# A Comparison Analysis of Different Data Mining Techniques for Weather Prediction

Muhtasim Shafi Kader
Department of Computer Science and Engineering,
Daffodil International University, Dhaka, Bangladesh
muhtasim15-10809@diu.edu.bd

***Abstract*—Comparison analysis of several weather prediction data mining techniques involves forecasting and weather calculation for a future weather for days or times coming using data sets. This is one of the great contributions from sophisticated science in calculating prior quantitative data sets of weather status. Predict the weather status as well as weather the atmosphere manages towards the next 15 days. The work will efficient to prognostication of the weather of Dhaka City. We are using different Data Mining techniques for weather prediction. A long time ago people used physical components like an anemometer and a barometer for predicting the upcoming weather. Although it takes a huge period of time for prediction there are some bargains to maintain this equipment. Sometimes the prediction accuracy is very poor. As time moves on people have to find a way to predict the weather in the right way. With the help of Machine learning algorithms and Data Mining approaches the weather prediction can be more accurate. The purpose of this work is to predict the weather of Dhaka City for the next 15 days by using Machine approaches for data mining. We use the algorithms for machine learning, which follow linear and logistic regression as well as the method of Naïve Bayes. The datasets are collected from the last year's weather results of Dhaka City.**

***Keywords— prediction, Barometer, Anemometer, Data Mining, Machine learning, Linear regression, Naïve Bayes classifier, Logistic regression and so on.***

## I. INTRODUCTION

The weather affects everyday living in a significant way. It can be changed without additional warning at any moment. This instability occurs basically for various differences in the climate. But weather status prediction is an important element of everyday living. The predictability of weather is therefore highly essential. Today, supercomputers interpret the weather data. They get raw data from a satellite launched by space. However, the data obtained in raw format provide no insights. Thus, to enter the mathematical model, the cleansing of such data is necessary. This is called data mining. It is used to enter and predict the weather after cleansing the data in the mathematical model. Data from the last months are collected and a dataset is generated in this study report. They are classified in summer, fall, and winter to avoid complications. To prevent complexity. Each season consists of four months. Then the data sets include techniques such as linear regression besides logistic regression and Gaussian Naïve Bayes.

Machine learning algorithms used to make this project efficient. There will be a comparison analysis of different types of data mining approaches for predicting the weather for the next 15 days. The data mining-based project algorithm is given below, here we are using machine learning algorithms like KNN (K Nearest Neighbor) algorithm which can be imported by K Nearest Neighbor Classifier. We are using Decision Tree Classifier algorithm which can be imported by Decision Tree Classifier. We are using Random Forest Classifier which can be imported by Random Forest Classifier. We are using Gradient Boosting Classifier which can be imported by Gradient Boosting Classifier. And finally, we are using the Support Vector Machine Classifier which can be imported by SVC.

All of these algorithms may allow us to perfectly anticipate Dhaka City's whole weather condition for the following 15 days. In the dataset we utilize data from certain machine learning techniques for modeling to forecast weather conditions by analyzing Temperature (°C), Wind (KM/H), Gust (KM/H), Rain (MM), Humidity (%), Pressure (MB) data. By taking these data from the data set we will be able to find out a complete multiple weather status of Dhaka City for the next 15 days. After processing the data, we will find out the particular accuracy of different machine learning approaches.

The comparative analysis of weather prediction has been a challenging problem in Data mining test fields. We are working on the prediction of Dhaka City with proper dataset. The current data is using for research a model with the help of Data Mining algorithms that will be used to predict weather for next 15 days of Dhaka City. After predict we will measure the best accuracy with a comparative analysis. The Improvements of weather

Prediction with the use of machine learning algorithms currently has an enormous weather forecast influence on the progress of Artificial Intelligence and Data Mining techniques. Our objective is to get the greatest predictive accuracy.

We will try to collect the most accuracy rate among all of these Machine Learning algorisms. In weather prediction, the prediction of rain is the most common structure. There are also more components of the weather like lightings, storm, cloudy, thundershower, heavy rain, thunder outbreaks, cloudy, moderate or heavy rain, clear or sunny, light rain, partly cloudy and many more. But we are gathering information about a complete weather status with all this particular weather information. Here the accuracy will not remain the same in every term. It will change by the algorithms. Which algorithm performs the best accuracy that will be the best case for this prediction method and then there will be a comparison analysis.

## II. BACKGROUND STUDY

### *A. Data Mining Approaches:*

Here we are using different Data Mining techniques and calculate a comparative analysis of weather prediction. Predictive techniques of artificial information and data mining using machine learning algorithms are effective and advanced with a great deal of weather influence currently a few days. Our objective is to get the greatest forecast accuracy. Unfortunately, as the number of information increases, it becomes harder for meteorological instruments were being refined during the previous centuries. There are many data mining development around the world, weather prediction is one of the most interesting platforms among them. [2] Therefore, the changes weather condition is risky for human society. It affects human society in all possible ways. We are gathered data from the last two years to make a clear prediction of the next 15 days of Dhaka City.

### *B. Machine Learning Algorithms:*

After the prediction, we will make a comparative analysis of them The objective is to discover the best machine learning algorithm feasible in order to improve accuracy. Satellite-based systems are costly and need comprehensive support systems. For the data mining project, there will be used various machine learning algorithms. There will be great opportunities to make a big step in weather science with this project.

As mentioned earlier that we will be using various machine learning algorithms as well. The fact is that we will create a comparative analysis of different Data Mining techniques are being used to discover the most accurate methods for predicting Dhaka City's weather. The process has been done with help of datasets gathered with different weather information. We have datasets of two years of weather broadcast information as well. After applying the machine learning algorithms, the best result will be extraordinary.

### *C. Weather prediction:*

The procedure is to predict the weather of Dhaka City with some dataset of last couple pf years. After predicting the weather status with data mining techniques then we will make a comparative analysis of these techniques. Data mining is the process of extracting or mining knowledge from massive quantities of data. In other words, data mining is the process of obtaining meaningful, non-obvious information from a large amount of data. It mines massive datasets for hidden prognostic data. It's a fascinating new technique with a lot of promise for data analysis and decision-making. [11]. In typical data mining jobs, data mining functions are used to specify the kind of patterns that will be discovered. Here we are predicting weather components from dataset and calculate the best accuracy machine learning algorithm suitable for the study.

