Predicting Weather Forecasting State Based on Data Mining Classification Algorithms

Article *in* Asian Journal of Research in Computer Science · June 2021

DOI: 10.9734/AJRCOS/2021/v9i330222

3 authors:

Fairoz Q Kareem
Duhok Polytechnic University
5 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Adnan Mohsin Abdulazeez
Duhok Polytechnic University
203 PUBLICATIONS **5,034** CITATIONS
SEE PROFILE

Dathar Abas Hasan
Duhok Polytechnic University
22 PUBLICATIONS **415** CITATIONS
SEE PROFILE

Asian Journal of Research in Computer Science

9(3): 13-24, 2021; Article no.AJRCOS.68636
ISSN: 2581-8260

Predicting Weather Forecasting State Based on Data Mining Classification Algorithms

Fairoz Q. Kareem[1*], Adnan Mohsin Abdulazeez[2] and Dathar A. Hasan[3]

[1]Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.
[2]Research Center of Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.
[3]Shekhan Technical Institute, Duhok Polytechnic University,Duhok – Kurdistan Region, Iraq.

Authors' contributions

This work was carried out in collaboration among all authors. Author FQK designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript. Author AMA managed the analyses of the study. Author DAH managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v9i330222
Editor(s):
(1) Dr. Xiao-Guang Lyu, Huaihai Institute of Technology, P. R. China.
Reviewers:
(1) Ananthi Sheshasaayee, Quaid-E-Millath Government College for Women (Autonomous), India.
(2) Paulus Agus Winarso, Indonesia.
Complete Peer review History: http://www.sdiarticle4.com/review-history/68636

Original Research Article

Received 18 March 2021
Accepted 28 May 2021
Published 05 June 2021

ABSTRACT

Weather forecasting is the process of predicting the status of the atmosphere for certain regions or locations by utilizing recent technology. Thousands of years ago, humans tried to foretell the weather state in some civilizations by studying the science of stars and astronomy. Realizing the weather conditions has a direct impact on many fields, such as commercial, agricultural, airlines, etc. With the recent development in technology, especially in the DM and machine learning techniques, many researchers proposed weather forecasting prediction systems based on data mining classification techniques. In this paper, we utilized neural networks, Naïve Bayes, random forest, and K-nearest neighbor algorithms to build weather forecasting prediction models. These models classify the unseen data instances to multiple class rain, fog, partly-cloudy day, clear-day and cloudy. These model performance for each algorithm has been trained and tested using synoptic data from the Kaggle website. This dataset contains (1796) instances and (8) attributes in our possession. Comparing with other algorithms, the Random forest algorithm achieved the best performance accuracy of 89%. These results indicate the ability of data mining classification algorithms to present optimal tools to predict weather forecasting.

**Corresponding author: E-mail: Fairoz.kareem@dpu.edu.krd;*

Keywords: Weather forecasting; DM; random forest; naïve bayes; K-nearest neighbor; neural networks.

1. INTRODUCTION

Nowadays, technology has developed a lot, especially in the field of Machine Learning (ML), which is useful for reducing human work. In the field of artificial intelligence, ML integrates statistics and computer science to build algorithms that get more efficient when they are subject to relevant data rather than being given specific instructions [1,2]. Data Mining (DM) has developed and turned out to be a powerful and strong tool because extracting beneficial records out of tons of engineering and commercial data, the use of multi strategies to analyze data certain as much classification and clustering [3,4]. Weather prediction is a challenge in meteorology that has been a major subject of meteorological research [5]. Research on weather prediction has been done by various methods which each method has deficiencies and advantages [6]. The approach in weather prediction can be done by an empirical or dynamic method. Short-term weather predictions have been using dynamic methods which are an analytical approach based on the principles of fluid dynamics, while empirical methods performed with statistical and mathematical approaches are more widely used for long-term weather predictions. Both approaches have their flaws and advantages [7-9]. The use of empirical methods in BMKG for short-term weather prediction is not much done yet. Related to this, the researchers are interested to examine more about how the use of empirical methods, especially DM techniques for short-term weather prediction [10-12].

Knowledge discovery from databases (KDD) or databases is known as DM (DM), and it focuses on data that hasn't already been discovered and insight that might be obtained from data [13-15]. In comparison to most statistical approaches, DM, intriguing phenomena can be found by data tasks, such as anomaly detection, association, correlation, and classification. The creative retrieval of inferred, previously hidden and theoretically valuable data from datasets is referred to as data mining. data-mining can perform various functions: certain functions analyze the data and identify a model that works with the data [16,17]. Technologies of DM embrace idea discovery, explanation, classification, forecast, pattern detection, relation detection, and separation, and distance computation [18]. Researchers in the field of meteorology has studied how to get an accurate prediction method with DM techniques to build a weather prediction model using DM [19].

DM can't automatically generate valuable information from a large dataset by itself. Taking time to assemble a productive data set, evaluating the appropriateness of the data, and deciding on the best approach to analyzing it is essential [20,21]. Predictive analytics uses datasets. Predictive models disclose previously hidden trends based on the convergence of various datasets, giving you an accurate picture in the future [22,23]. The classification is one of the predictive DM tasks and used to find a model that describes and distinguishes classes or concepts. The weather parameters discussed in this study are temp min, temp max, summery, desc, cloud cover, visibility, humidity, wind speed, using Naïve Bayes, KNN, Random Forest and Neural Network algorithm with a good result in predicting the weather.

