VOL. 14, NO. 18, SEPTEMBER 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PREDICTION OF HUMIDITY IN WEATHER USING LOGISTIC REGRESSION, DECISION TREE, NEAREST NEIGHBOURS, NAIVE BAYESIAN, SUPPORT VECTOR MACHINE AND RANDOM FOREST CLASSIFIERS

G Sujatha, Dr Chinta Someswara Rao and T Srinivasa Rao
Department of Computer Science Engineering, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India
E-Mail: chinta.someswararao@gmail.com

ABSTRACT

The ultimate objective of this system is to predict the variation of humidity in the weather over a given period. The weather condition at any instance is described by using different kinds of variables. Out of these variables, significant variables only are used in the weather prediction process. The selection of such variables depends strongly on the location. The existing weather condition parameters are used to fit a model and by using the machine learning techniques and extrapolating the information, the future variations in the parameters are analyzed.

Keywords:

1. INTRODUCTION

Weather forecasting is an essential application in meteorology. It has been one of the most scientific challenging problems around the world. Weather Humidity prediction is mainly concerned with the prediction of weather condition for a given environment. The prediction of weather condition is essential for various purposes like climate monitoring, drought detection, severe weather prediction, agriculture and production, planning in energy industry, aviation industry, communication, pollution dispersal, and so forth. In military operations, there is a considerable historical record of instances where weather conditions have altered the course of battles. Accurate prediction of weather conditions is a difficult task since weather is non-linear and dynamic process i.e., it varies from day to day and even from minute to minute. The accuracy of the prediction depends on the knowledge of previous weather conditions over large areas and over large period. Weather forecasts provide critical information about the future weather. There are various approaches available in weather forecasting, from relatively simple observation of the sky to the highly complex computerized mathematical models.

The weather condition at any instance may be represented by some variables. Out of those variables, the most significant are selected to be involved in the process of prediction. The selection of variables is dependents on the location for which the prediction is to be made. The variables and their range always vary from place to place. The weather condition of any day has some relationship with the weather condition existed in the same tenure of previous year and previous week. Rainfall is a form of precipitation. Its accurate forecasts can help to identify possible floods in future and to plan for better water management. Weather forecasts can be categorized as present forecasts are forecasts valid up to few hours, short-range forecasts are like mainly rainfall forecasts for 1 to 3 days, medium range are valid forecasts for 4 to 10 days and long-range forecasts are valid for more than 10 days. Short range and Medium Range forecasts like rainfall are important for flood forecasting and water resource management.

Traditionally, weather forecasting has always been performed by physical simulation of the atmosphere treating it as a fluid. The current state of the atmosphere is sampled. The future state of the atmosphere is computed by solving numerical equations of thermo dynamics and fluid dynamics. However, this traditional system of differential equations that govern the physical model is sometimes unstable under disturbances and uncertainties while measuring the initial conditions of the atmosphere. This leads to an incomplete understanding of the atmospheric processes, so it restricts weather prediction up to a 10 day period, because beyond that weather forecasts are significantly unreliable. But Machine learning is relatively robust to most atmospheric disturbances as compared to traditional methods. Another advantage of machine learning is that it is not dependent on the physical laws of atmospheric processes.

2. LITERATURE SURVEY

Significant work has been done in forecasting of Temperature & Pressure. Some of these are presented here. This lead us to make better understanding of research work. Some basic concepts, findings & facts of such work will be extracted to make some conventions for our project.

Lynn Houthuys, Zahra Karevan and Johan A.K. Suykens., [1] proposed a data-driven modeling technique for temperature prediction. They have used Soft Kernel Spectral Clustering (SKSC) to discover similar samples to the test point, for learning purpose. Due to its high dimensionality, they have used Elastic net as a basis for the selection required features. For each cluster, features are chosen independently and then Least Squares Support Vector Machines (LS-SVM) regression is used to fit the data. Finally, the predicted values by LS-SVM are averaged out based on the membership of the test point to each cluster. Experimental results proved that, their

proposed system's performance is better when compared to "weather underground" and performed as good as all other existing weather temperature prediction methods.

Kuei-Chung Chang *et al*., [2] proposed Arduino weather box to collect the weather data from the living space, and by using Support Vector Machine (SVM) and Neural Network (NN), they analyzed the collected data to predict thermal conditions and chances for rainfall. Finally they build a model to predict temperature and probability of rainfall. The experiments show that they achieved best results in predicting the temperature by using back propagation network (BPN) and predicting the rainfall by using support vector machine (SVM).

Jos´e L. Aznarte and Nils Siebert [3] presented a dynamic line-rating forecasting (DLR) study and an accurate procedure to predict future values of the rating for two conductor lines. This procedure involves data transformations and the results suggest using multivariate adaptive regression splines (MARS) as the core regression method. Their approach obtains good error results which rank far below the key project indicators defined in the Twenties Project for the DLR problem. These results prove the feasability of computing line-rating forecasts that can be used in the daily operation of the power system to lift some constraints while maintaining safe and reliable operating conditions.

Jos´e R. Andrade and Ricardo J. Bessa [4] proposed a forecasting framework to discover information from a grid of numerical weather predictions (NWP) applied to both wind and solar energy. They have extracted maximum information from the NWP grid by combining the gradient boosting trees algorithm with further engineering techniques. Experimental results show that their proposed system is somewhat better than a model that only considers one NWP point for a given location.

Arief Koesdwiady *et al*., [5], proposed a wide-ranging prediction architecture that incorporates deep belief networks (DBN) and data fusion to obtain more accurate traffic flow prediction in San Francisco Bay Area using traffic flow history and weather data through correlation analysis. In their proposed architecture, they have differentiated several scenarios to highlight the merit of using data fusion at decision level. Experimental results show that the state-of-the-art techniques were outperformed by their data driven urban traffic system.

Doreswamy *et al*., [6] proposed a new method to handle missing values in weather data using machine-learning algorithms by executing experiments with NCDC dataset to evaluate the prediction error of five methods namely the kernel ridge, linear regression, random forest, SVM imputation and KNN imputation procedure. By using each method, the missing values are calculated and compared to the observed value. Results of the proposed method were compared with existing ones.For evaluating the results, the performance metrics considered are standard deviation of error (STDE), variance of error (VARE), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), bias and coefficients of determination R2.

Rushika Ghadge *et al.*, [7] proposed a model to assist farmers by predicting, whether the crop is suitable for cultivation or not based on soil type and soil quality. This helps farmers to maximize their crop yield by using the suggested appropriate fertilizers. The proposed system gives best results with the help of supervised and unsupervised machine learning algorithms. Depending on best and accurate output, one of the two algorithms is chosen. Thus, the system will help to reduce the difficulties faced by the farmers, stop them from losses, and even lose the life by attempting suicide.

Sun Choi *et al*., [8] proposed a model to forecast the airline delays due to the sudden changes in the weather by using data mining and supervised machine learning algorithms. They considered US domestic flight data and the weather data from 2005 to 2015 to train the model. They have implemented Decision trees, random forest, the AdaBoost and the k- Nearest-Neighbors algorithms to build a model that can forecast possibility of the airline delays. Based on the accuracy and ROC curves of these algorithms, the best can be chosen and used to classify whether the given flight will be on time or being delayed.

Soumya Tiwari *et al.*, [9] have used gradient boosting regression and bootstrap aggregation machine learning models to build a model that can predict the short-term solar irradiance at a given location. They considered parameters such as spatial parameters (elevation, latitude, longitude) and seasonal parameters (day and month of the year). Effectiveness of the proposed method will be evaluated based on Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) indices

Navin Sharma *et al*., [10] have used machine learning techniques to build an automatically creating site-specific prediction model for solar power generation from national weather service (NWS) weather forecasts. Regression techniques like linear least squares and support vector machines were compared to generate prediction models. For each of these models, they evaluate the accuracy by using NWS forecasts and solar intensity readings from a weather station deployment. Experimental results show that the predictions done by the SVM based prediction model are more accurate compared to all other existing models

Zhao Liu and Ziang Zhang [11] proposed a model that can provide the estimated photo voltaic (PV) output using K-Nearest Neighbors algorithm. This model uses numerical weather and solar irradiance prediction data and also includes a weather condition classification process and a physical model of PV units. Experimental results show that the uncertainity of PV output forecasting can be reduced by the proposed system. This can also enhance the reliability and operation efficiency of the model.