All of these Machine Learning Algorithms can assist us in obtaining a flawless forecast of Dhaka City's complete weather state for the following 15 days. We utilized data from several machine learning techniques to construct models to forecast weather status by evaluating Temperature (°c), Wind (km/h), Gust (km/h), Rain (mm), Humidity (percent), and Pressure (mb) data in the dataset. We will be able to obtain a comprehensive multiple weather status of Dhaka City over the following 15 days by using the data from the data set. We'll determine the accuracy of various machine learning techniques after we've processed the data.

## III. LITARATURE REVIEW

There have been several studies comparing data mining approaches for weather prediction.

Olaiya, F. et. al [1] has presented in their study that, they wanted to predict weather parameter like most of the lowest temperatures, rainfall, and wind at the time of particular period, respectively. Between 2000 and 2009, the data was for the Ibadan metropolis station. The comparison result demonstrates how these variables affect the weather.

Chauhan, D. et. al [2] proposed that in their research, they look at data mining approaches for weather prediction. The article was a survey that used data mining techniques to forecast weather. The paper demonstrates the ability of several machine learning algorithms for forecasting meteorological conditions such as temperature, rain fall, thunderstorms, and so on. with some basic machine learning approaches like artificial neural network, clustering, regression algorithm and so on. The comparison is made in this particular paper.

Sheikh et. al [3] has presented in their study that, data mining is using for weather prediction. The analysis of Naïve bayas and Decision Tree algorithm has used to predict weather here. The weather data gathered This collection includes work completed during a two-year period. The C4.5 (J48) decision tree method's performance was shown to be significantly superior to that of the Nave Bayes algorithm. C4.5 had an accuracy of 88.2 percent when it came to accurately categorizing the occurrences. When it came to categorizing the cases, however, Nave Bayes performed poorly, with a score of 54.8 percent. The above- mentioned remark about C4.5 being a higher performance in the case of weather dataset was also validated by the confusion

matrix. The number of true positives, that is, true cases that were also predicted true by C4.5, was higher than that of Nave Bayes, whereas the number of true negatives was lower.

Mandale, et. al. [4] has delivered in their study that, A decision tree classification technique for the categorization of meteorological data, including maximum temperature, lowest temperature, precipitation and evaporation, and wind speed, was employed in establishing decision treaties and regulations by month and year. The apparent trend over time may be studied with adequate data, and significant variations can be found showing changes in climate patterns. Artificial neural networks can connect input variables and produce outputs based on data patterns, without programming or complicated equations to represent such relationships. Decisive tree classification techniques for categorization by month and year of weather information such as maximum and minimum temperatures, rainfall, evaporation, and wind speed have been developed using C5 Decisions Tree Classification technology. With sufficient data, it is possible to examine the observed trend across time and detect major variations in climate change trends. Artificial Neural Networks can detect links between input variables and generate outputs based on data patterns without programming or complex calculations.

[5] Pandey et. al [5] proposed that, to forecast weather data The Hadoop tool in this study has been attempted to link to ANFIS and FL methods. Weather data are first collected using Beautiful SOUP and Python scripts from the website of the weather agency. The data collected is pre-processed using Hadoop's Wordcount technique. The final dataset is created after preprocessing and is then utilized in the weather prediction process. To predict weather data, two data mining methods are employed, namely ANFIS and FL are acronyms for acronyms for acronyms for acronyms for acronym The ANFIS approach forecasts weather data more correctly, according to the findings. Other soft computing approaches will be studied in the future for improved weather data prediction.

Xu, Q. et. al [6] proposed that, Applying the NWP data correction to short-term wind power predictions is the key contribution of this study. First, data mining algorithms are used to discover and cluster NWP problems, and then the incorrect raw NWP data is corrected before it is delivered to the WPF motor. The results of the simulation show that the approach presented successfully reduces the overall WFO error. However, issues remain in the proposed project

Radzuan et. al [7] has presented in their study paper that, in this review research, Methods have been examined in unclear time series. Previous research and tests shown how techniques may be used to extract information from specific time series data for future projects. On uncertain time series data, there were techniques used, but the prediction procedures remained constrained. The forecast relied heavily on uncertain time series. Monte Carlo simulation, PSO method, and so on for prediction.

A regression model in data mining technique was used to compare studies to assess the output of the prediction of uncertain time series. Through data mining methods, the approaches investigated brought benefits to the weather area.

Perez, R. et. al [8] described in their study that, there have Three validation tests comparing multi-day NWP irradiance forecasts for USA, Canada and Europe were completed. The comparison focused on the final use accuracy of different models, which included as a foundation global NWP model, multiscale and mesoscale, together with different postprocessing techniques for hourly site-specific forecasts ranging from simple interpolation to advanced statistic post-processing. The ECMWF global model and the GFS-driven WRF mesoscale model were all applied to three validation attempts and both were implemented in different configurations.

Lorenc, et. al [9] has presented at their paper that, The NWP-based multi-day irradiance forecasts for the United States, Canada, and Europe were compared by three validation experiments. In order to provide hourally-specified predictions from basic interpolation to complex statistical postprocessing, the confronts concentrated on the end-use accuracy of the different models using global, multi-scale and mesoscale MWP models as a basis.

All three validation efforts employed the ECMWF global model and the GFS-driven WRF mesoscale model, both of which were run in varied configurations.

Reddy et. al [10] has described in their study that, they cover a survey of several prediction approaches for early rainfall forecasting by different scholars in this article. The study also discussed the limits and difficulties that need to be addressed when using various rainfall forecasting systems. The review demonstrates that the Map Reducing Algorithm and Linear Regression Methods outperform other prediction techniques in terms of accuracy. NN also operates effectively on a big scale, according to the assessment yearly. However, NN delivers less accurate findings on a medium size and on a daily basis. FFNN delivers decent results on a monthly basis (medium size), TDNN performs better on a big scale basis, and FFNN generates better results on a short scale basis, according to the review weekly. In comparison to the other classification approaches, the Bagging Classification methodology is superior at forecasting the crop. FCM has the maximum accuracy of 93 percent, but SVM has a worse detection performance owing to outliers.

Nikam, V. et. al [11] has presented in their paper that, for predictive precipitation, the data mining technique is data-intensive instead than computational. In comparison to proven computational-intensive approaches our model seems virtually correct. Using a data mining technique reduces the overhead for computation, which makes it possible to analyze extremely big volumes of data in very little time and is said to be very efficient. The model may work on common hardware and does not require a high-performance cluster or supercomputer. The model is simple, highly predictive and can foresee binary and multi-class issues.