In this research, our four algorithms (KNN, Neural Network, Random Forest, and Naïve Bayes) were used to analyze meteorological data collected from the Kaggle website for developing classification Rules for a weather parameter through the study interval and for forecasting the weather conditions in the future. The objective of forecasting weather is those weather changes that affect our daily life such as changes in maximum and minimum temperature, wind speed, humidity and rainfall.

2. LITERATURE REVIEW

In recent years, various researchers have used DM techniques in meteorology and weather forecasts. The following is a study of the usage of various DM techniques in the field of weather prediction classification analysis over the last few years.

In [24] they proposed that the DM solution which applies the Naive Bayes algorithm is the basis for the current weather forecast and the C45 prediction method. when guessing a situation Although the results of comparisons between the Naïve Bayes, K-Nearest Neighbor, and C45 classifications of weather forecasting have shown that KNN classification is the most accurate, Naïve Bayes got 68.77% in the calculations.

In [25], the authors improve short-term weather forecasting for the model to which weather data are trained in the northwestern part of Bangladesh, this region, this research studies expands on the use of machine learning. Extreme machine learning was used to aid meteorologists in their weather predictions. These seven local weather stations located in the length of the northwestern part of Bangladesh were used to represent 30 years of temperature, wind, and humidity to simulate these three variables for this study (BMD). The output of the Extreme Learning Model (ELM) is much greater than that of an artificial neural network with an accuracy of 95%.

In [26] they suggest a method based on historical weather parameters to forecast the frequency and non-controversy of rainfall in La Trinidad, Benguet. The predictive model for the weather dataset has been developed with five machine learning grade algorithms: Fine Decision Tree, Linear Discriminant, K-Nearest Neighbor Course, Gaussian Support Vector Machines and Neural Networks. The findings of the 5 models show that the K-Nearest Neighbor course delivers the best value in all test measurements. KNN, the best predictor for the prediction of rainfall at La Trinidad, Benguet with a reasonable precision of 81.1%. KNN model assessment course shows that the classification of machine learning is possible to forecast rainfalls and non-incidents.

In [27] They use LSTM as a basis for the implementation of weather forecasting to achieve a data-driven prediction model. Furthermore, Transudative LSTM (T-LSTM), which uses local knowledge to forecast time series, offers experiments over two separate timeframes of one year. The findings show that T-LSTM improves better prediction accuracy.

In [28] There have been experiments and comparisons between DM technologies, including Naive Bayes (NB), KNN and Decision Trees, for predicting various forms of air dust. Cairo Airport study data were gathered; pressure, temperatures, humidity, wind and wind velocity and direction were all the analyzed variables within the data. The data has been processed in an open software portal for data scientific research, machine learning, in-depth learning, text mining and predictive analysis. 23 different models were evaluated with the confusion matrix, correlation and root-mean-square error. The results showed the decision tree to define and model data more successfully, followed by a classifier theorem from Bayes.

In [13] the Decision Tree, K-NN and Random Forest algorithm estimates and the best accuracy outcome of those three algorithms suggest a straightforward solution to future years' weather predictions by using the previous data analyses. In everyday uses, weather prediction plays a major role and is performed based on temperature variations in some areas. The mean values, median, confidence values, likelihood and the variance between the plots of all three algorithm values are calculated from all of these algorithms. In this job, they will determine whether the temperature goes up or down with the use of these equations, whether or not it is a rainy day. The results suggest that they are best accurate using the Random Forest algorithm.

In [29] they used techniques such as Deep Neural Network modelling with optimization to forecast rainfall. an examination of results After data preprocessing and function extraction, model parameter optimization was performed to measure the performance of the algorithms. The Adam optimizer was used in the optimization, which demonstrated that deep neural networks outperform machine learning algorithms for weather prediction.

In [30] Their research work is carried out using a recurrent neural network(RNN) with LSTM technique, the model suggested for a weather forecasting scheme. The data was trained with the LSTM algorithm in the model. The experimental findings show that the neural network Long-Short Term memory produces significant results with high precision amongst other weather forecasting techniques.

In [31] The authors used three algorithms to forecast rain: Naive Bayes, K-Nearest Neighbors, and Classification Tree, along with validation parameters such as the Confusion Matrix, ROC curves, and Brier Score. The input data collection consists of synoptic data from Kemayoran Meteorological Station, Jakarta (96745) over ten years (2006 - 2015), and it contains 3528 datasets and eight attributes. Following a sequence of data analysis, collection, and model testing, it was determined that the Naive Bayes Algorithm has the highest accuracy rate of 77.1% in the category of equal classification, indicating that it has considerable potential for practical use.

In [32] aims to solve weather problems by developing a rain model based on more complete and plentiful rainfall data in Indonesia. The study compares and contrasts four DM techniques: C4.5/J48, Random Forest (RF), Naive Bayes (NB), and Multilayer Perceptron (MLP). The experimental findings demonstrated that the MLP and J48 algorithms are capable of providing the highest level of precision (up to 78,4%).

3. METHODOLOGY

In this section, we focused on building, training and testing the proposed weather forecasting prediction system based on the four DM classification techniques. In addition, we presented a clear explanation about the utilized dataset and the tools that are used for model building.