Sam Sanders *et al*., [12] developed a model to make predictions of solar radiation. The data was collected from the Georgia Automated Environmental Monitoring Network (GAEMN) and the National Oceanic and Atmospheric Administration (NOAA) for five cities in Georgia. They observed that early predictive models made

use of historically recorded solar radiation data only and other weather phenomena as inputs, while later models incorporated weather forecasts for the target area and surrounding areas.Their results proved that Random Forecasts method achieved the lowest error rate compared to other machine learning techniques.

Rastislav Rusn´ak and Rudolf Jakˇsa [13] tested stochastic weights update methods on forecasting of real world weather data. They observed that the performance of proposed methods is comparable with plain back-propagation and thus they came to know that these methods are inter-changeable. They have introduced two new methods stochastic weights selection and stochastic neurons selection. Their results support the results of Salvetti and Koscak2010 shuffle δ updates method.

Sara Landse *et al.*, [14] presented a case study in which they look at injuries during 713 Major League Soccer games across the 2015 and 2016 seasons. Their dataset consists of 713 regular season games out of which 548 games recorded at least one injury. Totally their dataset contains the information on 1,238 separate in-game injuries. In this paper based on local weather and playing surface they compared the performance of 9 different classifiers for predicting the chance of getting injury. They find that Support Vector Machine (SVM) is the best suitable algorithm for their dataset.

Man Galih Salman *et al*., [15] studied deep learning techniques for weather prediction. They compared forecasting performance of Recurrence Neural Network (RNN), Conditional Restricted Boltzmann Machine (CRBN) and convolutional Network (CN) models. They tested these models using weather dataset provided by BMKG (Indonesian Agency for Meteorology, Climatology, and Geophysics). They are expected to contribute their study to different areas such as flight navigation, agriculture and tourism.

T.R.V.anandharajan *et al.*, [16] developed an intelligent weather-predicting module. In this paper they considers parameters such as maximum temperature, minimum temperature and rainfall.They have taken these parameters for sampled period of days. They predicted the available data using linear regression technique, which predicts the next day's weather with good accuracy.

Navadia, Sunil, *et al*. [17] proposed a system that serves as a tool, takes the rainfall data as input, and predicts the future rainfall with min, max and average rainfall in an efficient manner.

3. METHODOLOGY

The methodology of framework is shown in the Figure-1, which consists of three modules namely data pre-processing, machine learning and prediction.

Data pre-processing: The data pre-processing is a technique by which the data is refined, transformed and cleaned for improving the quality of the training data on which the data model is prepared for decision making or prediction. This module consists of three parts data collection, data cleaning and data annotation.

Historic data from csv file: The historical data is collected from kaggle.com as CSV file and on this CSV file we applied data cleaning.

Data cleaning: Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or common data. Data cleaning may be performed interactively with data wrangling tools, or as batch processing through scripting. The actual processing data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities.

Data annotation: Data annotation is the task of labelling different types of data like images, audio, text or video. It is done by selecting a zone of the data and adding a label to this specific zone.

Normalized Data: It is the process of organizing data to minimize redundancy.

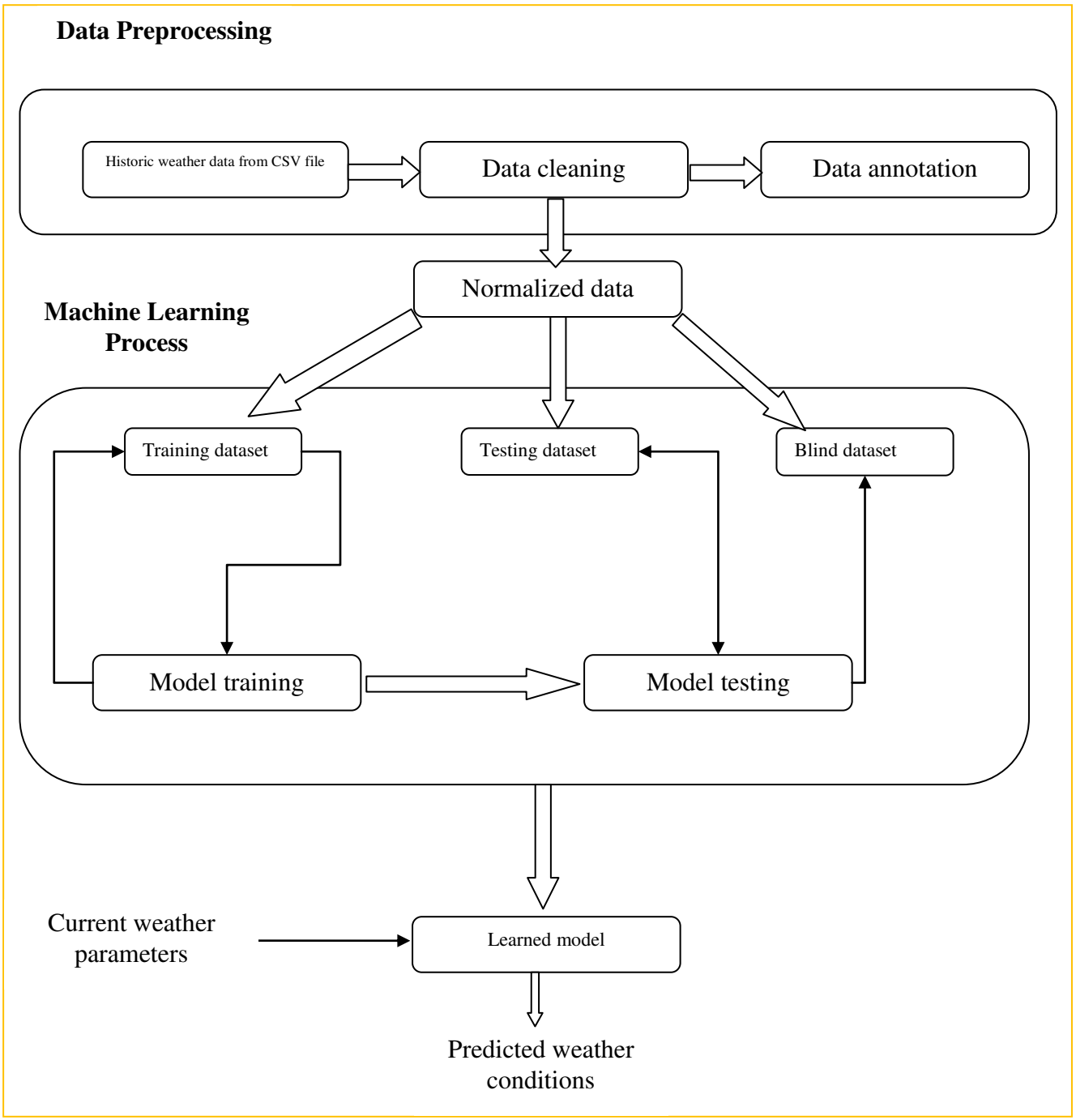

Figure-1. System structure.

Machine learning process: The machine learning process goes through the following functions.

Training dataset: Training Data is labelled data used to train machine learning algorithms and increase accuracy. A model is generally provided with a data set of known data, called the training data set.

Model Training: The process of training a Machine Learning model involves providing a Machine Learning algorithm (that is, the learning algorithm) with training data to learn from. The term Machine Learning model refers to the model artefact that is created by the training process. The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs a Machine Learning model that captures these patterns.

Testing dataset: A set of unknown data against which the model is tested, known as the test data set. The test dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Model testing: The testing dataset is given to simulation model for testing purpose.

Blind dataset: The blind dataset is used for cross validation. A round of cross-validation comprises the partitioning of data into complementary subsets, then performing analysis on one subset. After this, the analysis is validated on other subsets (testing sets). To reduce variability, many rounds of cross-validation are performed using many different partitions and then an average of the results is taken. Cross-validation is a powerful technique in the estimation of model performance.