New classes may be learned fast via the Bayesian prediction model. The accuracy improves with the collection of additional learning data. improve. The model offers trustworthy predictions because of the enormous size of the training dataset. The drawback of the model is that if a predictor range lacks training data, a new record with that range is unlikely. This might be a big problem if this uncommon predictor value is required. The hybrid model can improve the accuracy of the model by combining different techniques to data mining or even the combination of computation-based models with data mining models.

Ali, M. et. al [12] has described in their study that in this work, they conducted tests and compared data mining approaches including Naive Bayes, KNN, and Decision Tree for weather predicting phenomena. The findings showed that decision trees were more successful in categorizing and modeling data sets, and that they were also effective in both classification and perdition. The KNN algorithm has the worst behavior of the three algorithms. Naive Bayes, a basic classifier based on Bayes theory, is a simple classifier to apply and shows to be efficient in performance when compared to the other two classifiers since it produces results that are almost identical to those of the Decision tree method.

Sharma et. al [13] has described in their paper that, the primary contribution of this study is the use of a closed loop nonlinear autoregressive artificial neural network to produce cost- effective and accurate one-day-ahead PV power predictions using NWP data. A network of neurons the results of this method's forecasts are contrasted. With other data-driven models that utilize and don't use NWP. The proposed strategy outperforms existing data-free NWP methods. Models improved by about 9% in terms of RMSE, while models improved by around 8% in terms of RMSE. MAE.A comparison of the CL-NAR-ANN and CL-NARXANN models, as predicted, the later model outperforms the former. Since the CLNAR-NARXANN.The ANN model is only based on historical data. To make the prediction, you'll need two days. It fails when the weather conditions suddenly change compared to the preceding two days. This modification is not a problem for the CL-NARX-ANN. As a result, it can be stated that the CL- NAR-ANN may be used as a suitable backup model in places with changing weather and as a primary forecasting model in areas with stable weather.

Medar, R. et. al [14] has delivered in their study that, A succinct and reliable weather forecasting model was created by comparing various current and proposed weather forecasting models. According to the results, the Hybrid MLR ANN model matches quite well. Interest in computing multi-linear regression weather parameters, ARIMA, ANN, and other ARIMA hybrid models of the MLR. The more ambiguity is particular, the better. Forecasts will improve in accuracy. A grouping of models can be utilized instead of only one, and by doing so, we will be able to improve our results. Cover the forecast's uncertainty.

Findawati, Y. et. al [15] has presented in their paper that, the findings of the comparison of algorithms employed in weather predictions of the K-Neighbor, Nave Bayes and C4.5 Study shows that the KNN classification algorithm from the specific dataset has the greatest accuracy and results are k = 7 and fold = 5 in weather forecast compared with the Nave Bayes classification method fold = 3 and C4.5 classification algorithm. Finally, Nave Bayes received 68.77 percent of the vote.

Yu et. al [16] has described in their study that, A hybrid short term wind speed forecasting technology was developed for this work using WPD-DBSCAN-ENN Hybrid Model. The WPD was the first to decompose the wind speed time series. The results were obtained DBSCAN was used to process deconstructed subseries, which were subsequently utilized to GBRT-supported ENNs should be established. Finally, all of the sub-predictions are combined. As a result, the final results were totaled up. Five numerical examples were used in this study to see if the proposed approaches' predictions are accurate. The Based on the aforementioned analysis, the following findings were obtained.

Choi et. al [17] has presented in their study that, This study created a prediction system that can classify aircraft delays caused by poor weather. The model created for using classical weather and prediction data. Machine learning techniques were utilized to evaluate data from individual OD pairings. Supervised machine learning algorithms have been put in place. Random forests, AdaBoost and k-means are among the methods used in this study. Decision Trees Nearest-Neighbor Due to the imbalance of information, the combination of SMOTE and random samples has been employed. Installation Precision predicted by the model We evaluated both the test and validation sets. The accuracy of the model has been checked both the test and the validation sets. In the future model, there are still alternatives for improving the issue. The optimal performance of classifiers may be easily established when the costs of the false positive or false negative are considered. In consequence, a decision maker that predicts the arrival of planes could give a strong basis. The predicted performance would also be improved by a comprehensive analysis of prediction uncertainties.

Sun et. al [18] has included in their study, The real achievement on the article is the implementation of the paper forecasting methodology that incorporates a hybrid clustering component as well as the prediction technique for IPSO-WNN. A comparable measurement feature, which combines Euclidean and Angle Cosine to identify days with equal wind speeds, is used in the hybrid clustering component. The relief method adapts the sample characteristics to their relevance, and then the IPSO-K means algorithm is utilized to obtain the best clustering result. An IPSO-WNN forecasting model's training samples are identified as daily samples that are comparable to the anticipated days. In comparison to conventional ARIM, Prediction errors such as RMSE and NMA are significantly reduced. When the prediction period horizon is extremely short, FFNN's and the recommended approach are predicting ability is near. When the prediction period horizon is expanded, therefore, the proposed strategy's action to foresee improves substantially.

The collected results show that the forecasting engine employed in this study has a reasonable prediction capacity. In the clustering stage, the variable characteristics are prohibited to regular air flow; next research would have target on other particular data calculation, such as air flow, air power, and so on, with the objective of discovering and establishing a more accurate air flow calculation method.

Wang, Z. et. al [19] has presented in their study that, in this work, forecasting model algorithms are utilized to categorize Weather such as highest temperature, wind speed and lowest temperature by month and year. Using the data mining method to the problems of wind farm production, particularly the prediction of wind speed. The system also includes the ARIMA time series prediction method, which is a data mining prediction tool. The platform can store large amounts of meteorological data, perform quick queries and analyses, and forecast weather, among other things. We also used the Artificial Neural Networks technique in this study, It may identify correlations between given data and produce result based on observed patterns in values expect the requirement for coding classification to express these co-relations.