3.1 Dataset

In this research, we used a dataset for weather forecasting which is taken from the Kaggle website, because it supplies many different datasets so, it is a common source for many datasets. In our research the weather forecasting dataset consists of 1796 objects and 8 variables for prediction, the dataset features are:

- temp min
- temp max
- summery
- desc
- cloud cover
- visibility
- humidity
- wind speed

The (desc) variable is the target value for the dataset which consists of values such as (fog, partly-cloudy-day, rain, clear-day and cloudy) without gaps in the results. So, the dataset contains 769 rain records, 566 partly-cloudy-day records, 210 clear-day records and 68 cloudy records for weather prediction as Table 1.

3.2 The Proposed Algorithms

3.2.1 Naïve Bayes

Naïve Bayes is a wonderful general-learning algorithm for all fields of machine-learning and data analysis. the Naïve Bayes' presumption of attribute freedom is bad for business (no attribute linkage). In most cases, we assume that our attributes are independent of each other, but in a few, the output of the Naïve Bayes classification is higher than others indicate [33,34]. The prediction of Naïve Bayes is based on the Bayes theorem with the following classification format [35,36]:

$$\boldsymbol{p(y|x)} = \frac{\boldsymbol{p(y)\Pi_{i=1}^{q} p(\chi_i|y)}}{\boldsymbol{p(x)}} \quad \textbf{(1)}$$

While Naïve Bayes with continuous features has a formula:

$$\boldsymbol{p(y|x)} = \frac{\boldsymbol{1}}{\boldsymbol{\sqrt{2\pi\sigma}}}\boldsymbol{exp}\,\frac{\boldsymbol{-(x-\mu)^2}}{\boldsymbol{2\sigma^2}} \quad \textbf{(2)}$$

Where $P\ (Y \mid X)$ refers to the data probability with X vector on Y class and (Y) is the Initial probability for Y Class, $\boldsymbol{\Pi_{i=1}^{q}\, p(\chi_i|y)}$ is the Independent probability of Y class from all features from X vector, μ is the mean value of the attribute with continue features and σ is the standard deviation.

3.2.2 K-Nearest Neighbor (KNN)

KNN method is a Machine Learning algorithm that is considered as a simple method to be applied in data analysis with many dimensions [37,38]. Although this method is simple, this method has advantages compared to other methods, which can generalize a relatively small set of training data [39-41]. The k-nearest neighbour (KNN) method is a classifying approach that is nearest to the object based on learning results. Learning data is projected into a multi-dimensional space in which each dimension depicts data characteristics [42]. This room is split into parts depending on the study data graduation. One point in this area is labelled as class c when class c is the most commonly used in the nearest k of the dot. Near or distant neighbours are normally measured based on the following equations on Euclidean distances [43]:

$$\boldsymbol{Untup} = \boldsymbol{(p_1, p_2, \ldots\ldots, p_n)}\,\boldsymbol{dan}\ \boldsymbol{Q} = \boldsymbol{(q_1, q_2, \ldots\ldots, q_n)} \quad \textbf{(3)}$$

$$\boldsymbol{Jarak} = \sqrt{\boldsymbol{(p_1 - q_2)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}} \quad \textbf{(4)}$$

$$\boldsymbol{Jarak} = \sqrt{\boldsymbol{\sum_{i=1}^{n}(p_i - q_i)^2}} \quad \textbf{(5)}$$

3.2.3 Neural Network

A neural network is a system that models the way the brain functions Neural Networks are equipped to contribute to a given goal outcome by a specific input [44]. The network is balanced according to a comparison of the output and the aim before the output is near the goal. Different

types of networking architecture are available: single layer transmission networks, several layer transmissions networks, recurring networks etc. ANN is an alternative method to efficiently predict time series. [45,46]. For modelling and predicting complicated time series, ANN may be used for researchers in many time series applications, such as seasonal prevision [18], Weather forecasting, prediction of electric demand, air emissions, etc. In precipitation modelling, artificial neural networks were used for various network configurations [18,47]. The network was educated using the algorithm for the propagation of errors in various hidden layers. Exhaustive simulation determined optimal parameters [48,49]. The work has been monitored to provide the optimal response to the network. The input data were introduced to the network and then adjusted to change the neurons' weights to predict with desired precision the next point in the input data. This is done using a preview of the input data to train the network. The qualified network has been used to forecast a point in the test series. After a variety of simulations, learning rate and momentum have been set [50,51].

3.2.4 Random forest

All the decision trees (decision graphs) trees are combined into a single model that determines the prediction value and has its results divided into a few distinct trees by applying the join function of all of the individual trees [52-54]. This algorithm eliminates the data's overfitting problem and trains the data quickly using test data. individual decision trees were used to generate the bootstrap samples. As a consequence, we use the random forest function algorithm to generate results from the dataset.

3.3 Orange Tools

Nowadays, data science made it easy to test and evaluate models using an open-source machine learning software called Orange [31]. So, Orange is a component-based DM software. It includes a range of data visualization, exploration, preprocessing and modelling techniques. It can be used through a nice and intuitive user interface or, for more advanced users, as a module for the Python programming language.

In this study, we used two evaluation tools (Confusion Matrix and Roc Analysis) as in Fig.1 and we will discuss the two evaluation tools in the next section.

4. RESULTS AND DISCUSSION

In general, the process of making a model using the Random forest, Naïve Bayes, k-Nearest Neighbor and Neural Network algorithms looks like in Fig. 1.