Learned Model: It takes current weather data parameters as input and it provides the predicted weather conditions as the result.

3.1 ALGORITHMS EXPLANATION

3.1.1 K-Nearest Neighbour

Algorithm 1
Input: Weather dataset
Output: Humidity prediction

STEP 1:	Load the training and test data.
STEP 2:	Choose the value of K.
STEP 3:	for each point in test data
STEP 3.1:	Find the Euclidean distance to all training data points.
STEP 3.2:	Store the Euclidean distances in a list and sort it.
STEP 3.3:	Choose the first k points.
STEP 3.4:	Assign a class to the test point based on the majority of classes present in the chosen points.
STEP 4:	End.

3.1.2. Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal and interval independent variables.

Algorithm 2: LOGISTIC REGRESSION
Input: Weather dataset
Output: Humidity prediction

STEP 1: Take two variables x as feature vector and y as response vector.

STEP 2: Draw the regression line. The equation of regression line is represented as

$$\ln\left(\frac{P}{1-P}\right) = m + kx$$

Solving this equation for P, we get the logistic probability:

$$\left(\frac{P}{1-P}\right) = e^{m+kx}$$

$P = e^{m+kx} - P\,(e^{m+kx})$

$P + P\,(e^{m+kx}) = e^{m+kx}$

$P\,(1+e^{m+kx}) = e^{m+kx}$

$$P = \frac{e^{m+kx}}{1+e^{m+kx}}$$

$$P = \frac{1}{\frac{1+e^{m+kx}}{e^{m+kx}}}$$

$$P = \frac{1}{\frac{1}{e^{m+kx}} + 1}$$

$$P = \frac{1}{1+e^{-(m+kx)}}$$

In linear regression we add several dimensions to the problem.

$$P = \frac{1}{1+e^{-(m+k_1x_1+k_2x_2+.....+k_{N-1}x_{N-1}+k_Nx_N)}}$$

STEP 3: End.

3.1.3 Decision Tree

Algorithm 3: Decision Tree
Input: Weather dataset
Output: Humidity prediction

STEP 1: Read the data set.
Pick the best attribute from the dataset for splitting using information gain. The following is the equation for information gain

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

Where P_i is the non zero probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$.

STEP 2:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j).$$

Where the term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A.
Gain(A)= Info(D)- $Info_A(D)$
Where Gain(A) tells us how much would be gained by branching on A.

STEP 3: Repeat step2 until end of the tree construction.
STEP 4: End.

3.1.4 Random Forest

Algorithm 4: Random Forest
Input: Weather dataset
Output: Humidity prediction

STEP 1: Read the data set.
STEP 2: Select k number of decision trees.
STEP 3: Voting will be conducted based on the output from decision trees.
STEP 4: Calculate the votes for each predicted target.
STEP 5: Take the high voted predicted target as the final prediction.
STEP 6: Results will be given to the user.
STEP 7: End.

3.1.5 Naïve Bayesian

Algorithm 5: Naïve Bayesian
Input: Weather dataset
Output: Humidity prediction

STEP 1: Read the data set.
STEP 2: Convert the data set into frequency table.
STEP 3: Create Likelihood table by finding the probabilities.
STEP 4: Apply Naïve Bayesian equation to calculate the posterior probability for each class. The equation is:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

STEP 5: The class with the highest posterior probability is the outcome of prediction.
STEP 6: Results will be given to the user.
STEP 7: End.

3.1.6 Support Vector Machine

Algorithm 6: Support Vector Machine
Input: Weather dataset
Output: Humidity prediction

STEP 1: Read the data set.
STEP 2: Select the hyper-plane which segregates the two classes better.
STEP 3: Select the hyper-plane with higher margin is robustness.
STEP 4: Select the hyper-plane which classifies the classes accurately prior to maximizing margin.
STEP 5: Results will be given to the user.
STEP 6: End.

4. RESULTS

4.1.1 Evalution metrics

Accuracy: Accuracy is computed as "the total number of two correct predictions, True Positive (TP) + True Negative (TN) divided by the total number of a dataset Positive (P) + Negative (N)".

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

Precision: Precision is computed as "the number of correct positive predictions (TP) divided by the total number of positive predictions (TP + FP)". Precision is also known as a positive predictive value.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: Recall is computed as "the number of correct positive predictions (TP) divided by the total number of positives (P)". Recall is also known as the true positive rate or sensitivity.

$$\text{Recall} = \frac{\text{TP}}{\text{P}}$$

Confusion matrix: A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

Testing dataset's Confusion matrix for DTC is shown in Figure-2 and precision, recall, f-score and support of DTC is shown in Table-1. With the data of precision-recall from Table-1, a graph is drawn which is shown in Figure-3, precision-recall with threshold is shown in Figure-4 and ROC curve is shown in Figure-5.

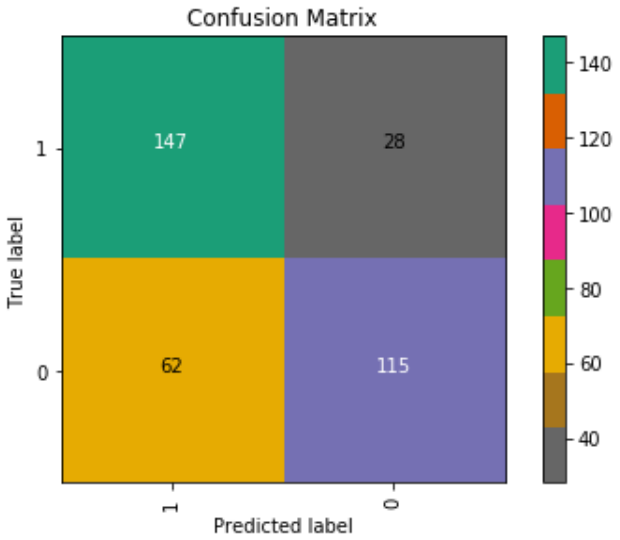

Figure-2. Confusion matrix of DTC for weather data.

From the Figure-2. We will get the true positive value as 147, true negative value as 115, false positive value as 62 and false negative value as28.

Table-1. Evolution metrics of DTC for weather dataset.

	precision	Recall	f1-score	Support
0	0.7	0.84	0.77	175
1	0.8	0.65	0.72	177
avg/total	0.75	0.74	0.74	352

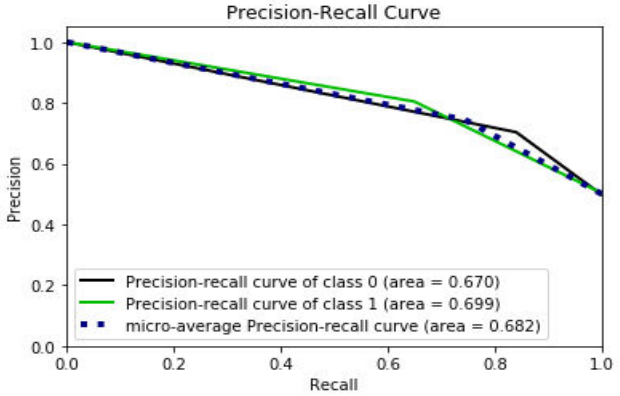

Figure-3. Precision-recall curve of DTC for weather data.

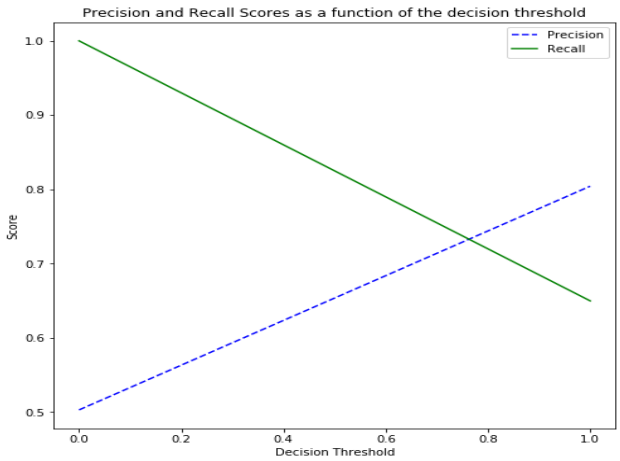

Figure-4. Precision-recall curve of DTC with threshold for weather data.