Delerce et. al [20] has described in their paper that, by evaluating large quantities of observer agriculture data in combination with meteorological records, we were able to estimate the influence of climatic variability on rice. In 2 Colombian sites for a group of particular farms, we determined the main climate variables for rice production. The use of data mining tools, and the comparably high observational data spatiotemporal resolution, enables a thorough study of weather-yield relationships. Cropping episode clustering found apparent Relationships to climatic variability with the experience of the rice crop under various weather patterns as a collective knowledge basis. These figures show the highest yield in each weather pattern predictable as well as the cultivars which in certain conditions are best suited for each scenario.

Fugon et. al [21] has presented in their study paper that, the performance of several data mining techniques applied to short-term wind power forecasting is compared in this article. The utility of non-linear techniques is demonstrated, with performance comparable to that reported in the literature for wind farms in similar terrains. Random Forest outperforms the other models, according to the comparison. This model, which was initially developed for wind power forecasting, is intriguing since it does not require a lengthy architectural optimization phase; instead, just the number of trees in the forest must be optimized.
According to comparison, Random Forest exceeds other models. Initially created for the prediction of wind energy, this model is fascinating, because it needs no longer an architectural improvement; it has to be optimized only for the number of trees in the forest.

## IV. MEDTHOLOGY

The objective of this study is to anticipate the weather in Dhaka City for the next 15 days using Data Mining methods and Machine Learning algorithms. As a result, there will be some comprehensive explanations of the research effort in this area. The study subject and instrumentation will be discussed in further detail shortly. Weather is one of the most significant aspects in our everyday live. As a result, there should be effective techniques to anticipate weather in order to mitigate the damage caused by uncertainty in weather behavior. The most common approaches for weather forecasting are the utilization of vast amounts of data to gather information about future weather and the creation of equations to help predict weather by identifying distinct factors and replacing the values to get the desired outcome [3]. Information calculation is a critical component of machine learning. thus, this can be covered into that. Data mining have been the focus of decades of weather prediction research. Recently, academics have begun to emphasize the efficacy of data algorithms in predicting weather. After the formal buildup of the prediction, we will make a comparison of different data mining approaches to find the most accurate machine learning algorithm. Once the data has been collected, different data mining methods will be utilized to calculate the accuracy of the anticipated outcome. For 2019 and 2020, we are using data from the Dhaka weather report. However, while this study focuses in the "Summer," the approaches may be enhanced to the leftover information in terms for the lest. From this, you may predict any additional factors such as humidity, pressure, and wind speed. However, in this case, the research was conducted only for the purpose of forecasting temperature.

### *A. Research Subject and Instrumentation:*

The subject and dataset are the main instrument of this study cause this is a basic study endeavor, this must have good understanding. Even though, the study would differ then study since the outcome might change at any time. As a result, the work is successful in accurately understanding those variances. The equipment or gadgets utilized in this research are referred to as instrumentation. Here the weather status prediction will be happening with data mining approaches and machine learning algorithms to find a comparative analysis these algorithms.

### *B. Data Preprocessing:*

There is no work in data mining that could not be classified without data. So, the most important aspect of this research was data collection, which was also the most challenging aspect. Because discovering or acquiring data is not as simple as it appears. There was not one resource for information. For this project, several data mining approaches will be used after gathering the data to calculate the accuracy of the predicted outcome. Here the data is about the 2019 to 2020 weather report of Dhaka that we are using. However, this picture is titled "Summer." The approaches may also be applied to the remaining datasets. From here, predictions may be made on any additional factors such as humidity, pressure, and wind speed. However, in this case, the research was conducted only for the purpose of forecasting temperature.

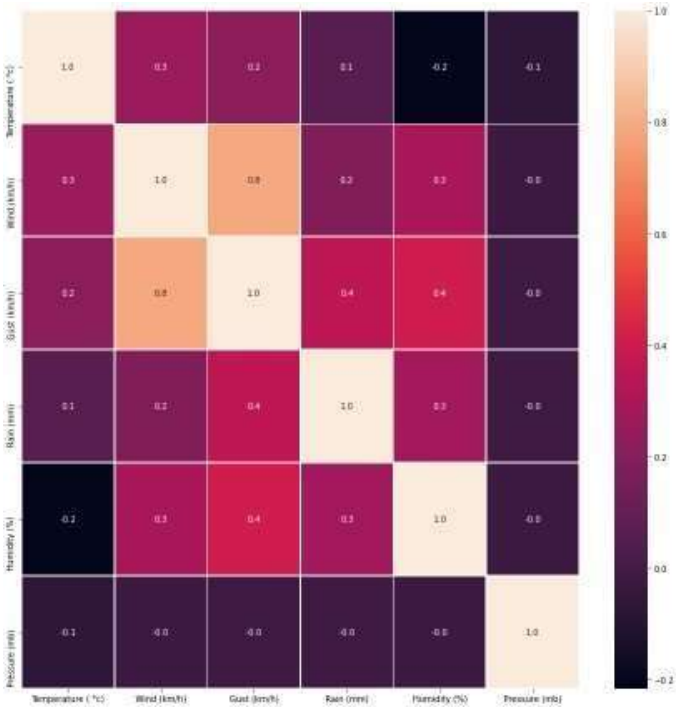


Fig. 1. The dataset description with visualization. (*figure caption*)

Figure Labels: Here we are plotting multiple weather status in this figure. For predicting weather status by analyzing Temperature (°C), Wind (KM/H), Gust (KM/H), Rain (MM), Humidity (%), Pressure (MB) data. By taking these data from the data set we will be able to find out a complete multiple weather status of Dhaka City for the next 15 days. After processing the data, we will find out the particular accuracy of different machine learning approaches

### *C. Statistical Analysis:*

The term "statistical analysis" refers to the use of quantitative data to investigate trends, patterns, and correlations. Scientists, governments, corporations, and other organizations utilize it as a research tool. Statistical analysis necessitates meticulous preparation from the outset of the research process in order to derive reliable results. You'll need to decide on your study design, sample size, and sampling process, as well as define your hypothesis. After you've collected data from your sample, you may use descriptive statistics to arrange and summarize it. You may next use inferential statistics to explicitly test hypotheses and create population estimates. Finally, you may put your findings into context and generalize them. This article provides students and researchers with a practical introduction to statistical analysis. Using two study examples, we'll guide you through the steps. The first looks into the possibility of a cause-and-effect link, whereas the second looks into the possibility of a correlation between variables.

When dealing with data, several problems occurred due to missing data in the dataset. These mistakes must be corrected since the effective deployment of data mining based after accurate data processing. As a result, correcting the dataset becomes the main goal for the work that we are wanted to be successed are so predictable for the study.

