Published in AJSE, Vol:22, Issue: 1
Received on 8th March 2022
Revised on 16th March 2023
Accepted on 30th April 2023

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 – 3679 (print) 2520 – 4890 (Online)

Prediction of Flood in Bangladesh Using Different Classifier Model

Md Sajid Hossain and Mohammad Zeyad

***Abstract*— Bangladesh is highly affected by climate change scenarios notably floods due to its location on the world map in the South Asian region. Besides due to monsoon rains and high upstream rainfall in several areas eventually turn into floods. Thus, early flood forecasting might save human lives as well as agriculture crops. In this paper, we have applied different machine learning classifier models (Decision tree, Naive bayes, k-NN and Random Forest) with a view to predicting the occurrence of flood. RapidMiner tool has been used extensively to perform data preparation, decision tree model generation, cross-validation, model selection and optimization of the model parameters. It is seen that the decision tree model has performed well by achieving an accuracy of 94.23% which is further optimized to reach 94.68%. Comparison of four widely used optimized classifier models as well as the feature selection criterion using 'correlation matrix' is also a good aspect of this work by which we have reached to a good result.**

***Index Terms*— RapidMiner, Flood, DataMining, ML Classifier**

I. INTRODUCTION

Weather forecasting heavily relies on atmospheric research. There were other theoretical and technical advancements, unrelated to our atmospheric weather understanding, which helped us gain new information. Prediction of weather may generally be done utilizing remote sensing satellite data. Meteorological satellites capture photographs of the sky, which are used to generate future weather patterns by analyzing various weather data including temperature, precipitation, and cloud cover. These satellite-based systems are typically not practical due to their high cost and considerable support requirements. To accurately anticipate meteorological conditions, new methods, such as artificial neural networks, regression, or data mining techniques, have to be used in addition to the existing statistical methods.

The process of extracting useful knowledge and information from vast quantities of muddled, noisy, random, and incomplete data is known as data mining. It is a groundbreaking new technology with significant promise to help businesses reduce the amount of information they deal with by focusing on critical facts. Forecasting the weather is a kind of data mining that explores vast amounts of publicly available meteorological data in quest of earlier discovered trends [1].

Md Sajid Hossain is an Assistant Professor at the Department of EEE, American International University-Bangladesh, 408/1, Kuratoli, Khilkhet, Dhaka-1229, Bangladesh. Email: sajidh@outlook.com
Mohammad Zeyad is engaged as the Head, Energy and Technology Research Division, Advanced Bioinformatics, Computational Biology and Data Science Laboratory Bangladesh (ABCD Laboratory, Bangladesh), Chattogram-4226, Bangladesh. Email: mohammad.zeyad.eee@gmail.com

Although there are different complex tools and methods to predict and analyze the weather datasets. However, the complex computational obligations for employing these tools are very cumbersome. Furthermore, the success of employing such intelligent algorithms mainly depends upon the experience of the researcher. Hence, this work aims to use a simple Graphical User Interface (GUI) based tool which is called RapidMiner Studio to predict and analyze flood datasets. RapidMiner was positioned in the leader quadrant of Gartner's Magic Quadrant for Data Science and Machine Learning Platforms in 2019 [2]. Furthermore, readers picked RapidMiner as one of the most popular data analytics tools (after python) in a KDnuggets annual software survey (2019) [3].

II. LITERATURE REVIEW

To gather weather information and uncover concealed patterns within an extensive dataset, M. A. Kalyankar and S. J. Alaspurkar utilized data mining techniques. Subsequently, they applied these discoveries to develop novel weather forecasting systems capable of categorizing and predicting weather conditions [4]. Data mining algorithms that can learn dynamically are necessary to construct dynamic data mining models for weather and events that change quickly.

Abhishek Saxena and coworkers conduct a review of the effectiveness of using artificial neural networks for weather forecasting [5]. According to this research, artificial neural networks (or deep learning, DL) is capable of accurately forecasting several meteorological phenomena, including temperature, thunderstorms, rainfall, wind speed, and it is possible to use key architectures like BP, MLP to do so.

P. Hemalatha employed data mining methods to direct the ships using their navigational trajectories [6]. The Global Positioning System (GPS) is utilized to locate the present location of the ship. The weather encompasses numerous factors, such as climate, humidity, and temperature. When checking the weather report for a particular area, the current database is examined to identify any resemblances. Data analysis is aided by strong interaction between the statistical and computational sectors.

Data mining methods were used to investigate ad hoc weather characteristics such as time of day, month of the year, wind direction, speed, and intensity on a specific location, which needed a meteorological data set from Subana Shanmuganathan and Philip Sallis [7-8]. The effectiveness of forecasting wind gust patterns for a vineyard was established through the discovery of new information. Data mining techniques were applied to meteorological data that was captured at varying

intervals. The study focused on the vineyard of Kumeu River, analyzing data from a repository spanning four years, between 2008 and 2012. All the readings that are outside of Kumeu's records are excluded from the data. All 86,418 incidents, along with their occurrence over the course of a year, are described in this report. Algorithms which are known as decision tree include C5, Quest, CRT, and CHAID are used. SOM is utilized for clustering. ANN with several layers and supervised learning is used for forecasting wind gusts. Practical uses of SPSS include data mining techniques and statistical approaches are also utilized. Ad hoc dataset analysis is made easy with the help of this application.

The innovative proposal put out by Amarakoon included the utilization of historical meteorological data and the "K-Nearest Neighbor (KNN)" data-mining technique, which classifies the historical data into a certain time frame [9]. The data on the current weather conditions are being gathered for a full year. Accurate findings within a realistic time frame for many months are generated. Feature selection strategies may provide even more precise results when used to feature integration.