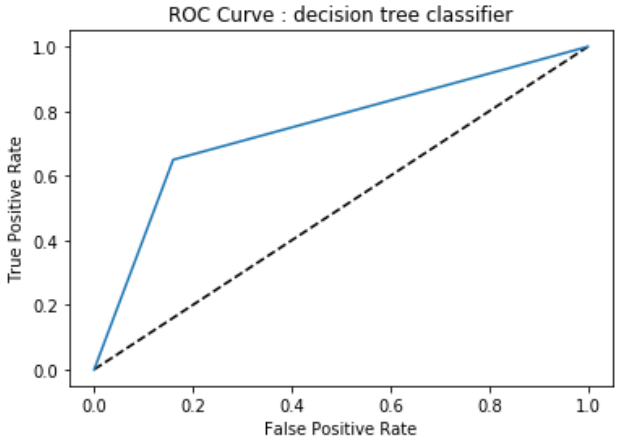

Figure-5. ROC curve of DTC for weather data.

Testing dataset's Confusion matrix for KNN is shown in Figure-6 and precision, recall, f-score and support of KNN is shown in Table-2. With the data of precision-recall from Table-2, a graph is drawn which is shown in Figure-7, precision-recall with threshold is shown in Figure 8 and ROC curve is shown in Figure-9.

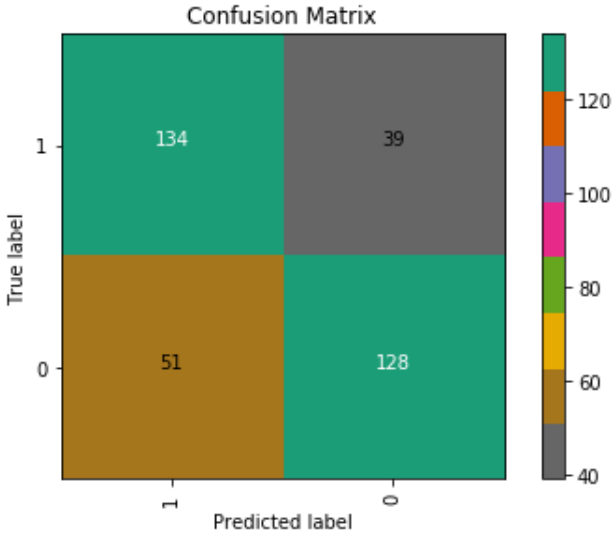

Figure-6. Confusion matrix of KNN for weather data.

From the Figure-6. We will get the true positive value as 134, true negative value as 128, false positive value as 51 and false negative value as 39.

Table-2. Evolution metrics of KNN for weather dataset.

	Precision	Recall	f1-score	Support
0	0.72	0.77	0.75	173
1	0.77	0.72	0.74	179
avg/total	0.75	0.74	0.74	352

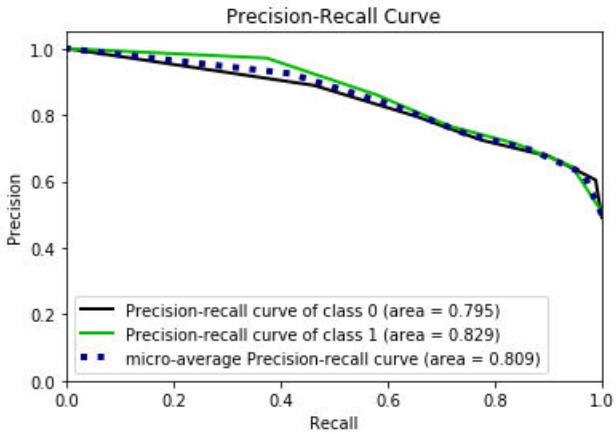

Figure-7. Precision-recall curve of KNN for weather data.

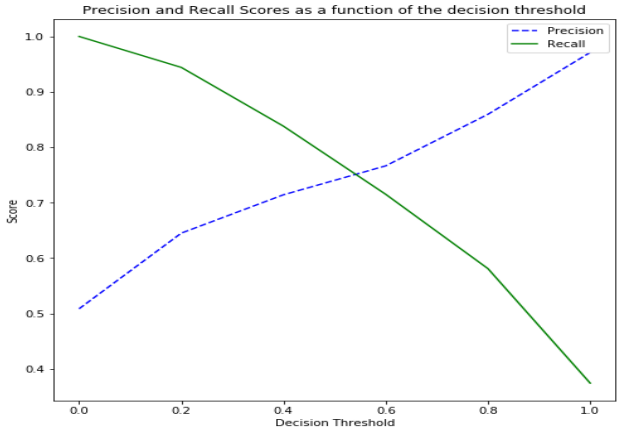

Figure-8. Precision-recall curve of KNN with threshold for weather data.

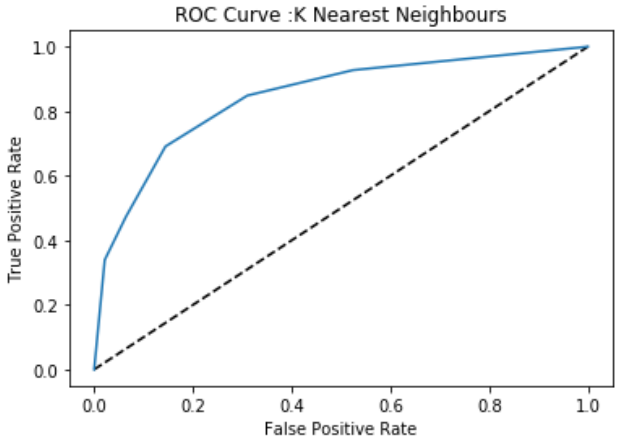

Figure-9. ROC curve of KNN for weather data.

Testing dataset's Confusion matrix for NAIVE BAYESIAN is shown in Figure-10 and precision, recall, f-score and support of NAIVE BAYESIAN is shown in Table-3. With the data of precision-recall from Table-3, a graph is drawn which is shown in Figure-11, precision-recall with threshold is shown in Figure-12 and ROC curve is shown in Figure-13.

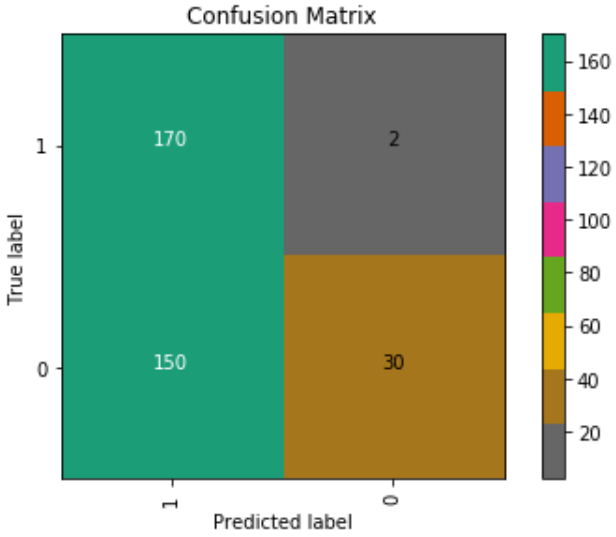

Figure-10. Confusion matrix of NAIVEBAYESIAN for weather data.

From the Figure-10. We will get the true positive value as 170, true negative value as 30, false positive value as 150 and false negative value as 2.

Table-3. Evolution metrics of NAIVE BAYESIAN for weather dataset.