In Fig. 1, flowchart process that is matters the outcome result of the work.

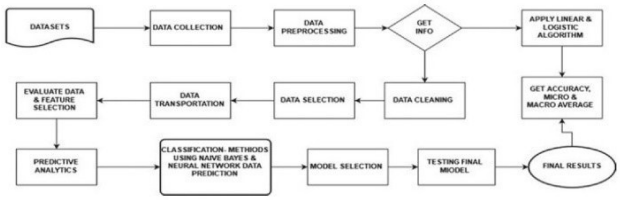


Fig 2. Flowchart for analysis (*figure caption*)

This the flowchart of the data mining techniques machine learning approaches of the research.

## V. REASULT AND DISCUSSION

The experimental model is constructed after training the data with four methods. The dataset was lacking certain values, which needed to be filled using the Panda's approach. As a result, the data can be more precise. The dataset was divided into two sections in order to create the regressor model:

- ✓ Train Dataset.
- ✓ Test Dataset.

There is the graphical interface, that is utilized to construct the structure. The first sections datasets are utilized to the training dataset, while the last one section is used as test data. The dataset had 2192 data points and 8 characteristics. Eighty percent of the infarction points were utilized to train, so the remaining twenty percent are using the test. Three distinct methods were employed to create the necessary model: linear regression, logistic regression, and Nave Bayes. Furthermore, with the assistance of,

### *A. Support Vector Machine (SVM):*

The purpose of the vector support technique is to locate a perfect forecast of n dimensions. To separate the two sets of There are different hyperplanes to picture from the data points. Our objective is to determine the greatest margin, or distance, for both groups of data points. Maximizing the range offers some consolidation and makes it easier to identify subsequent data points. Support Vectors are used to evaluate the best rest from dataset and it maintain the origin. Using them as a resource. We use maximize the classifier's margin by using vectors. In case the support vectors are being modified, then position of the line would change. These are the assets that will help us forecast using SVM. The "Support Vector Machine" is a open source supervised machine learning method that may be works for solve methods and optimization issues. However, it is primarily used to address problems. Into this SVM algorithm, the data components are displayed in n dimensions of a certain coordinate. Then we classify the data by identifying the plane this is clearly separates then these classes. In the project after predicting accuracy with the following Support Vector Machine (SVM) Algorithm the predicted

accuracy in 70% instead of the next 15 days weather information of Dhaka City respectfully.

### *B. Random Forest:*

Random Forest is a classifier that we use to forecast weather data. Random forest is an educational approach monitored. It builds a 'forest' with a set of decision trees, which are usually trained via the 'bagging' approach. The key principle of the bagging technique is to improve the final outcome by combining numerous learning models. The random forest has the benefit of being able to solve issues of classification and regression, typical in current machine learning systems. As classification is usually seen as a key element of machine education, let us look at the random forest. Underneath lies a random woodland with two trees. The hyperparameters of a random forest are almost the same as those in a tree of choice. Fortunately, instead of combining a decision tree with bagging classificatory, you may use the random forest classificatory class. You may use the algorithm regressor to deal with regression problems using random forests.

In the project after predicting accuracy with the following Random Forest Algorithm the predicted accuracy in 81% instead of the next 15 days weather information of Dhaka City.

### *C. Decision Tree:*

One of the algorithms of decision tree is controlled machine learning algorithms. Either a classification issue or a regression problem can be utilized. The decision tree addresses the question by means of the representation of the tree where the node of the leaf matches the class label, and the internal tree node shows features, in order to create a model which predicts the value of a target variable. As you can see in the accompanying diagram, it divides our data into two branches based on cholesterol: high and normal. Assume our new patient has high cholesterol. Based on the given statistics, we can't determine whether Drug B or Drug A will help.

Assume our new patient has high cholesterol. Based on the above data breakdown, we cannot say whether Drug B or Drug A will be appropriate for the patient. Also, if the patient's cholesterol is normal, we still don't have enough information to establish if Drug A or Drug B is appropriate for the patient. Look Let's at another criterion: age. As we can see, age is divided into three categories: young, middle age, and senior.

Assume our new patient has high cholesterol. Based on the above data breakdown, we cannot say whether Drug B or Drug A will be appropriate for the patient. Also, if the patient's cholesterol is normal, we still don't have enough information to establish if Drug A or Drug B is appropriate for the patient. Look Let's at another criterion: age. As we can see, age is divided into three categories: young, middle age, and senior.

In the project after predicting accuracy with the following Decision Tree Algorithm the predicted accuracy in 74% instead of the next 15 days weather information of Dhaka City.

### *D. KNN (K Nearest Neighbor):*

The KNN method is a kind of supervised machine learning approach to address classification problems and prediction of regression. However, in business it is mostly utilized to overcome difficulties with categorization and prediction. Because the training phase does not exist and all the data are used for training and classification, KNN is a lazy method for learning. It is also a non-parametric learning technique because it makes no assumptions about the underlying data. In Use of the KNN Algorithm

1. Determine the distance from each row of test data to the following approaches, employing one or more methods: Euclidean, Manhattan or Hamming. The Euclidean method is the most frequent way to determine distance.
2. Rearrange the values in ascending order based to the long-distance value.
3. After that, the top rows of the arranged array would be taken.
4. Depend on the main prevalent items of this rows, the test point will now be allocated a class.
5. Complete

In the project after predicting accuracy with the following KNN (K Nearest Neighbor) Algorithm the predicted accuracy in 67% instead of the next 15 days weather information of Dhaka City respectfully.

### *E. Gradient Boosting:*

Boosting is a way wherein weak students are transformed into powerful. In boosting, a modified version of the original data set is used to construct each new tree. Read about the AdaBoost algorithm for the first time to grasp the gradient boosting method. To begin with, for each observation the AdaBoost algorithm trains a decision tree with the same weight. After the examination of the first tree, we increase the weights of the data that are difficult to categorize and lower the weights of easy observations. The second tree is therefore built with weighted data.

As a consequence, Tree 1 Plus Tree 2 is our new model. This new 2-tree ensemble model will then compute the classification error and a third tree will be developed to anticipate the updated residuals. For a given number of iterations this process is repeated. Aid for the classification of observations following trees that were not appropriately classified by the previous trees. The prediction of the final ensemble model is based on the weighted sum of the forecasts given by the previous tree models.