When we process synoptic weather data using four classifiers, Random Forest, Naïve Bayes, Neural Network and k-NN, the resulting surface analysis (depth) is supposed the most accurate. Any input data will evaluate the performance model of each approach to determine the model's dependability Furthermore, the test results are compared to determine which algorithm has the highest accuracy, allowing the best algorithm to be calculated.

4.1 Confusion Matrix

The confusion matrix test aims to determine the precision, recall, accuracy and et of the test results. The 10-fold Cross-Validation approach was used to calculate the precision and Area Under Curve (AUC) of the determination. The following are the outcomes of each algorithm's tests:

The result of the Random Forest algorithm in the Table 1 Train time = 59%, AUC = 97%, CA = 89%, F1 = 89%, Precision = 89%, Recall = 89% and Specificity = 94%. Based on the results in Table 2 and Table 3, it can be seen that the level of accuracy using the Random Forest is 89% that has the best accuracy than other algorithms with the number of rain prediction 1145 dataset from the total amount of data tested that is 1796 datasets.

The second algorithm Neural Network in the Table 2 Train time = 190%, AUC = 49%, CA = 21%, F1 = 15%, Precision = 13%, Recall = 21% and Specificity = 78%. Based on the results in Table 2 and Table 3, it can be seen that the level of accuracy using the Neural Network is 21% that has the lowest accuracy than other algorithms with the number of rain prediction 520 dataset from the total amount of data tested that is 1796 datasets. Also, where the results of Naïve Bayes showed in the Table 2 Train time = 10%, AUC = 91%, CA = 60%, F1 = 70%, Precision = 86%, Recall = 60% and Specificity = 96%. Based on the results in Table 2 and Table 3, it can be seen that the level of accuracy using the Naïve Bayes is 60% and the level of training time is 10% so it takes less time for the training model than another algorithm with the number of rain prediction 650 dataset from the total amount of data tested that is 1796 datasets.

Table 1. Dataset instances samples

No	Temp Min	Temp Max	Summary	Desc	Cloud Cover	Humidity	Wind Speed	Visibility
1	7.53	12.23	Possible light rain until evening.	rain	0.8	0.89	14.69	4.43
2	3.58	7.15	Possible light rain throughout the day.	rain	0.62	0.79	15.04	5.64
3	-0.61	6.54	Clear throughout the day.	rain	0.31	0.84	4.48	6.2
4	-0.63	7.59	Mostly cloudy throughout the day.	partly-cloudy-day	0.78	0.85	4.35	6.22
5	6.51	10.43	Overcast throughout the day.	partly-cloudy-day	0.85	0.91	6.2	5.91
6	3.66	9.95	Possible light rain in the morning and afternoon.	rain	0.63	0.88	8.7	5.93
7	0.88	5.19	Partly cloudy throughout the day.	rain	0.68	0.83	11.19	6.22
8	1.39	7.52	Possible light rain starting in the afternoon.	rain	0.61	0.85	12.23	5.48
9	4.17	7.85	Possible light rain in the afternoon and evening.	rain	0.78	0.78	22.37	5.75
10	0.68	4.11	Foggy in the afternoon.	partly-cloudy-day	0.52	0.82	9.64	5.79
11	-2.49	2.43	Possible drizzle in the morning and afternoon.	rain	0.53	0.92	3.59	5.84
12	-2.45	3.9	Clear throughout the day.	partly-cloudy-day	0.37	0.81	4.13	5.7
13	-4.74	1.04	Clear throughout the day.	clear-day	0.13	0.85	2.13	6.22
14	0.11	3.95	Foggy until morning, starting again in the evening.	fog	0.88	0.93	1.14	1.99
15	-2.88	1.58	Mostly cloudy throughout the day.	partly-cloudy-day	0.72	0.87	1.62	4.08

The last results of KNN algorithms showed in the Table 2 Train time = 56%, AUC = 68%, CA = 53%, F1 = 53%, Precision = 52%, Recall = 53% and Specificity = 73%. Based on the results in Table 2 and Table 3, it can be seen that the level of accuracy using the KNN is 53% with the number of rain prediction 814 dataset from the total amount of data tested that is 1796 datasets.

The test results discover the precision, recall, accuracy, and AUC values for each experiment. The precision value of all tests is highest in the Random Forest algorithm which is 89%, where the lowest precision value is in the Neural Network algorithm which is 13%. The Recall value is the highest of all tests is in the Random Forest algorithm which is 21%, where the lowest Recall value is in the k-NN algorithm which is 72.3%. The highest F1 is in the Random Forest which is 89%. While the lowest F1 is in the Neural Network which is 15%.

Classifier accuracy metrics, or classification accuracy values, are the percentage of data records correctly categorized by an algorithm after performing a test on the classification results. Accuracy can also be characterized as the degree to which the predicted value is similar to the actual value. Where the highest or best accuracy is in the Random Forest this indicates that the algorithm is very good performance for predicting the weather. followed by the Naïve Bayes at 86% test results and the last algorithm KNN at test results 53%. The four algorithms have an equation that is accurate enough to be used for the prediction of weather.

