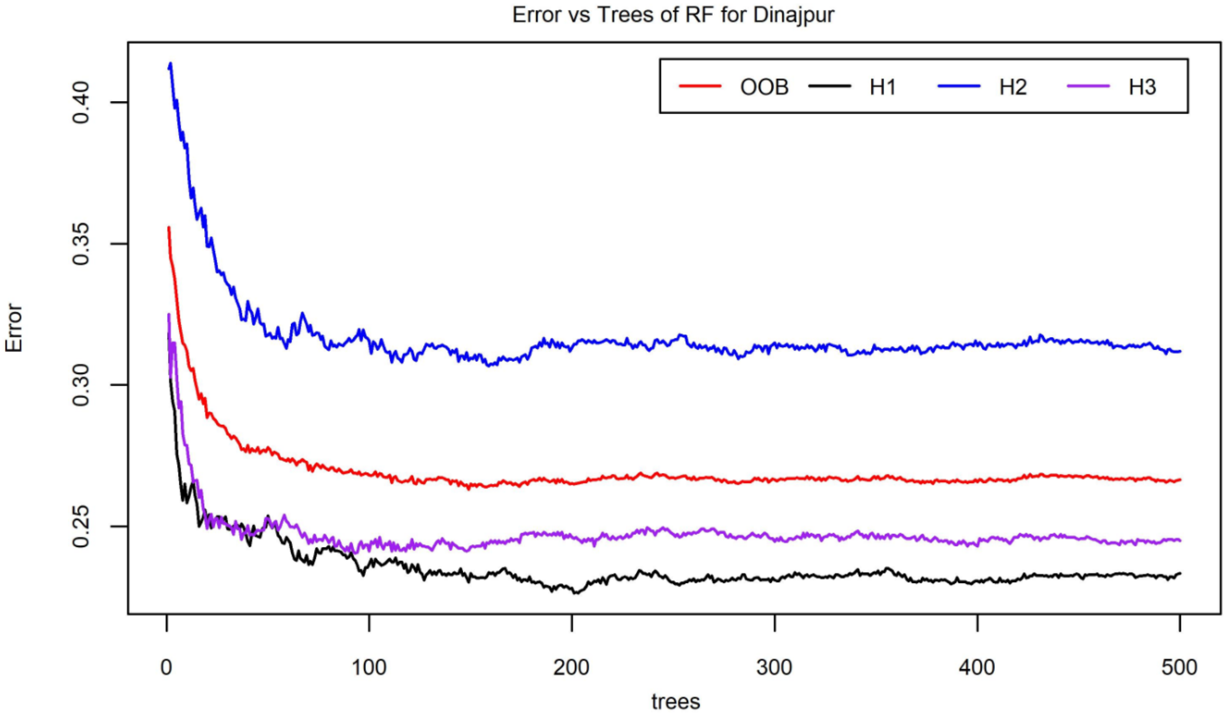

Figure 10. *Error rate of RF algorithm to forecast humidity for Dinajpur station.*

Table 7. *The confusion matrix of various ML algorithms for training data and test data at Dinajpur station using meteorological data, 1981-2020.*

Name	**Training Dataset**				**Test Dataset**			
	Category	**H1**	**H2**	**H3**	**Category**	**H1**	**H2**	**H3**
kNN	H1	2794	1078	130	H1	964	344	23
	H2	705	2886	557	H2	206	989	202
	H3	43	644	1971	H3	18	184	673
CART	H1	2873	1004	125	H1	972	313	46
	H2	1135	2455	558	H2	368	833	196
	H3	1135	2455	558	H3	42	249	584
C5.0	H1	3047	907	48	H1	1014	300	17
	H2	1801	2052	295	H2	594	702	101
	H3	364	1004	1290	H3	129	341	405
NB	H1	3165	787	50	H1	1095	220	16
	H2	1876	2062	210	H2	584	752	61
	H3	503	1105	1050	H3	177	356	342
RF	H1	3068	892	42	H1	1037	279	15
	H2	866	2854	428	H2	260	976	161
	H3	32	619	2007	H3	19	214	642
SVM	H1	2948	990	64	H1	977	333	21
	H2	808	2871	469	H2	236	994	167
	H3	20	671	1967	H3	4	241	630

Based on the training dataset, the RF provided 77.36% of sensitivity for class H1, 65.38% of sensitivity for class H2, and 81.03% of sensitivity for class H3. The specificity was 86.35% for class H1, 79.92% for class H2, and 92.19% for class H3. On the test dataset, the sensitivity was 78.80% for class H1, 66.44% for class H2, and 78.48%

for class H3. The specificity was 87.14% for class H1, 80.27% for class H2, and 91.63% for class H3. Table 8 demonstrates that the kappa coefficient values for the training and test datasets are 0.5919 and 0.5964, respectively, indicating moderate agreement. The proposed model achieves a classification accuracy of 73% for the training dataset and 74% for the test dataset. The test dataset exhibits the highest F_1-Score of 78.35% for the lowest category of humidity (H1), indicating its superior performance. RF achieves the maximum classification accuracy for forecasting humidity at the Dinajpur station and performs well for both datasets.