When the AdaBoost model identifies defects by the use of high weight data points, the increase in gradients results in the same result through use of losses. The loss function is a measure of how well the coefficients of a model match the data. The easiest approach to understand the loss function is to think about what we are trying to do. In example the loss function will be based on the difference between real and expected domestic prizes by applying regression at predicted sales prices. Similarly, In the project after predicting accuracy with the following Gradient Boosting

Algorithm the predicted accuracy in 79% instead of the next 15 days weather information of Dhaka City.
These are the algorithms that are used predict a complete weather information system including such weather components. We will make a comparison analysis of these various data mining approaches to calculate the weather of next 15 days.

## VI. Experimental Result among Algorithms:

The weather of various methods is included below following constructing models Accuracy. The following is the algorithm details:

### *A. Linear Regression:*

Machine learning, and more especially predictive modeling, is focused with reducing model error or producing the most accurate forecasts possible at the price of interpretability. We will borrow, reuse, and steal approaches from a multitude of settings, including statistics, and use them to these goal in applied machine learning. As a consequence, whereas linear regression in the field of statistics was established as a model to analyze the link between the number variable inputs and outputs, machine learning has taken it over. This is a statistical and data mining approach used at the same time. Linear regression is a fundamental and well-known machine learning technique. This is a method of building correlations between variables. Linear regression contains two kinds of variables: continuous and independent. The line's slope is m (y if x = 0) and the intercepts c. Fig. 3 shows MaxTemp and MinTemp. The data in Fig. 4 is displayed in a dispersion plot to demonstrate the link.
Comparison analyzes of the various Dhaka city weather forecast data mining algorithms...

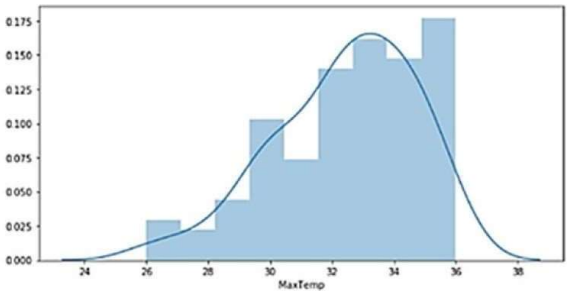


Fig 3. Linear Regression MaxTemp analysis (*figure caption*)

After using Linear Regression, the prediction of weather status will be as follows. The accuracy of the algorithm has been showing in the diagram. The dataset refers to the weather information of previous couple of years. With that dataset and help of data mining approaches the prediction have been made. And after prediction there will be a comparative analysis ofthe machine learning algorithms.

### *B. Logistic Regression:*

Your training data must be used to construct the logistic regression algorithm's coefficients. This is accomplished through the use of a maximum approximation. Although the distribution of your data is presupposed, maximum-like estimate is a way of learning. that is utilized by many machine learning algorithms. The optimal coefficients would yield a model that predicted a value for the default class very close to 1 and a value for the other class very close to 0. Maximum for logistic regression is based on the idea that an analytical approach seeks coefficient values that minimize the divergence between probabilities predicted by the model.

The assumptions that logistic regression makes about the distribution and connections in your data are quite similar to those that linear regression makes. These assumptions have been thoroughly researched, and exact probabilistic and statistical language has been employed. My recommendation is to use them as recommendations or rules of thumb when experimenting with various data preparation techniques. In the end, the goal of predictive modeling and machine learning initiatives is to make accurate predictions rather than analyze the findings. As a result, as long as the model is resilient and performs well, you can break some assumptions.
Logistic regression is another famous and widely used technique for dealing with classification issues that linear regression cannot solve. For example, If someone wants to extract positive and negative values from an array of random data, logistical regression must be applied. In this research, the logistical classification is used to categorize MaxTemp and MinTemp as a mid-value, as shown in Fig. 6. The logistic function accepts every integer between 0 and 1. The aim is,

Let's use t as a linear function in a regression model,

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \quad (1)$$

$$t = \beta o + \beta 1 x \quad (2)$$

The logistic equation would then become,

$$p(x) = \frac{1}{1 + e^{-(\beta o + \beta 1 x)}} \quad (3)$$

After the evaluation of Logistic Regression, the expected outcome will be come in the comparative analysis. As the following diagram the analysis would be given. The weather prediction will be calculated by the given dataset. After the prediction of the weather status there will be a comparison of all the data mining approaches that has been used in the procedure.

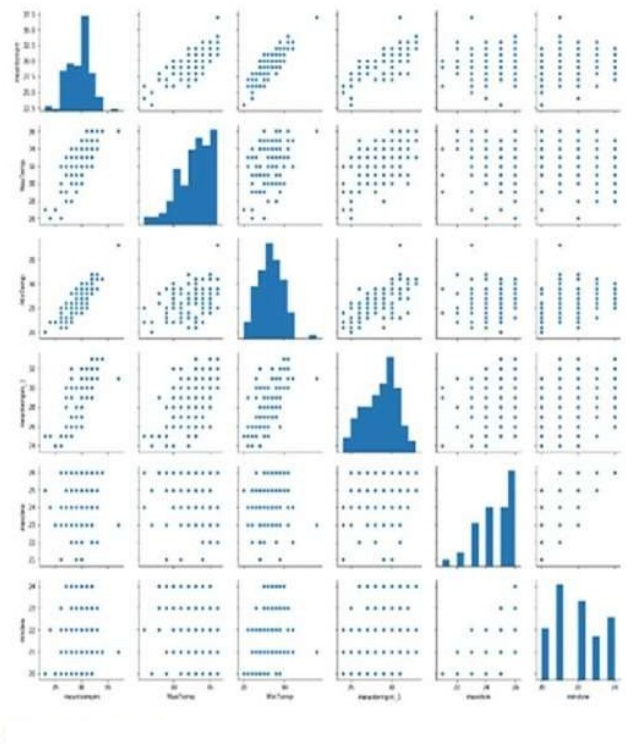

Fig 4. Scatterplot result (*figure caption*)

Machine Learning algorithms are commonly employed in prediction studies such as this one. Plotting is a helpful tool for display of the dataset as well as prediction in the graphical interface. It may also be extremely useful for comparing various data mining approaches.

*C. Naïve Bayes:*

The Bayes Hypothesis probabilistic algorithm Naïve Bayes is used in a wide range of classification applications. You will get a clear and full explanation of the Naïve Bayes algorithm, along with all essential notions, in this post, so that there are no doubts or gaps in your understanding The Nave Bayes approach is a probabilistic machine learning method that may be used to a variety of classification tasks. Filtering spam, identifying documents, and forecasting sentiment is sources of potential use. It is based on Rev. Thomas Bayes' writings, hence the name. The term naïve assumes that the characteristics that make up the model are independent of one another. That is, altering the value of one feature has no direct impact on the value of the other features included in the algorithm. This is due to the fact that Naïve Bayes has a huge edge. Because it is such a probabilistic model, the method might well be successfully turned and predictions produced. Quick throughout real-time. It is so easily scalable, and it is frequently the algorithm of choice for real-world applications which must respond to user requests immediately.

This is a well-known classification approach depends in the Bayes theory.

The Bayes presupposes in that each classifier in the analysis has freedom. This algorithm is extremely beneficial when dealing with big datasets. The Bayes theorem calculates P(a) as the likelihood of P(a), P(b), follows by P(b). So, Bayes theorem equation is shown below:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} \tag{4}$$

A Comparison of different weather forecasting data mining techniques from Dhaka City... 301

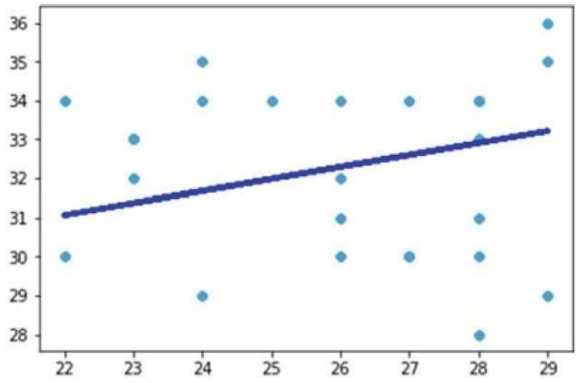


Fig. 5 Visualization about prediction (*figure caption*)

| | Actual value | Predicted value |
|---|---|---|
| 0 | 36 | 33.212321 |
| 1 | 30 | 32.597770 |
| 2 | 33 | 31.368670 |
| 3 | 29 | 33.212321 |
| 4 | 33 | 33.905045 |

Table 1. Actual value versus predicted value (*figure caption*)

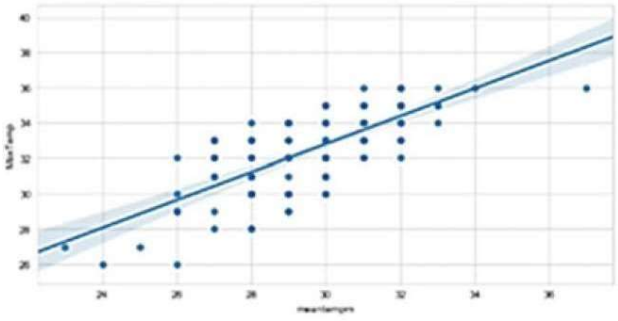

Fig. 6 Logistic Regression Implementation (*figure caption*)

Then, P(a|b) = P(b1|a) × P(b2|a) ×· · ·× P(b|a) × P(a) (5)

As this technique is mostly utilized in text classification and multiple class issues, a simple Naïve Bayes technique is performing to estimate into likelihood of various groups based on different inputs. The precap measured the column, which is seen in Fig.7. Naïve Bayes method is depicted. The performance of the Naïve Bayes classifier is measured in Table 2. A classifier's precision establishes its accuracy, whereas recall defines the whole model. So, the accuracy is 59 percent and the recall is 25 percent in this case.

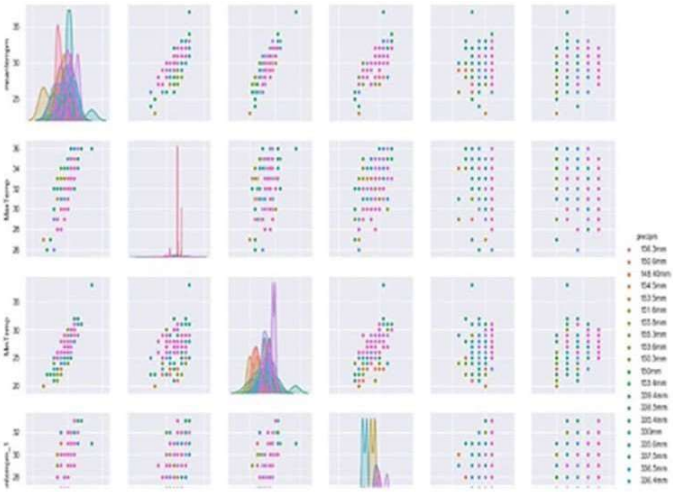

Fig. 7 Naïve Bayes Algorithm Implementation (*figure caption*)

| Matric | Score (%) |
|---|---|
| Precision | 58.70 |
| Recall | 25 |
| Accuracy | 29 |

Table 1. Naïve Bayes accuracy prediction (*figure caption*)

Over all Naïve Bayes classifier is used here for the weather prediction. Possibly it will provide a major accuracy for the prediction. It isn't just the crucial to know what taken place. it's also crucial to know how likely it is that it will happen in future. Randomness vs. likelihood is the main theme of probability theory. In a nutshell, it determines the likelihood of an event in a random space. If I flip a coin and expect heads, there is a 50%, or 12 percent, chance that my expectation will be satisfied, presuming the act of flipping is neutral a fair, or unbiased coin has the same probability to get head or tail. This expectation of fairness is assigned to randomness, and my probability is the possibility of achieving the expectation.

After the prediction all the prediction there will be comparative analysis from the data mining techniques that which approaches is providing the most accurate results.

## VII. Accuracy Comparison:

Maintaining all the data mining procedures with among the machine learning algorithms the project has come to end with a result analysis of the weather. The target was able to make a prediction of weather status of Dhaka City for next 15 days. After the prediction there is the accuracy comparison among all the data mining approaches that have been used for this study. So, In Fig. 7, there is a comparison of the accuracy of three methods used to create the model. When compared to the other methods, linear regression provides the highest accuracy. The comparison analysis shows that linear regression predicts the most by its algorithms.

So, the comparison analysis has been referring that Random Forrest Classifier of Linear Regression has the best accuracy. That is the prediction of weather information for Dhaka City. The time line is for next 15 days.