By utilizing clustering and K-Nearest Neighbor (KNN) techniques, S. S. Badhiye examined an extensive dataset on weather patterns. The objective was to reveal concealed patterns that could be employed to classify and forecast climatic conditions [10]. A high level of temperature and humidity forecast accuracy is achieved. A sensor-embedded data logger system may be used to gather, analyze, and forecast characteristics that are located at locations too far away for conventional sensors.

Data mining techniques were utilized by Pinky Saikia Dutta and Hitesh Tahbilder to forecast monthly rainfall in the state of Assam [11]. This methodology uses multiple linear regression. This dataset consists of data from the Regional Meteorological Center in Guwahati, Assam, India during the time period of 2007 to 2012. During each season, data is separated into four-month intervals. the lowest, maximum, mean sea level pressure, wind speed, and rainfall are among the model's parameters.

Neha Khandelwal and Ruchi Davey successfully predicted a year's worth of rainfall in Rajasthan by considering four meteorological factors: temperature, humidity, pressure, and sea level. Additionally, they utilized data analysis to identify potential water scarcity in the region [12]. Data mining algorithms extract certain elements. When correlating the dataset, correlation analysis is then used to identify correlations. In order to analyze the correlation of the specified components, they are picked and used for regression analysis. Regression analysis is used to forecast rainfall using MLR.

Soo-Yeon Ji correctly predicted the hourly rainfall time and location for each provided area [13]. To begin with, it is established whether or not it will rain. Hourly rainfall forecast is only done if there is any likelihood of rain. Most of the techniques used to forecast hourly predictions, however, they are constrained by data variability and the lack of data.

The weather information has a significant impact on the occurrence of floods. Bangladesh Meteorological Department (BMD) is responsible to track and record the weather information. An updated weather information dataset has been extracted from the source [14]. This dataset consists of 20544 instances which includes important attributes for predicting the flood such as Flood, Rainfall, Relative Humidity, Cloud Coverage, Maximum, Minimum Temperature, Bright Sunshine and Wind Speed. RapidMiner tool has been used extensively to prepare the dataset, feature selection, generate decision tree, model validation, model selection and optimize the parameter of the selected model.

III. METHODOLOGY AND RESULTS

A. Preparing Dataset

First, the 'Retrieve' operator has been used to import raw weather data from the csv file extracted from the source [14]. The dataset extracted contains weather information on various districts of Bangladesh. The dataset includes vital weather attributes such as Flood, Rainfall, Relative Humidity, Cloud Coverage, Maximum and Minimum Temperature, Bright Sunshine, Wind Speed etc. As our main target of this work is to predict flood so the 'Map' operator has been used to transform the binary 'Flood' attribute data into polynomial (Yes? or No?) form. Later, the 'Set Role' operator is used to make the 'Flood' attribute as a label attribute which we will predict. 'Select Attributes' has been used to filter out the important attributes from the raw data, those are Flood, Rainfall, Relative Humidity, Cloud Coverage, Maximum, Minimum Temperature, Bright Sunshine and Wind Speed. That is the end of the dataset preparation in the RapidMiner studio. Preparing the dataset is the most important step in the process of prediction or classification-related problems as the accuracy of the prediction depends largely on the good quality of data as well as the precise tunning of attribute parameters in the dataset. The whole process of the dataset preparation in RapidMiner studio has been depicted in the following Fig. 1.

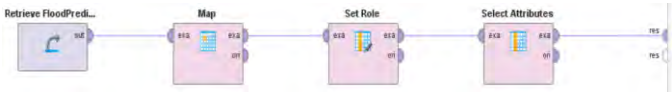


Fig. 1. Preparing Dataset using RapidMiner tool.

Dataset attributes and their unit have been given in the following table-1.

TABLE 1: DATASET ATTRIBUTES AND THEIR UNIT

Attributes	Unit
Max Temp	Celsius or °C
Min Temp	Celsius or °C
Rainfall	cm
Cloud Coverage	Okta
Bright Sunshine	Hours per day
Wind Speed	Meters per second or m/s
Relative Humidity	Percentage

The statistical view of the considered attributes in the dataset has been shown in Fig. 2.

Name	Type	Missing	Statistics	Filter (8 / 8 attributes):	
Flood?	Polynominal	0	Yes (4132)	No (16412)	No (16412), Yes (4132)
Max_Temp	Real	0	21.600	44	33.451
Min_Temp	Real	0	6.200	28.100	21.167
Rainfall	Integer	0	0	2072	198.777
Relative_Humidity	Integer	0	34	97	79.497
Wind_Speed	Real	0	0	11.200	1.415
Cloud_Coverage	Real	0	0	7.900	3.486
Bright_Sunshine	Real	0	0	11	6.419

Fig. 2. The Statistical view of the dataset

The considered attributes in the dataset have no missing values. The 'maximum temperature' attribute contains a dataset in the range of 21.6°C to 44°C on average 33.45°C. The 'minimum temperature' attribute contains a dataset in the range of 6.2°C to 28.1°C on average 21.17°C. Besides, the 'rainfall' dataset includes in the range of 0 cm to 2072 cm, the 'relative humidity' dataset contains in the range of 34% to 97%. Furthermore, the 'wind speed' attribute contains a dataset in the range of 0 m/s to 11.2 m/s, the 'cloud coverage' dataset includes in the range of 0 okta to 7.9 okta. Finally, the 'bright sunshine' dataset comprises in the range of 0 hours/day to 11 hours/day.

B. Feature Selection

The process of picking the right attributes to employ in the model's construction is known as feature selection. For feature selection, correlation analyses have been performed. The statistical link between any two attributes is called correlation. For feature selection, a 'Correlation Matrix' operator has been used along with pre-processed data in RapidMiner Studio as shown in Fig. 3. Pre-processed data block contains the data preparation steps as mentioned in section A.

Fig. 3. Feature Selection Method of Pre-processed Data using 'correlation matrix'

After correlation analysis, any correlation **above 0.7 and below -0.7** have been considered significant for this pre-processed dataset. Thus, three main features have been selected from the correlation matrix: Rainfall, Cloud Coverage and Bright Sunshine. As our main objective is to predict floods thus the features i.e. Rainfall, Cloud Coverage and Bright Sunshine must have a significant impact on the occurrence/non-occurrence of the flood. Bright Sunshine and Cloud Coverage are negatively correlated which means the brighter the sunshine the lesser the cloud coverage. Besides, Rainfall and Cloud Coverage are positively correlated which means the higher the rainfall hence the higher the cloud coverage. Unrelated attributes may influence the model adversely thus producing undesirable outcomes [15]; hence they were not chosen. Fig. 4 shows the result of the feature selection using the correlation matrix. Pearson Correlation method has been used in this case to relate between two sets of data. The boxed attributes in black color have been considered significant as those parameters lie in the range above 0.7 and below -0.7.

Attributes	Max_Temp	Min_Temp	Rainfall	Relative_H...	Wind_Spe...	Cloud_Co...	Bright_Su...
Max_Temp	1	0.699	0.257	0.029	0.308	0.463	-0.127
Min_Temp	0.699	1	0.595	0.538	0.383	0.828	-0.514
Rainfall	0.257	0.595	1	0.592	0.320	0.766	-0.676
Relative_Humidity	0.029	0.538	0.592	1	0.096	0.660	-0.667
Wind_Speed	0.308	0.383	0.320	0.096	1	0.389	-0.174
Cloud_Coverage	0.463	0.828	0.766	0.660	0.389	1	-0.743
Bright_Sunshine	-0.127	-0.514	-0.676	-0.667	-0.174	-0.743	1

Fig. 4. Feature Selection using 'correlation matrix'

C. Machine Learning Classifier: Decision Tree Model

The decision tree is a very common classification algorithm for predicting the label attribute. As our objective is to predict the flood occurrence thus a decision tree model has been used as a machine learning classifier. For that selected featured attributes have been used to model the algorithm as shown in Fig. 5.

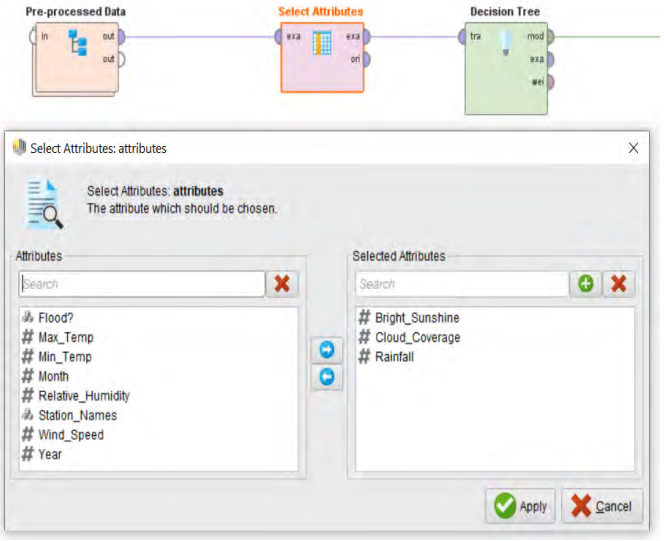


Fig. 5. Preparing a Decision tree model in RapidMiner Studio using the selected featured attributes.

According to feature selection using the correlation matrix, bright sunshine, cloud coverage and rainfall have been considered significant attribute. As shown in Fig. 5, we have only included those attributes in the decision tree classifier model. Unrelated attributes may have a negative impact on the model, resulting in unsatisfactory outputs. Hence, maximum or minimum temperature, relative humidity, wind speed, etc. parameters have been excluded in the decision tree classifier model. Fig. 6 shows the decision tree based on selected features (Rainfall, Cloud Coverage and Bright Sunshine) using the gain ratio criterion.