The objective of the Precision is used to compare (True Positive (TP) and False Positive (FP)) entities. It can be measured in the following manner:

$$\textbf{Precision} = \frac{TP}{(TP+FP)} \quad (6)$$

(TP) is used for entities that are correctly categorized, while (FP) is used for entities that are incorrectly classified.
The recall is used to compare True Positive entities to False Negative entities (FN) that are not labelled at all. It can be measured in the following manner:

$$\textbf{Recall} = \frac{TP}{(TP+FN)} \quad (7)$$

There may come a point where output estimation with recall and precision is no longer possible; for example, if one DM method has a higher Precision but a lower Recall than another, the issue of which algorithm is better emerges. The solution to this problem is to use the F-measure, which takes the precision and recall values and averages them. The F-measure can be calculated as follows:

$$\textbf{F-measure} = \frac{Precision * Recall * 2}{(Precision+Recall)} \quad (8)$$

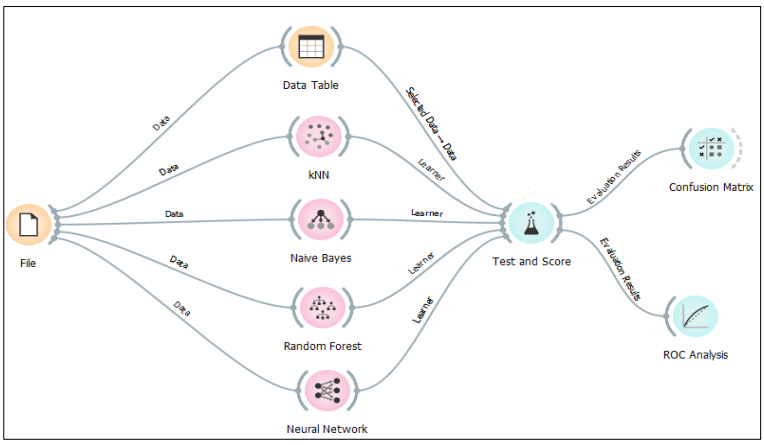
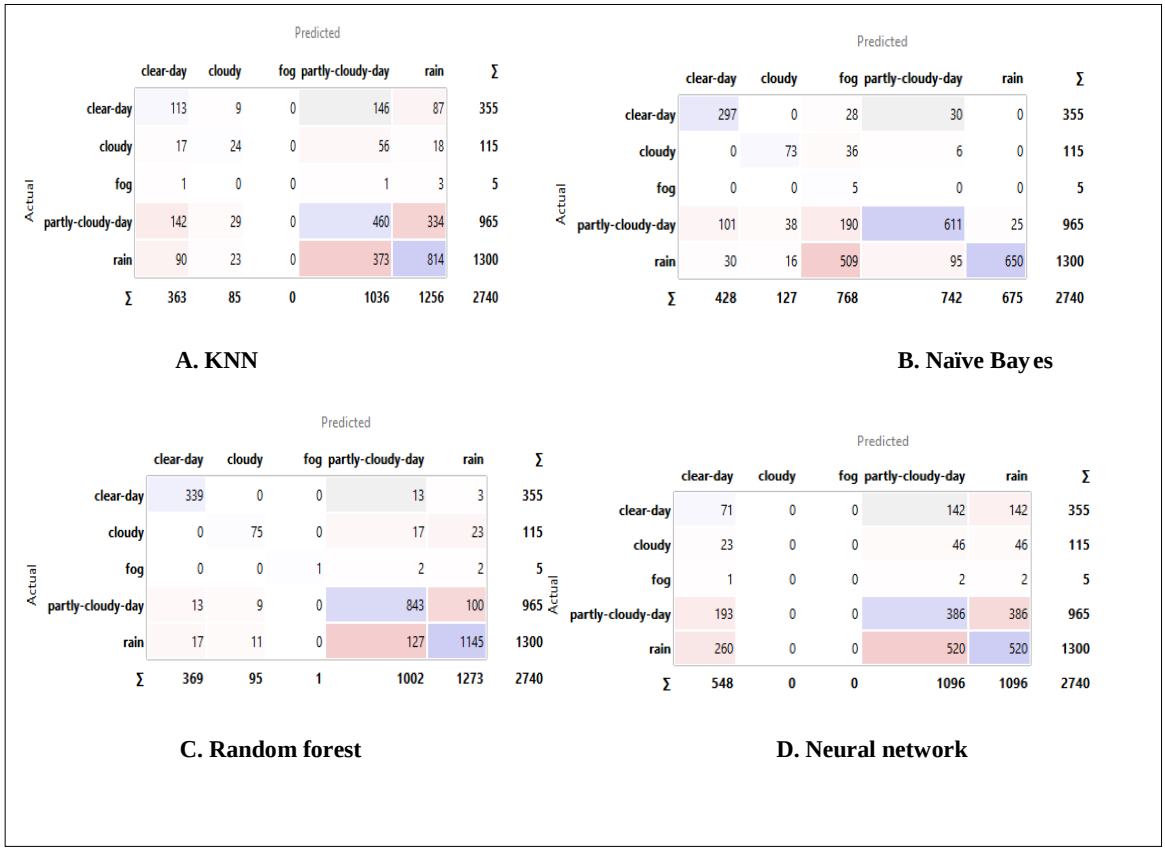


Fig. 1. The classification process of DM software (Orange Ver 3.28.0)

Table 2. Classifiers Weighted Average Detailed Accuracy

Model	Train time[s]	AUC	CA	F1	Precision	Recall	Specificity
Random Forest	59%	97%	89%	89%	89%	89%	94%
Neural Network	190%	49%	21%	15%	13%	21%	78%
Naive Bayes	10%	91%	60%	70%	86%	60%	96%
KNN	56%	68%	53%	53%	52%	53%	73%