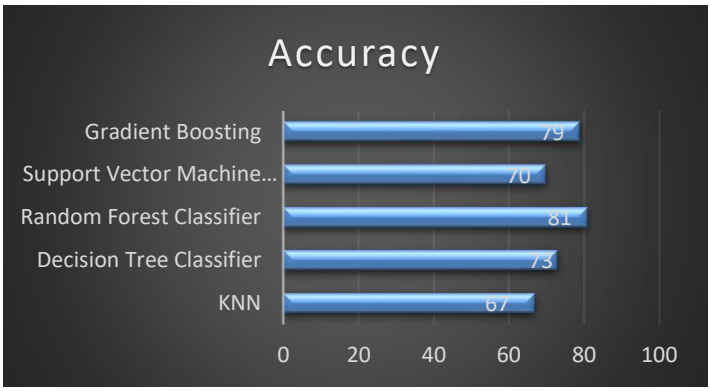


Fig. 8 Accuracy comparison among various algorithms (*figure caption*)

## VIII. Conclusion

The goal of the research is development of weather forecast prediction that would provide a comparison of several data mining techniques for complete weather status of Dhaka City during the next 15 days. As the weather status in Dhaka City has changed, complete dataset converted into presently being included into the model. In this investigation, five algorithms were used. Weather Status Prediction in Dhaka City was supplied by Support Vector Machine (SVM), Random Forest, Decision Tree, KNN, and Gradient Boost. In comparison to others, by using enough precision. The accuracy of the deployment for the leftover three information (summer as well as fall and winter) will be the primary issue in the future. Perhaps, along with some additional algorithms, there will be some comparison studies of other algorithms, allowing us to explore better weather forecast approaches. We feel it is critical to use for data mining researchers to employ data mining approaches. The study has been a big impact weather science as well. The comparison among data mining techniques can be use in different fields rather than weather prediction. There are more data mining algorithms such as Support Vector Machine (SVM) follows by Gradient Boosting, KNN, as well as Decision Tree are also used in weather prediction performance.

## REFERENCES


[1] Olaiya, F., & Adeyemo, A. B. (2012). Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1), 51.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Chauhan, D., & Thakur, J. (2014). Data mining techniques for weather prediction: A review. International Journal on Recent and Innovation Trends in Computing and Communication, 2(8), 2184-2189.K. Elissa, "Title of paper if known," unpublished.

[3] Sheikh, F., Karthick, S., Malathi, D., Sudarsan, J. S., & Arun, C. (2016). Analysis of data mining techniques for weather prediction. Indian Journal of Science and Technology, 9(38), 1-9..

[4] Mandale, M. A., & Jadhawar, B. A. (2015). Weather forecast prediction: a Data Mining application. Int J Eng Res Gen Sci, 3(2).

[5] Pandey, A. K., Agrawal, C. P., & Agrawal, M. (2017, February). A hadoop based weather prediction model for classification of weather data. In 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-5). IEEE.

[6] Xu, Q., He, D., Zhang, N., Kang, C., Xia, Q., Bai, J., & Huang, J. (2015). A short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining. IEEE Transactions on sustainable energy, 6(4), 1283-1291.

[7] Radzuan, N. F. M., Othman, Z., & Bakar, A. A. (2013). Uncertain time series in weather prediction. Procedia Technology, 11, 557-564.

[8] Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Van Knowe, G., Hemker Jr, K., ... & Pomares, L. M. (2013). Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Solar Energy, 94, 305-326.

[9] Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 112(474), 1177-1194.

[10] Reddy, P. C., & Babu, A. S. (2017, February). Survey on weather prediction using big data analystics. In 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-6). IEEE.

[11] Nikam, V. B., & Meshram, B. B. (2013, September). Modeling rainfall prediction using data mining method: A Bayesian approach. In 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation (pp. 132-136). IEEE.

[12] Ali, M., Asklany, S. A., El-wahab, M. A., & Hassan, M. A. (2019). Data Mining Algorithms for Weather Forecast Phenomena Comparative Study. International journal of computer science and network security, 19(9), 76-81.

[13] Sharma, V., Cali, U., Hagenmeyer, V., Mikut, R., & Ordiano, J. Á. G. (2018, June). Numerical weather prediction data free solar power forecasting with neural networks. In Proceedings of the Ninth International Conference on Future Energy Systems (pp. 604-609).

[14] Medar, R., Angadi, A. B., Niranjan, P. Y., & Tamase, P. (2017, August). Comparative study of different weather forecasting models. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1604-1609). IEEE.

[15] Findawati, Y., Astutik, I. I., Fitroni, A. S., Indrawati, I., & Yuniasih, N. (2019, December). Comparative analysis of Naïve Bayes, K Nearest Neighbor and C. 45 method in weather forecast. In Journal of Physics: Conference Series (Vol. 1402, No. 6, p. 066046). IOP Publishing.

[16] Yu, C., Li, Y., Xiang, H., & Zhang, M. (2018). Data mining-assisted short-term wind speed forecasting by wavelet packet decomposition and Elman neural network. Journal of Wind Engineering and Industrial Aerodynamics, 175, 136-143.

[17] Choi, S., Kim, Y. J., Briceno, S., & Mavris, D. (2016, September). Prediction of weather-induced airline delays based on machine learning algorithms. In 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC) (pp. 1-6). IEEE.

[18] Sun, G., Jiang, C., Cheng, P., Liu, Y., Wang, X., Fu, Y., & He, Y. (2018). Short-term wind power forecasts by a synthetical similar time series data mining method. Renewable energy, 115, 575-584.

[19] Wang, Z., & Mujib, A. M. (2017, October). The Weather Forecast Using Data Mining Research Based on Cloud Computing. In Journal of Physics: Conference Series (Vol. 910, No. 1, p. 012020). IOP Publishing.\

[20] Delerce, S., Dorado, H., Grillon, A., Rebolledo, M. C., Prager, S. D., Patiño, V. H., ... & Jiménez, D. (2016). Assessing weather-yield relationships in rice at local scale using data mining approaches. PloS one, 11(8), e0161620.

[21] Fugon, L., Juban, J., & Kariniotakis, G. (2008, March). Data mining for wind power forecasting. In European Wind Energy Conference & Exhibition EWEC 2008 (pp. 6-pages). EWEC.