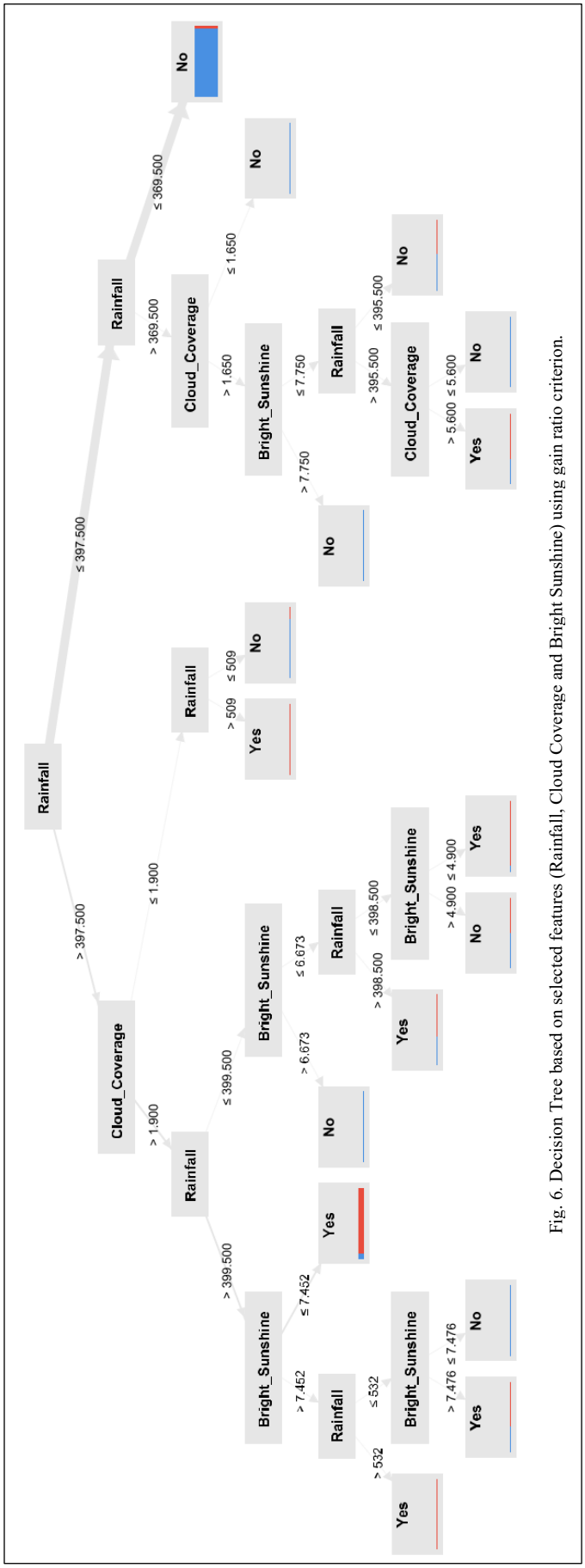


Fig. 6. Decision Tree based on selected features (Rainfall, Cloud Coverage and Bright Sunshine) using gain ratio criterion.

Our label attribute is the occurrence of floods (i.e. Flood?), thus a Decision Tree based on selected features (Rainfall, Cloud Coverage and Bright Sunshine) using gain ratio criterion has been generated using the RapidMiner tools. A tree description has been generated using the RapidMiner tool as shown in Fig. 7.

Tree

```
Rainfall > 397.500
|   Cloud_Coverage > 1.900
|   |   Rainfall > 399.500
|   |   |   Bright_Sunshine > 7.452
|   |   |   |   Rainfall > 532: Yes {No=0, Yes=8}
|   |   |   |   Rainfall ≤ 532
|   |   |   |   |   Bright_Sunshine > 7.476: Yes {No=15, Yes=26}
|   |   |   |   |   Bright_Sunshine ≤ 7.476: No {No=2, Yes=0}
|   |   |   Bright_Sunshine ≤ 7.452: Yes {No=257, Yes=3217}
|   |   Rainfall ≤ 399.500
|   |   |   Bright_Sunshine > 6.673: No {No=2, Yes=0}
|   |   |   Bright_Sunshine ≤ 6.673
|   |   |   |   Rainfall > 398.500: Yes {No=8, Yes=12}
|   |   |   |   Rainfall ≤ 398.500
|   |   |   |   |   Bright_Sunshine > 4.900: No {No=2, Yes=2}
|   |   |   |   |   Bright_Sunshine ≤ 4.900: Yes {No=1, Yes=10}
|   Cloud_Coverage ≤ 1.900
|   |   Rainfall > 509: Yes {No=0, Yes=3}
|   |   Rainfall ≤ 509: No {No=5, Yes=1}
Rainfall ≤ 397.500
|   Rainfall > 369.500
|   |   Cloud_Coverage > 1.650
|   |   |   Bright_Sunshine > 7.750: No {No=4, Yes=0}
|   |   |   Bright_Sunshine ≤ 7.750
|   |   |   |   Rainfall > 395.500
|   |   |   |   |   Cloud_Coverage > 5.600: Yes {No=5, Yes=9}
|   |   |   |   |   Cloud_Coverage ≤ 5.600: No {No=13, Yes=0}
|   |   |   |   Rainfall ≤ 395.500: No {No=219, Yes=202}
|   |   Cloud_Coverage ≤ 1.650: No {No=5, Yes=0}
|   Rainfall ≤ 369.500: No {No=15874, Yes=642}
```

Fig. 7. Decision Tree description

D. Decision Tree Model Validation (Cross-validation)

To validate the decision tree model, a 'Cross-Validation' operator has been used. In this operator, the number of folds is taken 10, which means it will take the pre-processed dataset and break it into 10 smaller subsets of data. Then it will go ahead and build a model of 9 of those keeping one of them for testing and then it will iterate. Fig. 8 shows the process where Pre-processed data is fed into the Cross-validation operator.

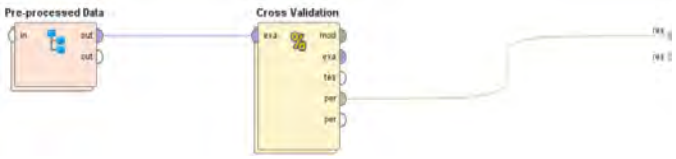


Fig. 8. Pre-processed data is fed into the Cross-validation operator.

The Cross-Validation Operator consists of two sub-processes, namely training and testing. In the training subprocess, a model is trained, which is then utilized in the testing subprocess. During the testing phase, the 'Performance' operator is used to evaluate the model's performance. A model's performance on independent test sets is an excellent indicator of how well it will

perform on unknown datasets [16]. Fig. 9 shows the cross-validation sub-process which comprises the training and testing phase.