Table 8. *The performance of the different statistics for forecasting humidity at Dinajpur Station.*

Name	**Statistics**	**Training Dataset**			**Test Dataset**		
		H1	**H2**	**H3**	**H1**	**H2**	**H3**
kNN	Sensitivity	0.7888	0.6263	0.7415	0.8114	0.6519	0.7494
	Specificity	0.8337	0.7965	0.9157	0.8480	0.8044	0.9253
	Precision	0.6982	0.6958	0.7415	0.7243	0.7079	0.7691
	F_1-Score	0.7407	0.6592	0.7415	0.7654	0.6788	0.7592
	Accuracy	0.7079			0.7288		
	Kappa	0.5537			0.5856		
CART	Sensitivity	0.6995	0.5933	0.7335	0.7033	0.5971	0.7070
	Specificity	0.8315	0.7462	0.9056	0.8384	0.7446	0.8952
	Precision	0.7179	0.5919	0.7073	0.7303	0.5963	0.6674
	F_1-Score	0.7086	0.5926	0.7202	0.7165	0.5967	0.6867
	Accuracy	0.6669			0.6631		
	Kappa	0.4907			0.4836		
C5.0	Sensitivity	0.5846	0.5178	0.7900	0.5838	0.5227	0.7744
	Specificity	0.8293	0.6938	0.8509	0.8301	0.6925	0.8474
	Precision	0.7614	0.4947	0.4853	0.7618	0.5025	0.4629
	F_1-Score	0.6614	0.5060	0.6013	0.6610	0.5124	0.5794
	Accuracy	0.5914			0.5887		
	Kappa	0.3647			0.3594		
NB	Sensitivity	0.5709	0.5215	0.80153	0.5900	0.5663	0.81623
	Specificity	0.8410	0.6957	0.83070	0.8649	0.7165	0.83260
	Precision	0.7909	0.4971	0.39503	0.8227	0.5383	0.39086
	F_1-Score	0.6631	0.5090	0.52923	0.6872	0.5519	0.52859
	Accuracy	0.5807			0.6075		
	Kappa	0.3448			0.3854		
RF	Sensitivity	0.7736	0.6538	0.8103	0.7880	0.6644	0.7848
	Specificity	0.8635	0.7992	0.9219	0.8714	0.8027	0.9163
	Precision	0.7666	0.6880	0.7551	0.7791	0.6986	0.7337
	F_1-Score	0.7701	0.6705	0.7817	0.7835	0.6811	0.7584
	Accuracy	0.7336			0.7369		
	Kappa	0.5919			0.5964		
SVM	Sensitivity	0.7807	0.6335	0.7868	0.8028	0.6339	0.7702
	Specificity	0.8501	0.7965	0.9168	0.8516	0.8020	0.9120
	Precision	0.7366	0.6921	0.7400	0.7340	0.7115	0.7200
	F_1-Score	0.7580	0.6615	0.7627	0.7669	0.6705	0.7442
	Accuracy	0.7204			0.7219		
	Kappa	0.5717			0.5730		

4.2.6. SVM to Forecast the Humidity of the Dinajpur

Table 8 summarizes the performance metrics, including sensitivity, specificity, precision, and F_1-Score. The SVM algorithm yields a maximum sensitivity score of 78% for H3 for the training and 81% for the test dataset. The specificity values that are at their highest for H3 are 91.68 and 91.20 percent, respectively. The maximum precision value (true positive rate) can be found in the training dataset and test dataset are approximately 74% for H3 and H1 categories. Moreover, class H3 displays the highest F_1-Score of 76%, implying that it possesses the greatest ability to accurately classify observations. Table 8 also shows that the highest kappa value is 57.3%, indicating that the agreement is considered moderate for the test dataset. In addition, SVM achieves a classification accuracy of 72.04% for the training dataset and 72.19% for the test dataset which is well.

4.2.7. *k-fold Cross-validation Results of the Best-fitted Two Algorithms: kNN and RF to Forecast the Humidity of Dinajpur*

k-fold cross-validation results of the best-fitted two algorithms: kNN and RF: Based on the findings presented above, this study has also determined that the kNN and RF algorithms produce greater efficiency than other algorithms for the Dinajpur station. These algorithms attain the maximum accuracy with the maximum kappa coefficient. The k-fold cross-validation methods, which are now being applied to the kNN and RF algorithms, are being used to assess the efficacy and efficiency of the best models that fit the data. The table 9 shows the results of the evaluation of the kNN and RF algorithms with the k-fold cross-validation technique. This study utilizes the numbers 2, 3, 5, 7, 10, and 11 for the parameter k to partition the provided data set.

Table 9. *Comparison of the performance of ML algorithms in percent for forecasting humidity at Dinajpur station. Here, NPV means Neg Pred Value, DER means Detection Rate, and BAC means Balanced Accuracy.*

Algorithm	k	Accuracy	Kappa	F_1 Score	Recall	Specificity	Precision	NPV	DER	BAC
kNN	2	0.718	0.569	0.725	0.722	0.854	0.731	0.854	0.239	0.788
	3	0.724	0.578	0.730	0.728	0.857	0.736	0.857	0.241	0.792
	5	0.725	0.580	0.732	0.730	0.858	0.738	0.857	0.242	0.794
	7	0.725	0.580	0.732	0.730	0.858	0.738	0.858	0.242	0.794
	10	0.726	0.581	0.732	0.730	0.858	0.738	0.858	0.242	0.794
	11	0.726	0.582	0.733	0.731	0.858	0.738	0.858	0.242	0.795
RF	2	0.732	0.589	0.738	0.734	0.860	0.742	0.861	0.244	0.797
	3	0.732	0.589	0.738	0.734	0.861	0.742	0.861	0.244	0.797
	5	0.735	0.594	0.741	0.737	0.862	0.747	0.862	0.245	0.800
	7	0.734	0.592	0.740	0.737	0.861	0.745	0.862	0.245	0.799
	10	0.737	0.598	0.743	0.740	0.863	0.748	0.864	0.246	0.802
	11	0.736	0.596	0.742	0.738	0.863	0.747	0.863	0.245	0.800

According to Table 9, the average accuracy of kNN and RF algorithms is higher than 70% after model training and validation completion. Every fold attains a degree of accuracy that does not exhibit a statistically significant difference. The kNN algorithm works effectively with k equal to 11 and has an average accuracy of 72.6 percent. The RF algorithm achieves its best accuracy on fold-10, with an average accuracy of 74%. Results from K-fold cross-validation testing show that the RF algorithm is more practical and valuable for humidity forecasting, where this research was conducted.

5. Conclusion

Forecasting humidity in Bangladesh poses a challenge due to the country's nonlinear trends and the unpredictable nature of the data in terms of both spatial and temporal dimensions. However, the presence of humidity is crucial in creating a conducive and salubrious habitat for living beings. Humidity is the term used to describe the existence of water vapor in the atmosphere. As mentioned earlier, the factor has a notable impact on the quality of air, management of climate, production of agriculture, and yields of crops. This research aims to apply machine learning algorithms to forecast humidity in the northern Bangladesh by making use of meteorological parameters. The Rangpur and Dinajpur stations in northern Bangladesh are considered research areas to forecast humidity. Through kNN, CART, C5.0, NB, RF, and SVM algorithms, the influence and relationship of meteorological factors such as RAN, SLP, CLA, DBT, MXT, MNT, and HUM are discussed. The information utilized in this research was obtained from the Bangladesh Meteorological Department. Throughout this extended duration, the Rangpur station recorded the highest daily average humidity level of approximately eighty, while the Dinajpur station recorded the lowest level of approximately seventy-seven. The correlation analysis conducted on the seven meteorological parameters of two stations indicates a moderate positive correlation between the ”daily humidity and daily average cloud amount”. The association between ”daily humidity and daily mean sea level pressure”, as well as ”daily humidity and daily maximum temperature”, showed a weak negative correlation and is deemed negligible. This study found that humidity reduces when temperature and mean sea level pressure rises and increases as cloud average rises. The kNN, RF, and SVM algorithms demonstrated good fitting performance on the training and testing set at each station. At Rangpur station, the kNN and RF algorithms demonstrated a good fit for the testing set. The accuracy achieved by the RF algorithm is seventy percent. The kappa coefficient has been calculated to be 0.533, indicating a moderate level of consistency for the algorithm. When compared to other algorithms on both training and testing datasets, the accuracy of k nearest neighbor, random forest, and support vector machine are better in Dinajpur station, never falling below seventy percent. Among the three models assessed at Dinajpur station, the random forest demonstrates the highest accuracy, reaching about seventy-four percent. The coefficient of agreement, commonly known as kappa, has been determined to be sixty percent, indicating a moderate to good consistency in the model. The study also found that ensemble-based learning techniques, such as the random forest algorithm, produce the maximum performance for both regions compared to other machine learning algorithms. It attained the most significant possible classification accuracy of seventy percent for the Rangpur station and seventy-four percent for the Dinajpur station while using the testing dataset. The k-fold cross-validation method is also used to assess the performance of this research and suggested kNN and RF classification algorithms as best. There is little to no discernible difference between the degrees of accuracy achieved by the various folds. Across all experiments, the performances of various classifications were compared, and the results indicated that the random forest algorithm performed better than other algorithms. Specifically, the random forest algorithm consistently achieved an accuracy rate of over seventy percent for all stations. Additionally, the k-fold cross-validation method shows that the random forest model is highly efficient. As a result, the evaluation of the random forest algorithm utilizing k-fold cross-validation as an evaluation technique has shown that it can effectively share training and test data and may be appropriate for forecasting the humidity in the northern area of Bangladesh. Bangladesh has experienced alterations and variations in its humidity levels in the last forty years. The impact of this phenomenon extends to various domains, including agriculture, plant and animal life, human health, climate change, drought detection, weather forecasting, energy management, and industrial applications. The results of this investigation suggest that the random forest algorithm will be appropriate for predicting the humidity of the northern part of Bangladesh, which could be applied to other meteorological stations. However, this research only considers two meteorological stations in the northern region of Bangladesh. In the future, an attempt will be made to cover all meteorological stations in Bangladesh to obtain a comprehensive understanding of changes in humidity across the nation.

Acknowledgments

The authors express their sincere gratitude to the Bangladesh Meteorological Department (BMD) for providing the data essential for this study. They also appreciate the valuable support and insightful feedback from the Editor and anonymous reviewers, which significantly improved the quality of the manuscript.

Ethics Approval

Not applicable.

Consent for Publication

Not applicable.

Availability of Data

The information presented in this article is sourced from the Bangladesh Meteorological Department (BMD), and the data can be accessed at http://live.bmd.gov.bd/ and http://www.bmddataportal.com/#/. The specific datasets utilized in this study are available from the corresponding author upon request. Please reach out to the corresponding author for additional details.

ORCID

0009-0006-5151-4388 (Most. Rubina Akter)
0000-0002-3972-3711 (Md. Habibur Rahman)

Funding

Not applicable.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

[1] Islam, M. M., 2014. Regional Differentials of Annual Average Humidity over Bangladesh. ASA University Review, 8(1), pp. 1-14.

[2] Abu-Taleb, A. A., Alawneh, A. J. and Smadi, M. M., 2007. Statistical analysis of recent changes in relative humidity in Jordan. American Journal of Environmental Sciences, 3(2), pp. 75-77.

[3] Arundel, A. V., Sterling, E. M., Biggin, J. H. and Sterling, T. D., 1986. Indirect health effects of relative humidity in indoor environments. Environmental health perspectives, 65, pp. 351-361. https://doi.org/10.1289/ehp.8665351

[4] Salim, M. J. N. P., 1989. Effects of salinity and relative humidity on growth and ionic relations of plants. New Phytologist, 113(1), pp. 13-20. https://doi.org/10.1111/j.1469-8137.1989.tb02390

[5] Assmann, S. M. and Grantz, D. A., 1990. The magnitude of the stomatal response to blue light: modulation by atmospheric humidity. Plant Physiology, 93(2), pp. 701-707. https://doi.org/10.1104/pp.93.2.701

[6] Chowdhury, M., Mondal, S. and Islam, J., 2018. Modeling and forecasting humidity in Bangladesh: box-jenkins approach. International Journal of Research, 6(4), pp. 50-60, https://doi.org/10.29121/granthaalayah.v6.i4.2018.1475

[7] Ruane, A. C., Major, D. C., Winston, H. Y., Alam, M., Hussain, S. G., Khan, A. S., Hassan, A., Al Hossain, B. M. T., Goldberg, R., Horton, R. M. and Rosenzweig, C., 2013. Multi-factor impact analysis of agricultural production in Bangladesh with climate change. Global environmental change, 23(1), pp. 338-350, https://doi.org/10.1016/j.gloenvcha.2012.09.001

[8] Rahman, M. H., Hossain, M. M., 2019. Classification and regression tree to predict the precipitation labels of north-west region in Bangladesh. Environment and Natural Resources Research, 9(3), pp. 117-126, https://doi.org/10.5539/enrr.v9n3p117

[9] Rahman, M. H., Matin, M., Salma, U., 2018. Analysis of precipitation data in Bangladesh through hierarchical clustering and multidimensional scaling. Theoretical and Applied Climatology 134, pp. 689-705, https://doi.org/10.1007/s00704-017-2319-y

[10] Rahman, M. H., 2022. Prediction of homogeneous region over Bangladesh based on temperature: a non-hierarchical clustering approach. Theoretical and Applied Climatology, 148(3-4), pp. 1127-1149. https://doi.org/10.1007/s00704-022-03955-3

[11] Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N. and El-Shafie, A., 2021. Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal, 12(2), pp. 1651-1663. https://doi.org/10.1016/j.asej.2020.09.011

[12] Yamac, S. S. and Todorovic, M., 2020. Estimation of daily potato crop evapotranspiration using three different machine learning algorithms and four scenarios of available meteorological data. Agricultural Water Management, 228, p. 105875. https://doi.org/10.1016/j.agwat.2019.105875

[13] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 2017. Classification and Regression Trees. Routledge. https://doi.org/10.1201/9781315139470

[14] Ghiasi, M. M., Zendehboudi, S. and Mohsenipour, A. A., 2020. Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine, 192, p. 105400. https://doi.org/10.1016/j.cmpb.2020.105400

[15] Atkinson, E. J., Therneau, T. M., 2000. An introduction to recursive partitioning using the rpart routines. Rochester: Mayo Foundation.

[16] Quinlan, J. R., 1986. Induction of decision trees. Machine learning, 1, pp. 81-106. https://doi.org/10.1007/BF00116251

[17] Williams, N., Zander, S. and Armitage, G., 2006. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Computer Communication Review, 36(5), pp. 5-16. https://doi.org/10.1145/1163593.1163596

[18] Ray, S., 2019. February. A quick review of machine learning algorithms. In 2019 International Conference on Machine Learning, Big Data, cloud and Parallel Computing (COMITCon) (pp. 35-39). IEEE. https://doi.org/10.1109/COMITCon.2019.8862451

[19] Parthiban, G., Rajesh, A. and Srivatsa, S. K., 2011. Diagnosis of heart disease for diabetic patients using naive Bayes method. International Journal of Computer Applications, 24(3), pp. 7-11. https://doi.org/10.5120/2933-3887

[20] Breiman, L., 2001. Random forests. Machine learning, 45, pp. 5-32. https://doi.org/10.1023/A:1010933404324

[21] Hastie, T., 2009. The elements of statistical learning: data mining, inference, and prediction. https://doi.org/10.1111/j.1541-0420.2010.01516.x

[22] Xu, W., Zhang, J., Zhang, Q. and Wei, X., 2017, February. Risk prediction of type II diabetes based on random forest model. In 2017 third International Conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB), pp. 382-386). IEEE. https://doi.org/10.1109/AEEICB.2017.7972337

[23] Ukil, A. and Ukil, A., 2007. Support vector machine. Intelligent systems and signal processing in power engineering, pp. 161-226. https://doi.org/10.1007/978-3-540-73170-2_4

[24] Suykens, J. A., De Brabanter, J., Lukas, L. and Vandewalle, J., 2002. Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing, 48(1-4), pp. 85-105. https://doi.org/10.1016/S0925-2312(01)00644-0

[25] Rohani, A., Taki, M. and Abdollahpour, M., 2018. A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I). Renewable Energy, 115, pp. 411-422. https://doi.org/10.1016/j.renene.2017.08.061

[26] Borna, N. J. and Rahman, M. H., 2024. Evaluating the degree of cloudiness using machine learning techniques based on different atmospheric conditions. Theoretical and Applied Climatology, pp. 1-30. https://doi.org/10.1007/s00704-024-05062-x

[27] Rahman, M. H., 2024. ANN-based and DT-based Classification Approaches to Predict the Rainfall Level of the Grid ($90°E - 92°E$, $23°N - 25°N$) in Bangladesh. International Journal of Data Science and Analysis, 10(6), pp. 109-128. https://doi.org/10.11648/j.ijdsa.20241006.11