Table 3. Confusion Matrix of Five Algorithms



Predicted

Actual	clear-day	cloudy	fog	partly-cloudy-day	rain	Σ
clear-day	113	9	0	146	87	355
cloudy	17	24	0	56	18	115
fog	1	0	0	1	3	5
partly-cloudy-day	142	29	0	460	334	965
rain	90	23	0	373	814	1300
Σ	363	85	0	1036	1256	2740

A. KNN

Predicted

Actual	clear-day	cloudy	fog	partly-cloudy-day	rain	Σ
clear-day	297	0	28	30	0	355
cloudy	0	73	36	6	0	115
fog	0	0	5	0	0	5
partly-cloudy-day	101	38	190	611	25	965
rain	30	16	509	95	650	1300
Σ	428	127	768	742	675	2740

B. Naïve Bayes

Predicted

Actual	clear-day	cloudy	fog	partly-cloudy-day	rain	Σ
clear-day	339	0	0	13	3	355
cloudy	0	75	0	17	23	115
fog	0	0	1	2	2	5
partly-cloudy-day	13	9	0	843	100	965
rain	17	11	0	127	1145	1300
Σ	369	95	1	1002	1273	2740

C. Random forest

Predicted

Actual	clear-day	cloudy	fog	partly-cloudy-day	rain	Σ
clear-day	71	0	0	142	142	355
cloudy	23	0	0	46	46	115
fog	1	0	0	2	2	5
partly-cloudy-day	193	0	0	386	386	965
rain	260	0	0	520	520	1300
Σ	548	0	0	1096	1096	2740

D. Neural network

In the case of the AUC (Area under the ROC Curve), all potential classification levels are considered. AUC can be seen in many ways. One way is to think about the likelihood that the model would score a random positive example higher than a random negative example. Where Specificity (True Negative Rate) quantifies the proportion of accurately defined negatives (i.e., the proportion of those that may not have the disease (unaffected) that are correctly identified as not having the Accuracy. The AC of a classifier is expressed as a percentage of accurate predictions divided by the total number of instances. In other words, if the classifier is given a reasonable mark for a given level of precision, it will classify all of the data that does not yet have a corresponding name.

4.2 ROC Analysis

The ROC curve shows precision and visually compares the rating. ROC exhibits a matrix of uncertainty. ROC is a 2D graph with true positives as vertical lines and false positives as horizontal lines. Another method to determine the discriminant strength of the four algorithms is by calculating the ROC Curve. Fig. 2 displays the ROC curve study results.

Based on Fig. 2 it is known that the curve has a shape that tends towards the Y line where the whole curve is above the dashed line. The closer the ROC curve is to the Y line (0.1), the better the model predicts the weather. To make sure the analysis can be seen from the AUC table. The value of AUC (area under the curve) will usually be at 0.5 and 1. If the value of the AUC approximation 1, the model is more accurate. The highest result of AUC value is in the Random Forest algorithm with a value of 97%, Naïve Bayes 91%, k-NN 68% and Neural Network with the lowest value of 49%. The (AUC) ROC curve for the four algorithms is shown in Table 2.

The test results showed for the area under the curve (AUC) value of the Naïve Bayes, k-NN, Neural Network and Random Forest algorithms can be seen in Table 2 with a fair classification level which means that the accuracy of the model is good enough so that these four algorithms can be used for weather forecasting.

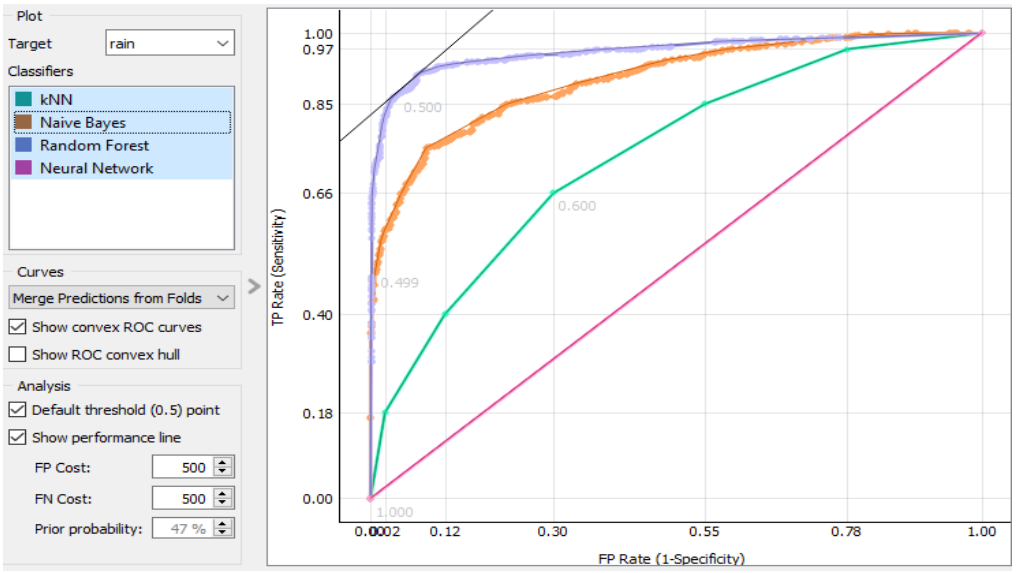


Fig. 2. Roc Analysis of (KNN, Naïve Bayes, Random Forest, Neural Network)

5. CONCLUSION

The following conclusions can be drawn from research on short-term weather prediction using a classification algorithm for rain prediction based on probabilistic supervised learning with Confusion Matrix, ROC curve analysis and Receiver Operating Characteristic test parameters:

- The four algorithms can be applied to weather data with a good category, based on the test results of the two parameters namely Confusion Matrix and ROC curve.
- The classification algorithms Comparison that is Naïve Bayes, k-NN, Random Forest and Neural Network test results show that the Random Forest has the best predictive probability for the weather, that its values are precision 89%, recall 89%, accuracy 89%, AUC (area under the curve) 97% and Training Time 59%. So, the Random Forest algorithm is quite the potential to be used practically.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Zeebaree DQ, Haron H, Abdulazeez AM. Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE). 2018;145-150: IEEE.
2. Zeebaree DQ, Haron H, Abdulazeez AM, Zebari DA. Trainable model based on new uniform LBP feature to identify the risk of the breast cancer. In 2019 International Conference on Advanced Science and Engineering(ICOASE).2019;106-111:IEEE.
3. Abdulqader DM, Abdulazeez AM, Zeebaree DQ. Machine Learning Supervised Algorithms of Gene Selection: A Review. Machine Learning. 2020;62(03).
4. Abdulkareem NM, Abdulazeez AM. Machine Learning Classification Based on Radom Forest Algorithm: A Review. International Journal of Science and Business. 2021;5(2):128-142, 2021.
5. Abdulkareem NM, Mohsin Abdulazeez A, Qader Zeebaree D, Hasan DA. COVID-19 World Vaccination Progress Using MachineLearning Classification Algorithms. Qubahan Academic Journal. 2021;1(2): 100-105.
6. Jamal S, Bappy TH, Pervin R, Rabby ASA. Weather Status Prediction of Dhaka City Using Machine Learning. In Computational Methods and Data Engineering: Springer. 2021;293-304.
7. Chauhan D, Thakur J. Data mining techniques for weather prediction: A review. International Journal on Recent and Innovation Trends in Computing and Communication. 2014;2(8):2184-2189.

8. Saeed J, Abdulazeez AM. Facial Beauty Prediction and Analysis Based on Deep Convolutional Neural Network: A Review. Journal of Soft Computing and Data Mining. 2021;2(1):1-12.
9. Kunjumon C, Nair SS, Suresh P, Preetha S. Survey on weather forecasting using data mining," in 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), 2018;262-264:IEEE.
10. Abhishek K, Kumar A, Ranjan R, Kumar S. A rainfall prediction model using artificial neural network," in 2012 IEEE Control and System Graduate Research Colloquium. 2012;82-87:IEEE.
11. Olaiya F, Adeyemo AB. Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business. 2012;4(1):51.
12. Abdulqadir HR, Abdulazeez AM. Reinforcement Learning and Modeling Techniques: A Review. International Journal of Science and Business. 2021;5(3):174-189.
13. Pavuluri BL, Vejendla RS, Jithendra P, Deepika T, Bano S. Forecasting Meteorological Analysis using Machine Learning Algorithms. In 2020 International Conference on Smart Electronics and Communication (ICOSEC). 2020;456-461:IEEE.
14. Rushing J, Ramachandran R, Nair U, Graves S, Welch R, Lin H. ADaM: A data mining toolkit for scientists and engineers. Computers & geosciences. 2005;31(5):607-618.
15. Yahia HS, Abdulazeez AM. Medical Text Classification Based on Convolutional Neural Network: A Review. International Journal of Science and Business. 2021;5(3):27-41,.
16. Dhore A, Byakude B, Sonar B, Waste M. Weather prediction using the data mining Techniques. Int. Res. J. Eng. Technol. (IRJET). 2017;4(5):2562-2565.
17. Mahmood MR, Abdulazeez AM, Orman Z. A new hand gesture recognition system using artificial neural network.
18. Jahnavi Y. Analysis of weather data using various regression algorithms. International Journal of Data Science. 2019;4(2):117-141.
19. Abas Hasan D, Mohsin Abdulazeez A. A modified convolutional neural networks model for medical image segmentation. Test Engineering and Management. 2020;83:16798-16808.
20. Sulaiman DM, Abdulazeez AM, Haron H, Sadiq SS. Unsupervised Learning Approach-Based New Optimization K-Means Clustering for Finger Vein Image Localization. In 2019 International Conference on Advanced Science and Engineering (ICOASE). 2019;82-87:IEEE.
21. Alankar B, Yousf N, Ahsaan SU. Predictive Analytics for Weather Forecasting using Back Propagation and Resilient Back Propagation Neural Networks. In New Paradigm in Decision Science and Management: Springer. 2020;99-115.
22. Sondwale PP. Overview of predictive and descriptive data mining techniques. International Journal of Advanced Research in Computer Science and Software Engineering. 2015;5(4):262-265.
23. Khorshid SF, Abdulazeez AM. breast cancer diagnosis based on k-nearest neighbors: A review. Palarch's Journal of Archaeology of Egypt/Egyptology.2021;18(4):1927-1951.
24. Findawati Y, Astutik II, Fitroni A, Indrawati I, Yuniasih N. Comparative analysis of Naïve Bayes, K Nearest Neighbor and C. 45 method in weather forecast. In Journal of Physics: Conference Series. 2019;1402(6):066046:IOP Publishing.
25. Rizvee MA, Arju AR, Al-Hasan M, Tareque SM, Hasan MZ. Weather Forecasting for the North-Western region of Bangladesh: A Machine Learning Approach. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020;1-6:IEEE.
26. Macabiog REN, Cruz JCD. Rainfall Predictive Approach for La Trinidad, Benguet using Machine Learning Classification. In 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). 2019;1-6:IEEE.
27. Karevan Z, Suykens JA. Transductive LSTM for time-series prediction: An application to weather forecasting. Neural Networks. 2020;125:1-9.
28. Ali M, Asklany SA, El-wahab M, Hassan M. Data Mining Algorithms for Weather Forecast Phenomena Comparative Study. International journal of computer science and network security. 2019;19(9):76-81.

29. Naik AR, Deorankar A, Ambhore PB. Rainfall Prediction based on Deep Neural Network: A Review. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). 2020;98-101:IEEE.
30. Fente DN, Singh DK. Weather forecasting using artificial neural network," in 2018 second international conference on inventive communication and computational technologies (ICICCT), 2018;1757-1761:IEEE.
31. Prasetya R. Data mining application on weather prediction using classification tree, naïve bayes and K-nearest neighbor algorithm with model testing of supervised learning probabilistic brier score, confusion matrix and ROC. JAICT. 2020;4(2):25-33.
32. Anwar MT, Hadikurniawati W, Winarno E, Widiyatmoko W. Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia. In 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020;83-88:IEEE.
33. Shadiq MA. Keoptimalan Naïve Bayes Dalam Klasifikasi. Bandung: Jurnal Universitas Pendidikan Indonesia; 2009.
34. Kareem FQ, Abdulazeez AM. Ultrasound Medical Images Classification Based on Deep Learning Algorithms: A Review.
35. Safri YF, Arifudin R, Muslim MA. K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. Sci. J. Informatics. 2018;5(1):18.
36. Najat N, Abdulazeez AM. Gene clustering with partition around mediods algorithm based on weighted and normalized Mahalanobis distance. In 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS).2017;140-145:IEEE.
37. Alkhatib K, Najadat H, Hmeidi I, Shatnawi MKA. Stock price prediction using k-nearest neighbor (kNN) algorithm. International Journal of Business, Humanities and Technology. 2013;3(3):32-44.
38. Ibrahim I, Abdulazeez A. The Role of Machine Learning Algorithms for Diagnosing Diseases," Journal of Applied Science and Technology Trends. 2021;2(01):10-19.
39. Danil M, Efendi S, Sembiring RW. The Analysis of Attribution Reduction of K-Nearest Neighbor (KNN) Algorithm by Using Chi-Square. In Journal of Physics: Conference Series. 2019;1424(1): 012004:IOP Publishing.
40. Hasan BMS, Abdulazeez AM. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining. 2021;2(1):20-30.
41. Salim NO, Abdulazeez AM. Human Diseases Detection Based On Machine Learning Algorithms: A Review. International Journal of Science and Business. 2021;5(2):102-113.
42. Rashid Ismael H, Mohsin Abdulazeez A, Hasan DA. Comparative Study for Classification Algorithms Performance in Crop Yields Prediction Systems. Qubahan Academic Journal. 2021;1(2):19-24.
43. Hapsari RI, Sugan BAI, Novianto D, Asmara RA, Oishi S. Predictability of Naïve Bayes classifier for lahar hazard mapping by weather radar. In IOP Conference Series: Earth and Environmental Science. 2020;437(1):012049:IOP Publishing.
44. Ahmad A-M, Chuan C-S, Fatimah M. Weather prediction using artificial neural network. In Proceedings of the IEEK Conference. 2002;262-264: The Institute of Electronics and Information Engineers.
45. Kakar SA, et al. Artificial neural network based weather prediction using Back Propagation Technique. Int. Journal of Advanced Computer Science and Applications. 2018;9(8):462-470.
46. Ahmed O, Brifcani A. Gene Expression Classification Based on Deep Learning," in 2019 4th Scientific International Conference Najaf (SICN). 2019;145-149:IEEE.
47. Omar N, Abdulazeez AM, Sengur A, Al-Ali SGS. Fused faster RCNNs for efficient detection of the license plates. Indonesian Journal of Electrical Engineering and Computer Science. 2020;19(2):974-982.
48. Sivanandam S, Deepa S. Introduction to neural networks using Matlab 6.0. Tata McGraw-Hill Education; 2006.
49. Zeebaree DQ, Abdulazeez AM, Zebari DA, Haron H, Hamed HNA. Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features.
50. Sheela KG, Deepa SN. Review on methods to fix number of hidden neurons in neural networks. Mathematical Problems in Engineering. 2013;2013.
51. Hussein HA, Abdulazeez AM. COVID-19 pandemic datasets based on machine

learning clustering algorithms: A review. PalArch's Journal of Archaeology of Egypt/Egyptology. 2021;18(4):2672-2700.
52. Fouilloy A, et al. Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. Energy. 2018;165:620-629.
53. Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2021;2(01):20 - 28.
54. Choi S, Kim YJ, Briceno S, Mavris D. Prediction of weather-induced airline delays based on machine learning algorithms. In 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). 2016;1-6:IEEE.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/68636