Fig. 9. Cross-validation Sub-process: Training and Testing

E. Machine Learning Classifier: Model Selection

In this section, we are interested to apply different machine learning classifiers (i.e. Decision tree, Naive bayes, k-NN and Random forest) to compare the performance among them. 'Compare ROCs' operator has been used as shown in Fig. 10 to compare the 'Area Under the Receiver Operating Characteristics' curve (AUC-ROC) of different classifier models.

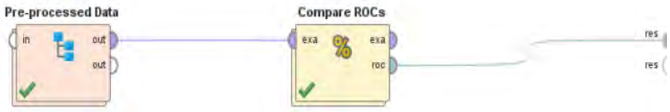


Fig. 10. Compare ROCs operator to generate ROC curve of different classifier model.

In the 'compare ROCs' operator, there are both the 'number of folds' parameter to do cross-validation and split ratio' parameter to be used to create a training and test set. So, the 'compare ROCs' operator loops over every algorithm it contains to determine the accuracy and generate the ROC curve. RapidMiner Studio will automatically perform the cross-validation so that there will be no training error. Fig. 11 shows the 'Compare ROCs' parameter selection window where cross-validation and split ratio settings are performed.

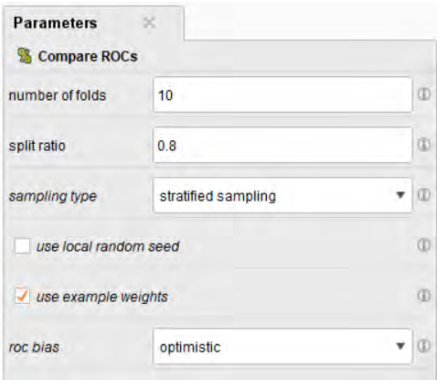


Fig. 11. Compare ROCs parameter selection

AUC - ROC curve is a performance measurement criterion for classification-based problems. Decision tree, Naive bayes, k-NN and Random forest classifier have been selected based on the suggestion provided by the rapid miner user community on classification-related problems [17]. The following Fig. 12 is showing the Compare ROCs Sub-process in the RapidMiner Studio tool.

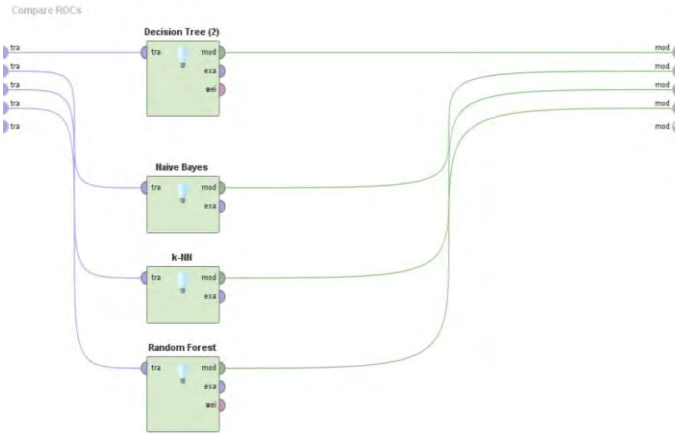
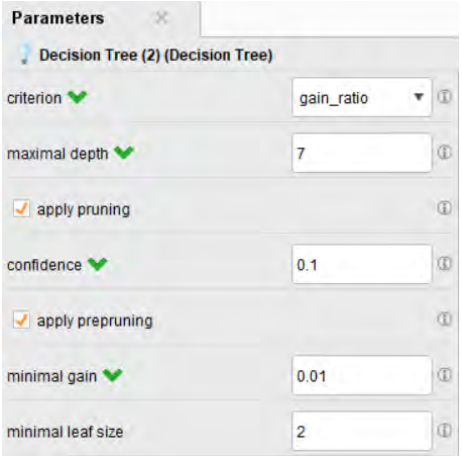
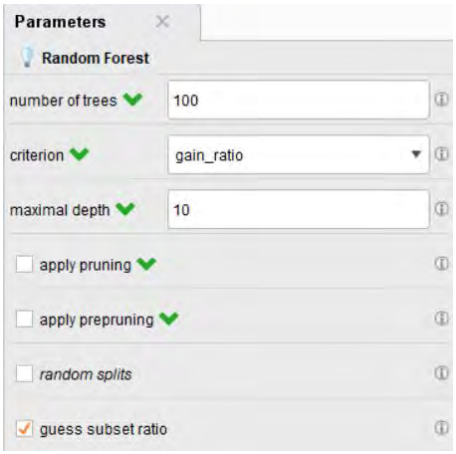


Fig. 12. Compare ROCs Sub-process: Decision tree, Naive bayes, k-NN and Random Forest

The below Fig. 13 shows the different parameters of four models set in the RapidMiner Studio tool.

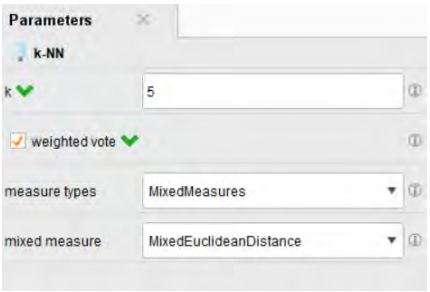


(i) Model parameter of Decision Tree



(ii) Model parameter of Random Forest

(iii) Model parameter of Naïve Bayes



(iv) Model parameter of k-NN
Fig. 13. Model parameters of four different classifiers

AUC is a probability curve, whereas ROC is a measure of separability. It describes how effectively the model can distinguish between different classes. The model is better at classifying data according to AUC, the higher the AUC (Yes or No). From the below Fig. 14, it is seen that the Decision Tree has a higher AUC than the other classifier model. Besides, according to Table 2, it is clear that the highest accuracy, precision and recall performance score of the models are 94.23%, 91.08% and 79.04% respectively. So, the Decision Tree classifier is a good choice for selecting the model of this problem. However, further optimization of the different parameters of the model can improve the ROC curve performance. In the next section, we will optimize the parameter of the Decision Tree model and compare the performance with and without optimization.

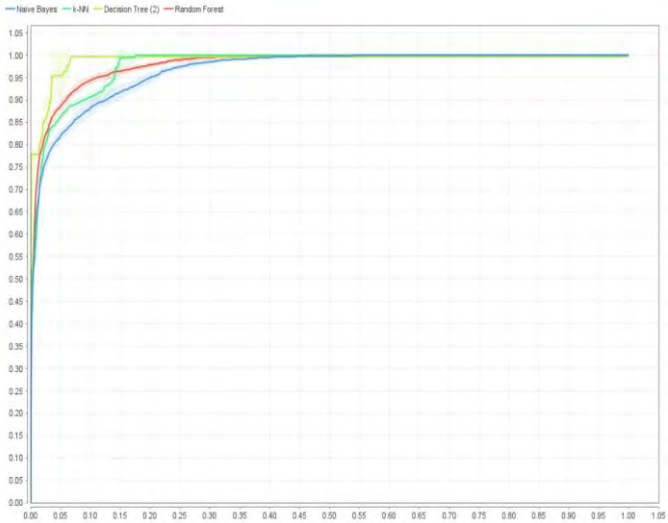


Fig. 14. Compare ROCs curve of Decision tree, Naive bayes, k-NN and Random Forest classifier.

TABLE 2: PERFORMANCE SCORE

Model	Accuracy	Precision	Recall
Naive Bayes	89.23%	90.36%	78.16%
k-NN	92.35%	90.84%	78.54%
Decision Tree	94.23%	91.08%	79.04%
Random Forest	87.23%	88.68%	77.14%

F. Optimize Parameter of the Selected model: Decision Tree

To optimize the decision tree model, the Optimize Parameter (Grid) operator is employed. The layered operator performs a subprocess for all possible parameter value combinations before sending the most effective parameter values to the parameter set port. The matching model is provided by the model port, while the performance port delivers the performance vector for the optimal parameter values. The output ports are used to present any supplementary results from the top-performing run. The performance value presented to the inner performance port determines the optimal parameters [16]. Fig. 15 shows the Optimize Parameter (Grid) which is used for optimizing the decision tree model.

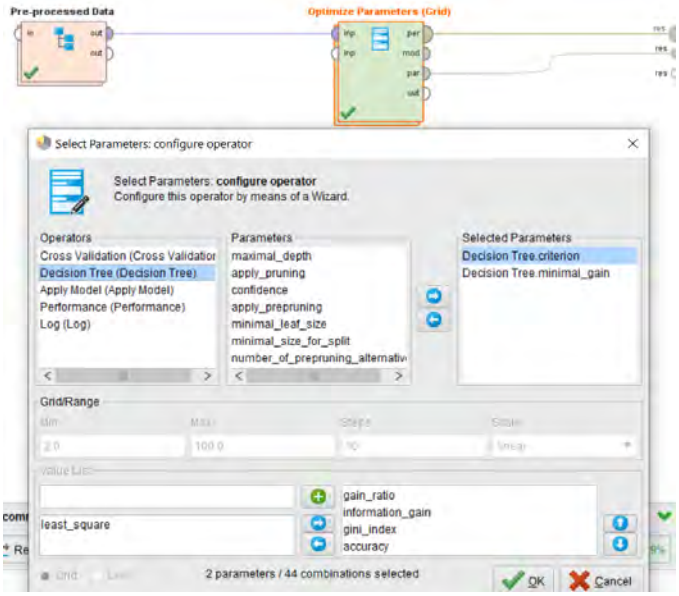


Fig. 15. Optimize Parameter (Grid) is used for optimizing the decision tree model

The Optimize Parameter (Grid) operator's complete setup has been executed through the edit parameter settings. It has been employed to optimize a specific parameter of the decision tree model. Parameters like Decision Tree criterion (gain ratio, information gain, gini index, accuracy) and Decision Tree minimal gain have been selected to determine the optimized conditions of these parameters. Decision Tree minimal gain controls how soon splits are being made so the lower the threshold value is the larger the tree grows. We have set the Decision tree minimal gain controls min and max range to 0.01 and 1 respectively with the step size of 100.

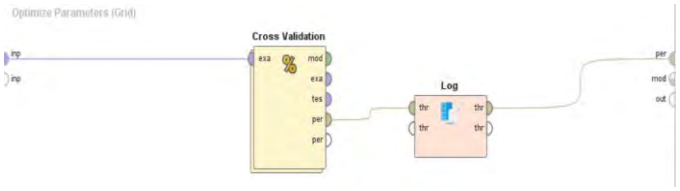


Fig. 16. Sub-process: Optimize Parameter (Grid)

Fig. 16 shows the sub-process of the optimize parameter (Grid) comprised of cross-validation (as shown in Fig. 17) and log operator.

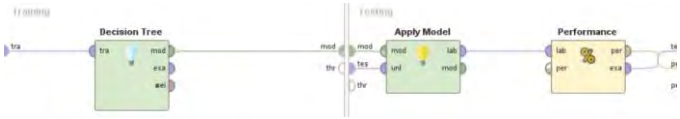


Fig. 17. Cross-validation Sub-process: Training and Testing

Comparison of performance vector analysis (i) without optimized and (ii) with optimized parameters have been performed to achieve better accuracy. From the analysis, it is seen that (as shown in Fig. 18) accuracy has been improved from 94.23% to 94.68%. Although the achieved accuracy is not a significant improvement, however, further optimization of parameters might help to increase the accuracy. We have already achieved a good accuracy (94.23%) from the unoptimized decision tree model thus further optimization will require more fine-tuning of parameters.

PerformanceVector

```
PerformanceVector:
accuracy: 94.23% +/- 0.35% (micro average: 94.23%)
ConfusionMatrix:
True:   No      Yes
No:     16092   866
Yes:    320     3266
precision: 91.08% +/- 0.90% (micro average: 91.08%) (positive class: Yes)
ConfusionMatrix:
True:   No      Yes
No:     16092   866
Yes:    320     3266
recall: 79.04% +/- 1.60% (micro average: 79.04%) (positive class: Yes)
ConfusionMatrix:
True:   No      Yes
No:     16092   866
Yes:    320     3266
AUC (optimistic): 0.981 +/- 0.004 (micro average: 0.981) (positive class: Yes)
AUC: 0.938 +/- 0.018 (micro average: 0.938) (positive class: Yes)
AUC (pessimistic): 0.896 +/- 0.037 (micro average: 0.896) (positive class: Yes)
```

(i) without optimized (accuracy: 94.23%)

PerformanceVector

```
PerformanceVector:
accuracy: 94.68% +/- 0.40% (micro average: 94.68%)
ConfusionMatrix:
True:   No      Yes
No:     16067   748
Yes:    345     3384
precision: 90.78% +/- 1.67% (micro average: 90.75%) (positive class: Yes)
ConfusionMatrix:
True:   No      Yes
No:     16067   748
Yes:    345     3384
recall: 81.90% +/- 1.26% (micro average: 81.90%) (positive class: Yes)
ConfusionMatrix:
True:   No      Yes
No:     16067   748
Yes:    345     3384
AUC (optimistic): 0.985 +/- 0.003 (micro average: 0.985) (positive class: Yes)
AUC: 0.970 +/- 0.016 (micro average: 0.970) (positive class: Yes)
AUC (pessimistic): 0.954 +/- 0.034 (micro average: 0.954) (positive class: Yes)
```

(ii) with optimized (accuracy: 94.68%)

Fig. 18. Performance vector analysis **(i)** without optimized and **(ii)** with optimized parameter

The 'log operator' has been implemented in our design to monitor the values computed during the optimization process. This is necessary to observe the values of various parameters in all iterations of any loop operator. Different information can be stored and viewed using the log operator [16]. We have considered the following parameters (as shown in Fig. 19) in the log operator intending to log the data based on performance.

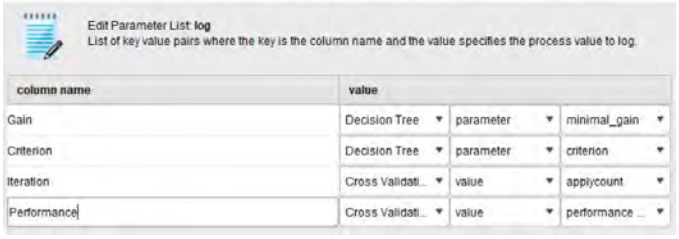


Fig. 19. Edit Log List Parameters

It is seen from the below log Table 3 that the information gain is the best split criterion in the decision tree model for this dataset because it has a higher performance percentage (94.7%). The top five results based on higher performance percentages have been shown in the below Table 3.

TABLE 3: LOG LIST TABLE DATA

Gain	Criterion	Iteration	Performance
0.010	information_gain	14	0.947
0.010	gini_index	21	0.945
0.020	information_gain	45	0.945
0.020	gini_index	52	0.945
0.505	accuracy	10	0.944

IV. CONCLUSION

Accurately forecasting floods is of paramount importance for Bangladesh, as the country is prone to frequent flooding due to its geographical location and topography. The impact of floods can be catastrophic, leading to loss of life and property, displacement of communities, and damage to infrastructure, agriculture, and the economy. Therefore, it is essential to develop and implement effective flood prediction models that can provide accurate and timely information to local authorities and communities. This will enable them to take necessary precautions and mitigate the adverse effects of flooding. Timely and precise flood forecasts can help save lives, reduce damage to property and infrastructure, and ensure sustainable development in flood-prone regions of Bangladesh. In this work, we have used the RapidMiner tool extensively to prepare the dataset, feature selection, generate decision tree, model validation, model selection and optimize the parameter of the model. Based on the ROC curve, we have seen that the Decision Tree has a higher AUC than the other classifier model. Thus, the decision tree classifier performs better in this data set. It is also observed from the optimization that the information gain is the best split criterion in the decision tree model for this dataset because it provides better performance. In our future work, we aim to use further sophisticated machine learning techniques for flood forecasting.

REFERENCES

[1] Sambare, Vidhya Vasantrao, and Arvind Jain. "The Application of Weather Forecast using Time Series Analysis." (2020).

[2] International Market research company "Gartner's report", available at https://rapidminer.com/news/rapidminer-named-a-leader-in-the-gartners-2019-magic-quadrant-for-data-science-and-machine-learning-platforms-for-sixth-consecutive-year/ accessed on March 8, 2022.

[3] KDnuggets annual software survey-2019, (accessed on March 8, 2022), available at https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

[4] Meghali A. Kalyankar, Prof. S. J. Alaspurkar, "Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering, February 2013, Volume3, Issue 2

[5] Saxena, Abhishek, Neeta Verma, and Krishna Chandra Tripathi. "A review study of weather forecasting using artificial neural network approach." International Journal of Engineering Research & Technology, vol. 2, no. 11, pp. 2029-2036, 2013.

[6] P. Hemalatha, "Implementation of data mining techniques for weather report guidance for ships using global positioning system" International Journal of Computational Engineering Research, vol. 3, no. 3, pp. 6-10, 2013.

[7] Shanmuganathan, Subana and P. Sallis, "Data mining methods to generate severe wind gust models," Atmosphere, vol. 5, no. 1, pp. 60-80, 2014.

[8] Mohammad Zeyad, Md Sajid Hossain. "A Comparative Analysis of Data Mining Methods for Weather Prediction", 2021 International Conference on Computational Performance Evaluation (ComPE), 2021

[9] A.R.W.M.M.S.C.B., Amarakoon, "Effectiveness of Using Data Mining for Predicting Climate Change in Sri Lanka", 2010.

[10] S. S. Badhiye, P. N. Chatur, and B. V. Wakode, "Temperature and humidity data analysis for future value prediction using clustering technique: an approach," International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 1, pp. 88-91, 2012.

[11] Dutta, Pinky Saikia and Hitesh Tahbilder, "Prediction of rainfall using data mining technique over Assam", IJCSE, Vol. 5, No. 2, 2014, pp. 85-90.

[12] Neha Khandelwal, Ruchi Davey, "Climatic Assessment of Rajasthan's Region for Drought With Concern Of Data Mining Techniques", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 5, September-October 2012, pp.1695-1697, 1695

[13] Ji, Soo-Yeon, Sharad Sharma, Byunggu Yu, and Dong Hyun Jeong. "Designing a rule-based hourly rainfall prediction model." In 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), pp. 303-308. IEEE, 2012.

[14] N. Gauhar. "Flood-prediction" github.com. https://github.com/n-gauhar/Flood-prediction (accessed on May 12, 2021)

[15] Noushin Gauhar, Sunanda Das, Khadiza, Sarwar Moury. "Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm", 2021 2nd international Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021

[16] Mierswa, I., & Klinkenberg, R., RapidMiner Studio (9.2) [Data science, machine learning, predictive analytics], 2018. Retrieved from rapidminer.com.

[17] RapidMiner user community suggestion, https://mod.rapidminer.com/ accessed on May 29, 2021

Md Sajid Hossain is currently serving as an Assistant Professor in the Department of Electrical and Electronic Engineering (EEE) at American International University-Bangladesh (AIUB). He obtained his B.Sc. Engg. (EEE) and M.Sc. Engg. (EEE) from American International University-Bangladesh, Dhaka, both with the 'Summa Cum Laude' academic distinction in 2012 and 2014 respectively. He also received a Master's degree in Smart Cities and Communities (MSc. SMACCs) from Spain, Greece, and the UK under the Erasmus Mundus Scholarship program by the European Union. With over 7 years of teaching experience in the EEE dept. at AIUB, Md Sajid Hossain is a proficient academic and professional. He is also a prolific researcher, having published several research articles in various journals and conferences. His primary research interest is focused on the UN Sustainable Development Goals (SDGs), specifically Goal No. 7: Affordable and Clean Energy. He believes that renewable energy technology is the key to addressing this goal and his research area mainly involves Renewable Energy, Grid Integration of Renewable Energy Sources (RES), Energy Efficiency, Sustainability, Energy Harvesting and Data Mining. He attended various seminars, workshops, and industrial tours in different places to broaden his knowledge and expertise. He is also a Graduate Member of IEEE USA and a Member of the Institution of Engineers Bangladesh (IEB), which reflects his commitment to professional development and lifelong learning.

Mohammad Zeyad is engaged as the Head of the 'Energy and Technology Research Division' at "Advanced Bioinformatics, Computational Biology and Data Science Laboratory Bangladesh (ABCD Laboratory, Bangladesh)". Although, he completed his Erasmus Mundus Joint MSc in Smart Cities and Communities (SMACCs) (2020-2022) from the Heriot-Watt University, Edinburgh, United Kingdom. Additionally, he completed his second Master's Degree in Research in Energy Efficiency and Sustainability in Industry, Transport, Building and Urbanism from the University of the Basque Country (UPV/EHU), Bilbao, Spain. He earned his B.Sc in Electrical and Electronic Engineering (EEE) at the American International University-Bangladesh (AIUB). Moreover, he has been working as an Editorial Board Member (Publication Committee Member) in IEEE Smart Cities since January 2021. Besides, he has received several invitations from the IEEE Flagship Conferences to join as a Session Chair, Program Committee, and Review Committee Member. Although, in 2019, he joined as an Advisory Panel Member in ELSEVIER and Convenor & Member (National Committee) of the International Electrotechnical Commission (IEC). He has 32 published papers in different scientific journals (ELSEVIER, Oxford University Press, Springer Nature, etc.) and conferences (IEEE flagship conferences) which are indexed by Web of Science, Scopus, and many more. Currently, almost 15+ national and international students and researchers are conducting their research under the supervision of Mohammad Zeyad in the ABCD Laboratory, Bangladesh. However, his primary research interest is in Smart Cities, Smart Grid, Building Energy systems, HVAC systems, Energy Efficiency, Optimization, Energy Economics, ML, AI, IoT, and Power Electronics.