	Precision	**Recall**	**f1-score**	**Support**
0	0.53	0.99	0.69	172
1	0.94	0.17	0.28	180
avg/total	0.74	0.57	0.48	352

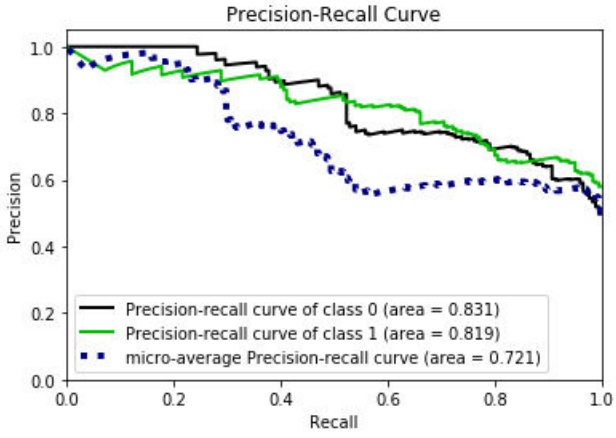

Figure-11. Precision-recall curve of NAIVE BAYESIAN for weather data.

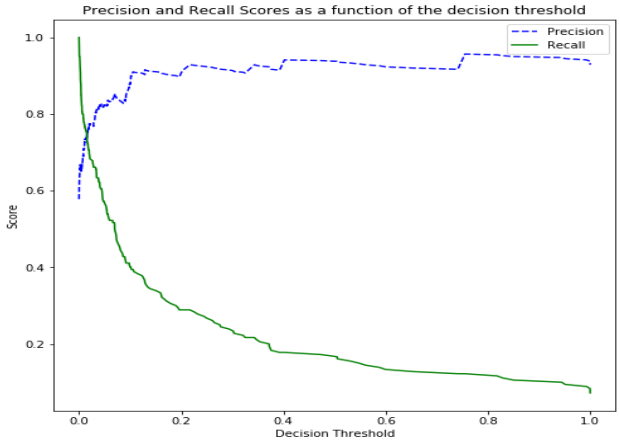

Figure-12. Precision-recall curve of NAVIE BAYESIAN with threshold for weather data.

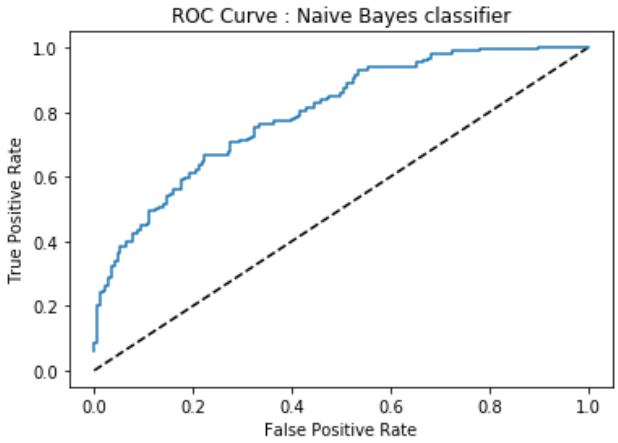

Figure-13. ROC curve of NAIVE BAYESIAN for weather data.

Testing dataset's Confusion matrix for RANDOM FOREST is shown in Figure-14 and precision, recall, f-score and support of RANDOM FOREST is shown in Table-4. With the data of precision-recall from Table-4, a graph is drawn which is shown in Figure-15, precision-recall with threshold is shown in Figure-16 and ROC curve is shown in Figure-17.

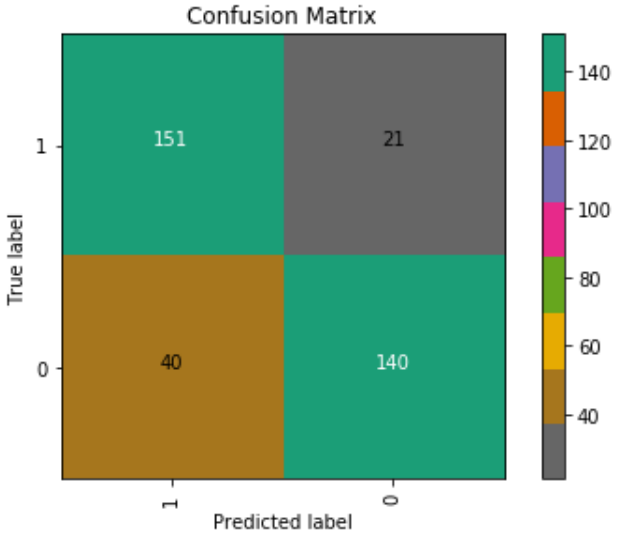

Figure-14. Confusion matrix of Random Forest for weather data.

From the Figure-14. We will get the true positive value as 151, true negative value as 140, false positive value as 40 and false negative value as 21.

Table-4. Evolution metrics of Random Forest for weather dataset.

	Precision	Recall	f1-score	Support
0	0.79	0.88	0.83	172
1	0.87	0.78	0.82	180
avg/total	0.83	0.78	0.82	352

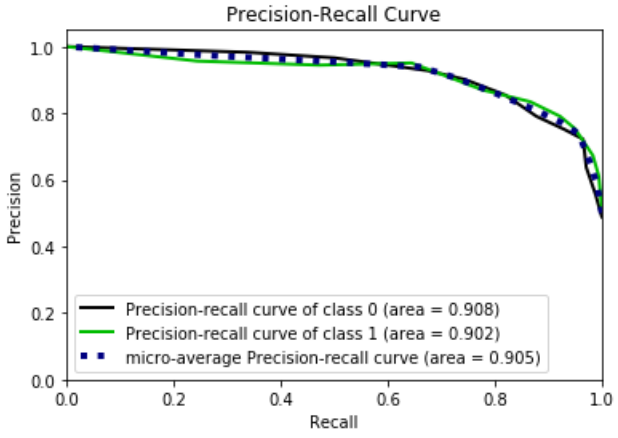

Figure-15. Precision-recall curve of Random Forest for weather data.

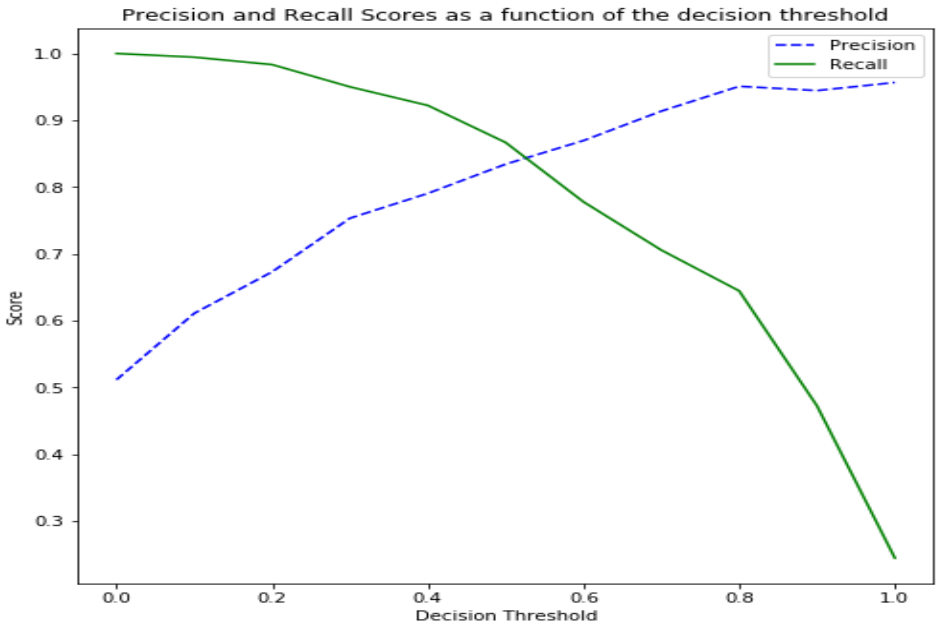

Figure-16. Precision-recall curve of Random Forest with threshold for weather data.

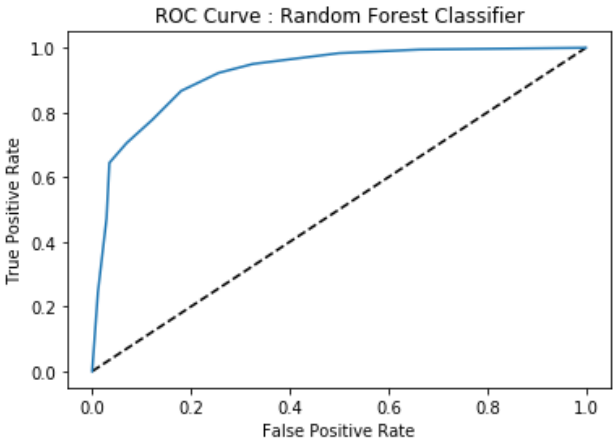

Figure-17. ROC curve of Random Forest for weather data.

Testing dataset's Confusion matrix for SVM is shown in Figure-18 and precision, recall, f-score and support of SVM is shown in Table-5. With the data of precision-recall from Table-5, a graph is drawn which is shown in Figure-19, precision-recall with threshold is shown in Figure-20 and ROC curve is shown in Figure-21.

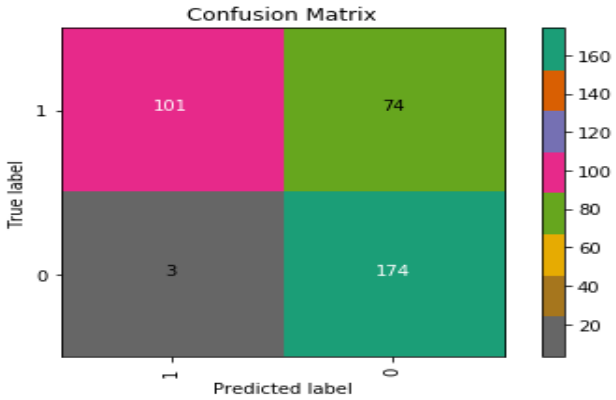

Figure-18. Confusion matrix of SVM for weather data.

From the Figure-18. We will get the true positive value as 101, true negative value as 174, false positive value as 3 and false negative value as 74.

Table-5. Evolution metrics of SVM for weather dataset.

	precision	**recall**	**f1-score**	**Support**
0	0.97	0.58	0.72	175
1	0.70	0.98	0.82	177
avg/total	0.84	0.78	0.77	352

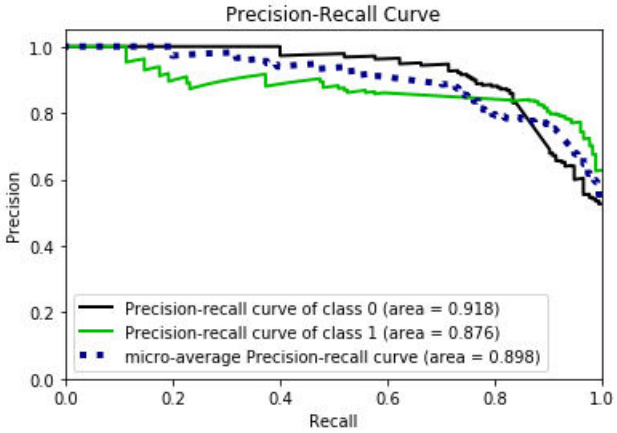

Figure-19. Precision-recall curve ofSVM for weather data.

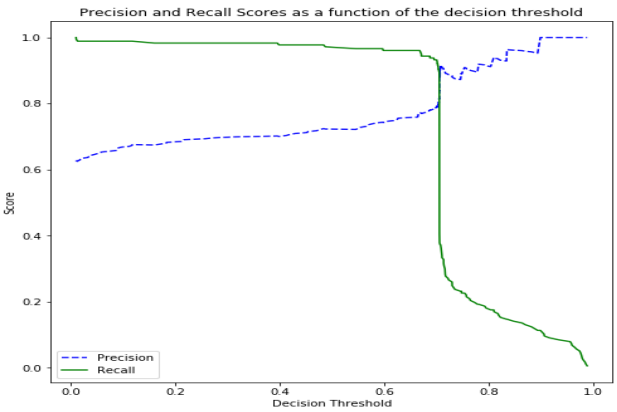

Figure-20. Precision-recall curve of SVM with threshold for weather data.

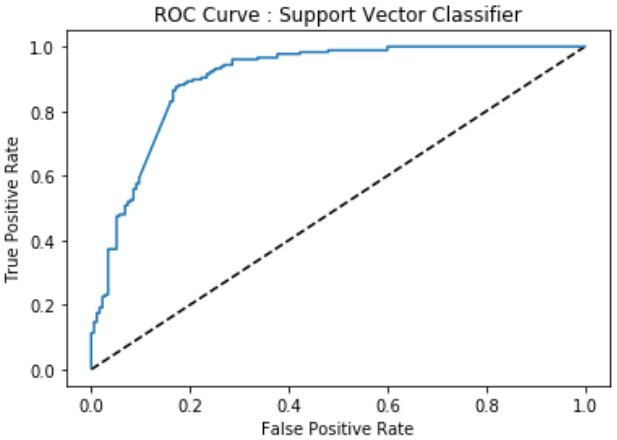

Figure-21. ROC curve of SVM for weather data.

Testing dataset's Confusion matrix for Logistic Regression is shown in Figure-22 and precision, recall, f-score and support of Logistic Regression are shown in Table-6. With the data of precision-recall from Table-6, a graph is drawn which is shown in Figure-23, precision-recall with threshold is shown in Figure-24 and ROC curve is shown in Figure-25.

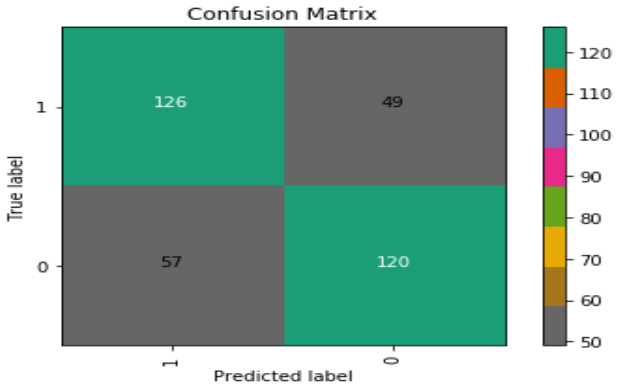

Figure-22. Confusion matrix of Logistic Regression for weather data.

From the Figure-22. we will get the true positive value as 126, true negative value as 120, false positive value as 57 and false negative value as 49.

Table-6. Evolution metrics of Logistic Regression for weather dataset.

	Precision	Recall	f1-score	Support
0	0.69	0.72	0.70	175
1	0.71	0.68	0.69	177
avg/total	0.70	0.70	0.70	352

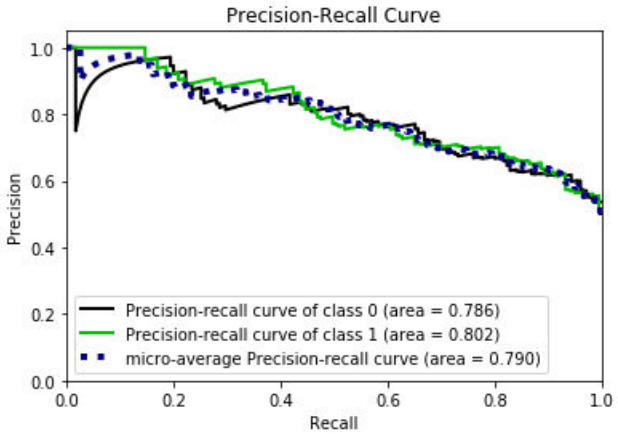

Figure-23. Precision-recall curve of Logistic Regression for weather data.

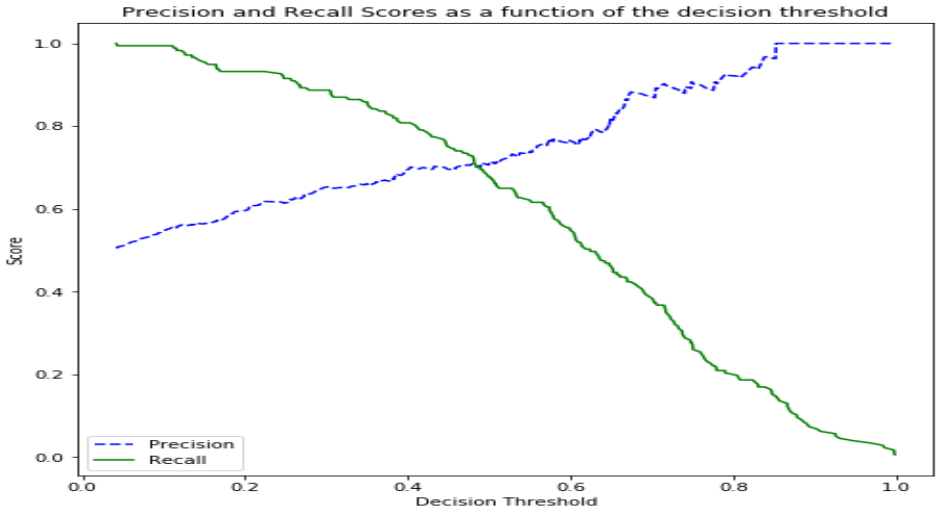

Figure-24. Precision-recall curve of Logistic Regression with threshold for weather data.

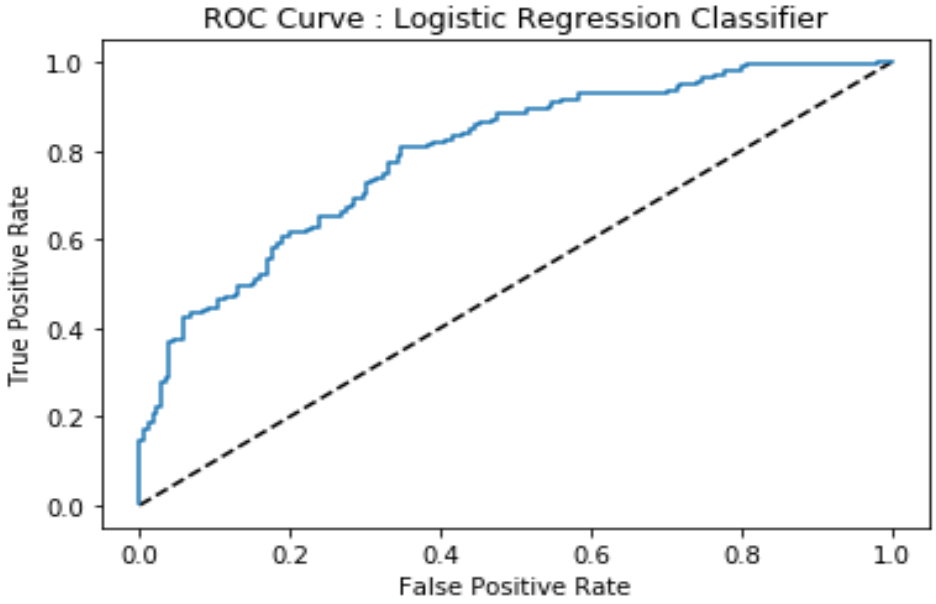

Figure-25. ROC curve of Logistic Regression for weather data.

The following Table-7 shows the accuracy, precision, recall, f1-score and support scores of Decision Tree, K-Nearest Neighbours, Naive Bayesian, Random Forest, Support Vector Machine, Logistic Regression algorithms.

Table-7. Accuracy, Precision, Recall and F1-Scores of Decision Tree, K-Nearest Neighbours, Naïve Bayesian, Random Forest, Support Vector Machine, Logistic Regression algorithms.

Method	Accuracy	Precision	Recall	F1-score
DTC	0.744318	0.75	0.74	0.74
KNN	0.744318	0.75	0.74	0.74
Naive Bayesian	0.568182	0.74	0.57	0.48
Random Forest	0.826705	0.83	0.78	0.82
SVM	0.78125	0.84	0.78	0.77
Logistic Regression	0.698864	0.70	0.70	0.70

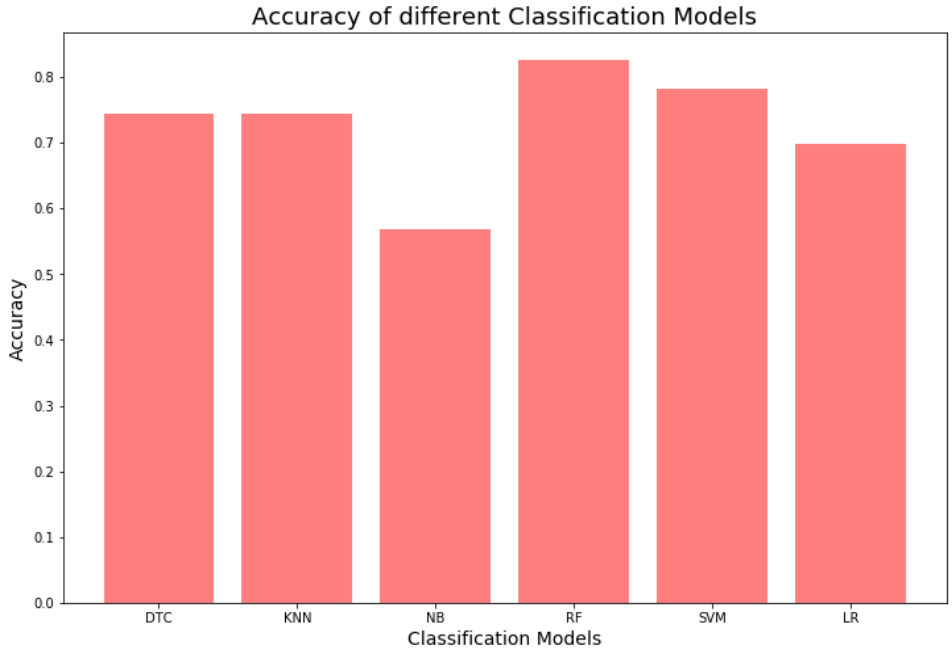

Figure-26. Accuracy graph for Decision Tree, K-Nearest Neighbours, Naive Bayesian, Random Forest, Support Vector Machine, Logistic Regression classification models.

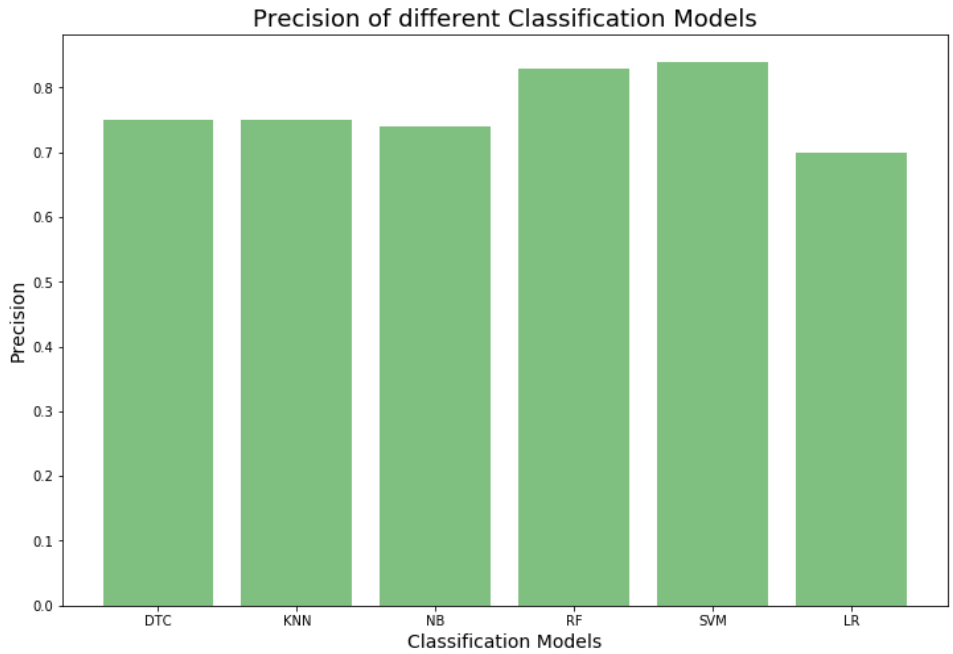

Figure-27. Precision graph for Decision Tree, K-Nearest Neighbours, Naive Bayesian, Random Forest, Support Vector Machine, Logistic Regression classification models.

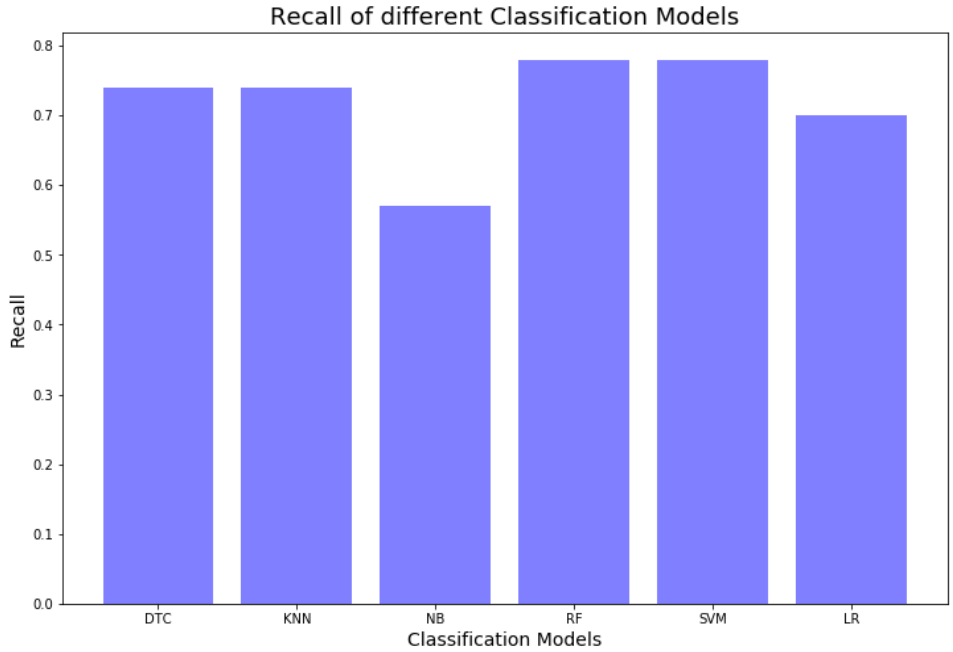

Figure-28. Recall graph for Decision Tree, K-Nearest Neighbours, Naive Bayesian, Random Forest, Support Vector Machine, Logistic Regression classification models.

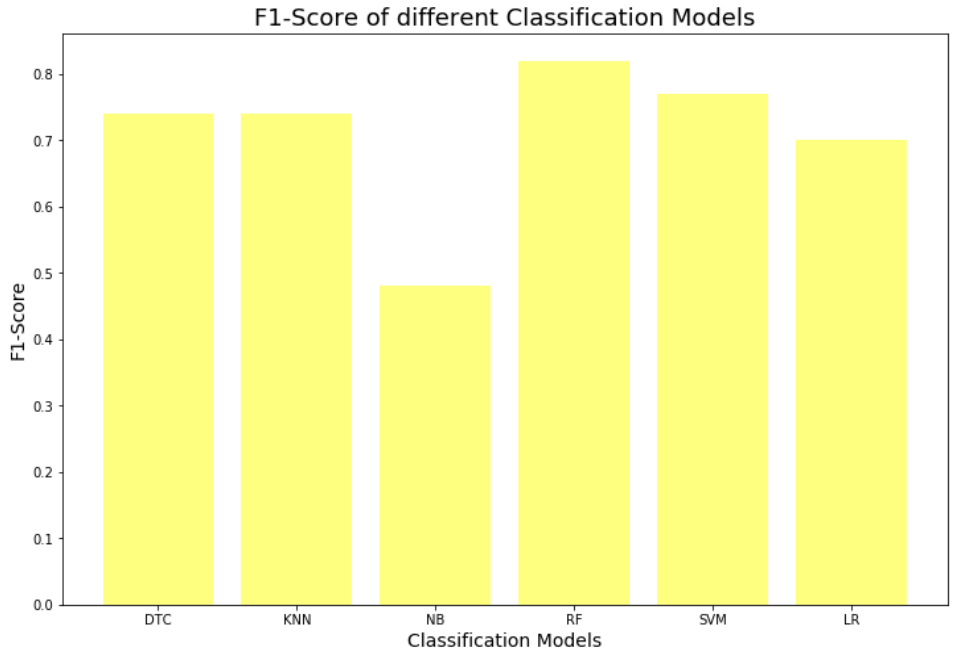

Figure-29. F1-Score graph for Decision Tree, K-Nearest Neighbours, Naive Bayesian, Random Forest, Support Vector Machine, Logistic Regression classification models.

5. CONCLUSIONS

In this paper, we designed a model to predict humidity in weather using weather data set and both supervised and unsupervised Machine learning algorithms, which gives best results based on accuracy. The results of the algorithms will be compared by using various evolution metrics like accuracy, precision, recall and confusion matrix. Based on the comparisons finally we conclude that ensemble learning method gives the more accurate predictions than other methods. The future work involves building a prediction model using Deep Neural Network.

REFERENCES

[1] Houthuys, Lynn, Zahra Karevan and Johan AK Suykens. 2017. Multi-view LS-SVM regression for black-box temperature prediction in weather

forecasting. International Joint Conference on Neural Networks (IJCNN).

[2] Chang Kuei-Chung, *et al*. 2018. Using Deep Learning Approaches to Predict Indoor Thermal and Outdoor Rainfall Probability by Embedded Weather Box. IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW).

[3] Aznarte José L. and Nils Siebert. 2017. Dynamic line rating using numerical weather predictions and machine learning: a case study. IEEE Transactions on Power Delivery. 32.1: 335-343.

[4] Andrade, José R and Ricardo J. Bessa. 2017. Improving renewable energy forecasting with a grid of numerical weather predictions. IEEE Transactions on Sustainable Energy. 8.4: 1571-1580.

[5] Koesdwiady Arief, Ridha Soua, and Fakhreddine Karray. 2016. Improving traffic flow prediction with weather information in connected cars: a deep learning approach. IEEE Transactions on Vehicular Technology. 65(12): 9508-9517.

[6] Gad Ibrahim and B. R. Manjunatha. 2017. Performance evaluation of predictive models for missing data imputation in weather data. International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[7] Rushika Ghadge*et al*. 2018. Prediction of crop yield using machine learning.5(2): 2237-2239.

[8] Choi Sun, *et al.* 2016. Prediction of weather-induced airline delays based on machine learning algorithms. Digital Avionics Systems Conference (DASC), IEEE.

[9] Tiwari, Soumva, Reza Sabzchgar, and Mohammad Rasouli. 2018. Short Term Solar Irradiance Forecast Using Numerical Weather Prediction (NWP) with Gradient Boost Regression. IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG).

[10] Sharma, Navin, *et al*. 2011. Predicting solar generation from weather forecasts using machine learning. IEEE International Conference on Smart Grid Communications (SmartGridComm)..

[11] Liu Zhao and Ziang Zhang. 2016. Solar forecasting by K-Nearest Neighbors method with weather classification and physical model. North American Power Symposium (NAPS).

[12] Sanders Sam, *et al.* 2017. Solar Radiation Prediction Improvement Using Weather Forecasts. International Conference on Machine Learning and Applications (ICMLA).

[13] Rusnák Rastislav and Rudolf Jakša. 2016. Stochastic weights and neurons selection in neural networks for weather prediction. International Symposium on Applied Machine Intelligence and Informatics (SAMI).

[14] Landset Sara, Michael F. Bergeron and Taghi M. Khoshgoftaar. 2017. Using Weather and Playing Surface to Predict the Occurrence of Injury in Major League Soccer Games: A Case Study. International Conference onInformation Reuse and Integration (IRI).

[15] Salman Afan Galih, Bayu Kanigoro, and Yaya Heryadi. 2015. Weather forecasting using deep learning techniques. International Conference on Advanced Computer Science and Information Systems (ICACSIS).

[16] Anandharajan T. R. V., *et al.* 2016. Weather Monitoring Using Artificial Intelligence. International Conference on Computational Intelligence and Networks (CINE).

[17] Navadia Sunil, *et al*. 2017. Weather prediction: A novel approach for measuring and analyzing weather data. International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC).