Advanced Machine Learning Approach for Weather Prediction

Mr.T.V. Mahendra
Assoc.Professor,NBKRIST

Mr.K.Raveendra Chaithanya
Asst.Professor,NBKRIST

Mr.V.Saicharan
Asst.Professor,NBKRIST

Abstract

Weather Prediction is a challenging task for all researchers of weather and the meteorological department. Many techniques are evolved in time for the prediction of weather since last many years. The advancement of science and tech- nology has helped the researchers to perform the prediction of weather simply and with less error rates. The soft computing techniques are the new technologies in computer science which are capable of making the weather prediction with promising output and less error rates. Weather prediction is done traditionally by use of many historical data in many models of physics. This prediction of weather is unsteady due to the change in weather condition. Because of change of weather system, the prediction is unstable. In this paper, we present different machine learning models that will make use of the historical data to train the models and then the model will be used to predict the weather whose accuracy is better than the traditional models. The evaluation of the models on the basis of accuracy shows that the models outperform and can be used as state-of-art technique to predict the weather in smarter way in less time.

Keywords Random forest · Decision tree · Support vector machine · Gradient boosting

1 Introduction

Climate modelling and weather prediction is the application of science and tech- nology to predict the state of the atmosphere for a given location is a challenging task for the researchers in this modern age. The use of machine learning for prediction of weather uses the dataset of 21 years having the parameters temp, dew, humidity, pressure, visibility and windspeed. The event of the place will be predicted by use of the models. The events may be “No Rain”, “Fog”, “Rain, Thunderstorm”, “Thun- derstorm”, “Fog, Rain”, etc. The machine learning models under consideration in this paper are random forest, decision tree, support vector machine, KNN, Adaboost, Xgboost, Gradient Boosting, naïve Bayes and logistic regression, etc. The evaluation of all these models are done on the basis of their performance compared as per their accuracy and f1 score.

2 Literature Review

Singh et al. [1] have made the weather forecasting using machine learning algo- rithms. They used different machine learning algorithms to predict the weather events. Khajure and Mohod [2] have deployed the future weather forecasting using soft computing techniques. Bhardwaj and Duhoon [3] have made use of soft computing techniques for forecasting of weather. Haghbin et al. [4] applied soft computing models for predicting sea surface temperature and made the review and assessment. Vathsala and Koolagudi [5] have used neuro-fuzzy model for quantified rainfall prediction using data mining and soft computing approaches. Balogh et al. [6] made a toy model investigate stability of AI-based dynamical systems. Jayasingh et al.
[7] have shown a novel approach for data classification using neural network. Litta et al. [8] have used artificial neural network model in prediction of meteorological parameters during premonsoon thunderstorms. Schultz et al. [9] have shown if deep learning beat numerical weather prediction. Sharma and Agarwal [10] have explained temperature prediction using wavelet neural network. Lin et al. [11] have discussed time series prediction based on support vector regression. Askari and Askari [12] have used time series grey system prediction-based models for gold price forecasting.

Lee and Lee [13] have constructed efficient regional hazardous weather prediction models through big data analysis. Jayasingh et al. [14] have made weather prediction using hybrid soft computing models. Sofian et al. [15] have done monthly rainfall prediction based on artificial neural networks with back propagation and radial basis function.

3 Methodology

In our research, we are trying to compare the different machine learning techniques for predicting the events. Machine learning models are designed to predict the events on the basis of the temp, dew, humidity, pressure, visibility and windspeed. In this research, we have taken the weather data from 1996 to 2017 to train our machine learning model. Here, we are using random forest, deceison tree, support vector regression, KNN, Adaboost, gradian-boost, Xgboost, naïve Bayes and logistic regres- sion to evaluate and train our model. The flow diagram of our proposed methodology is shown in Fig. 1.

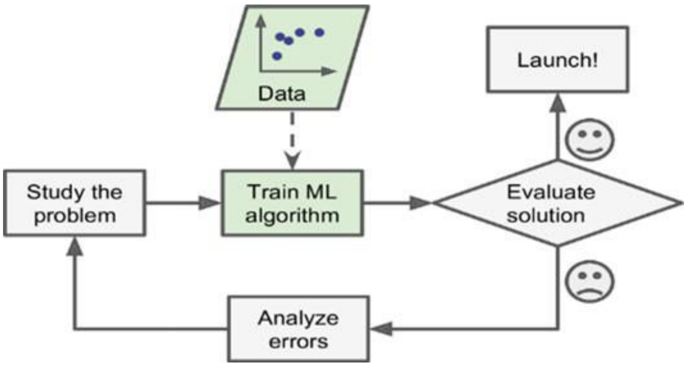


Fig. 1 The flow diagram of our proposed methodology

Initially, we preprocessed the dataset by eliminating any incomplete record. We apply feature selection and eliminate features that have no direct impact on the performance. Then we have to encode the categorical features into numerical feature because our machine learning algorithms will not be able to understand the cate- gorical value. Here when we try to encode the categorical feature into numerical using label encoder, our model not showed a good accuracy score that's why here we used get dummies method to encode the categorical features. We also checked that whether we have outliers or not but we did not find any outliers. And also we took a look at the multicollinearity and drop some highly co-related columns because if we keep some highly correlated features means it will affect our model accuracy and then we did some feature scaling.

Once our dataset is split using train test split on 80:20 ratio, the predictive models like linear regression, ridge and lasso regression, random forest, decision tree, support vector regressor. Adaboost, gradiantboost and xgboost are used to forecast upcoming match results. The machine learning process is shown in Fig. 2.

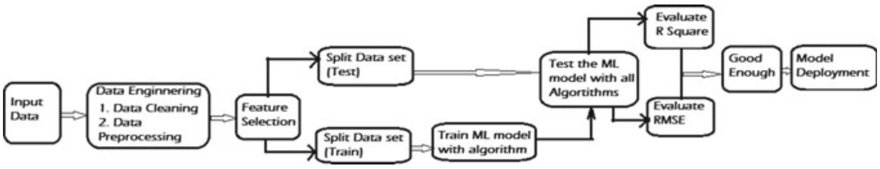


Fig. 2 Machine learning process overview

4 Techniques Used

K-nearest neighbours (KNN)

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

K-nearest neighbours (KNN) is a supervised learning method that may be used for both regression and classification, however it is usually utilized for classification. KNN aims to predict the right class of testing data given a set with various classes by calculating the distance between both the testing data and all the training points. It then chooses the k points that are the most similar to the test.

After the points have been chosen, the algorithm calculates the likelihood (in the classification phase) that the test point belongs to one of the *k* training point classes, and the class with the greatest probability is chosen. The regression model in a regression issue is the mean of the *k* chosen training points data as shown in Fig. 3.

Lazy learners are a phrase used to describe KNN algorithms. Eager learners refer the techniques such as Bayesian classification, logistic regression, SVM and others. These methods generalize over the training set before getting the test data that is the model is trained using training dataset and then test data is received by the model, and then predict/classify the test data.

With the KNN method, however, this is not the case. For the training set, it does not construct a generalized model; instead, it waits for the test data. Only when test data has been supplied it begins generalizing the training data in order to categorize

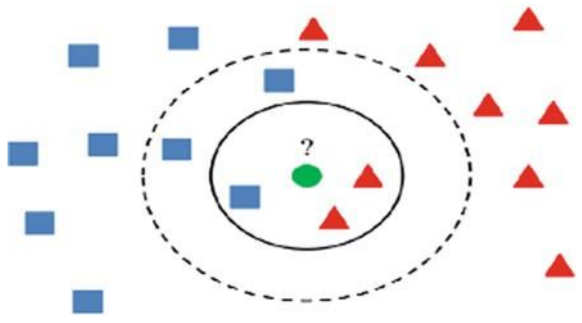
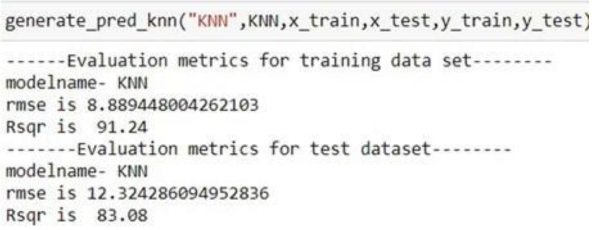


Fig. 3 K-nearest neighbours

the test data. As a result, a lazy learner just saves the training data and waits for the test set to arrive. Such algorithms are more effective at categorizing a particular test dataset than they are during training.

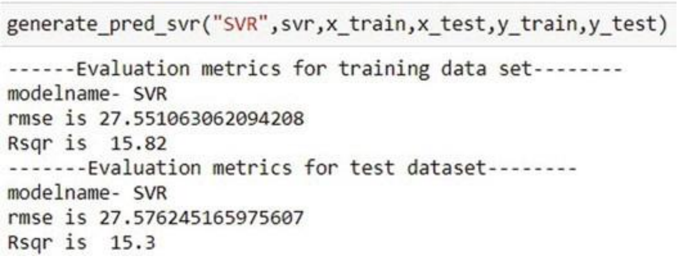


```
generate_pred_knn("KNN",KNN,x_train,x_test,y_train,y_test)

------Evaluation metrics for training data set--------
modelname- KNN
rmse is 8.889448004262103
Rsqr is  91.24
-------Evaluation metrics for test dataset--------
modelname- KNN
rmse is 12.324286094952836
Rsqr is  83.08
```

Support Vector Machine

By the computer hardware advancements, as well as the increasing availability and low cost of technology elements such as RAM and GPU, a substantial chunk of labelled data is readily available and created on a regular basis. To make the firm less sensitive to unanticipated situations, we focus on predictive data processing. The support vectir machine as shown in Fig. 4 is used for the pupose of classification.



```
generate_pred_svr("SVR",svr,x_train,x_test,y_train,y_test)

------Evaluation metrics for training data set--------
modelname- SVR
rmse is 27.551063062094208
Rsqr is  15.82
-------Evaluation metrics for test dataset--------
modelname- SVR
rmse is 27.576245165975607
Rsqr is  15.3
```

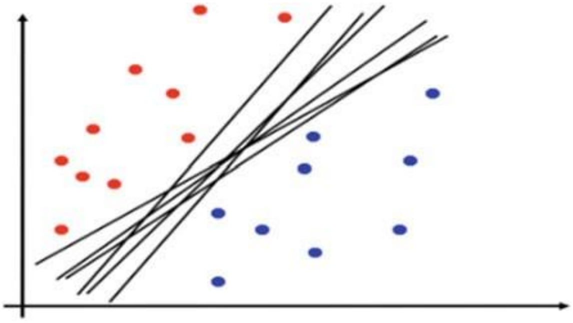
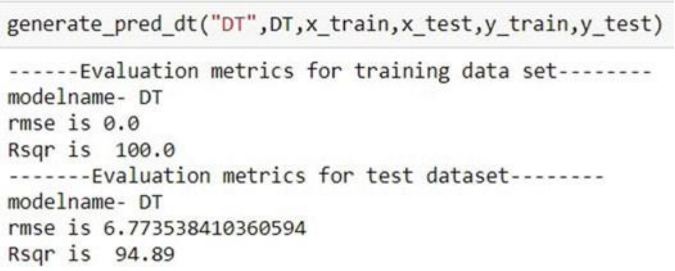


Fig. 4 Support vector machine

Decision Tree

A decision tree accepts an item or scenario with a set of attributes as input and provides a yes/no “decision”. It is one of the best used effective techniques of machine learning. We explain how to acquire a good hypothesis by first describing the depiction hypothesis space.



```
generate_pred_dt("DT",DT,x_train,x_test,y_train,y_test)
```

```
------Evaluation metrics for training data set--------
modelname- DT
rmse is 0.0
Rsqr is  100.0
-------Evaluation metrics for test dataset--------
modelname- DT
rmse is 6.773538410360594
Rsqr is  94.89
```

Each node examines the value of one or more input attributes. Leaf nodes give the values to be delivered if that leaf is encountered. Branches from the node correspond to possible attribute values in the **Decision Tree.**

Random Forests

It is a collection of decision trees that have been bagged trained as shown in Fig. 5. The strategy generates more tree variety, which trades a higher bias for fewer switches,

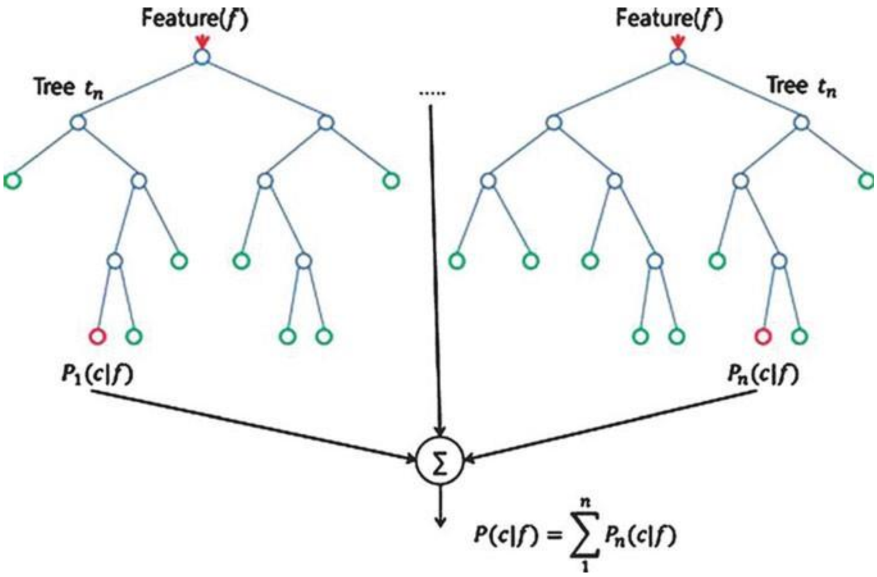
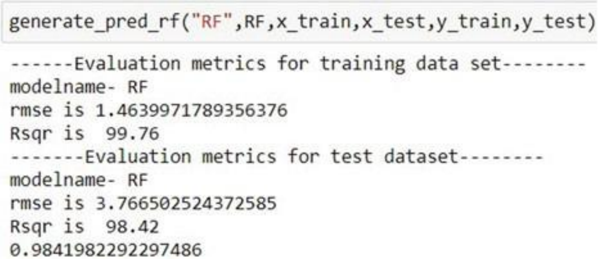


Fig. 5 Random forest

resulting in a more robust overall model. The random forest regression that comes after it is similar to the bagging regression that came before it.



```
generate_pred_rf("RF",RF,x_train,x_test,y_train,y_test)

------Evaluation metrics for training data set--------
modelname- RF
rmse is 1.4639971789356376
Rsqr is  99.76
-------Evaluation metrics for test dataset--------
modelname- RF
rmse is 3.766502524372585
Rsqr is  98.42
0.9841982292297486
```

XGboost

It stands for "Extreme Gradient Boosting". Boost is a toolkit for distributed gradient boosting that has been optimized for speed, versatility and portability. Machine learning algorithms are built using the Gradient Boosting framework as shown in Fig. 6. It employs parallel tree boosting to quickly and accurately solve a number of data science problems.

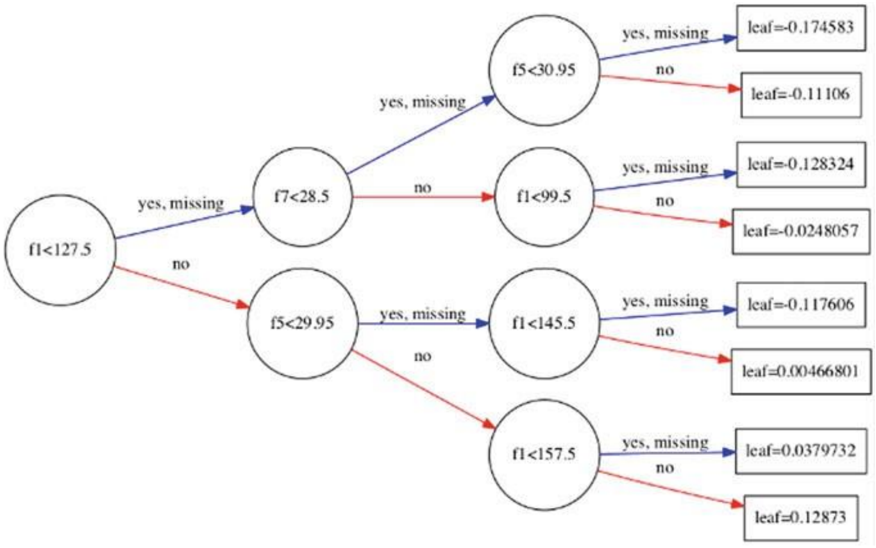


Fig. 6 Xgboost

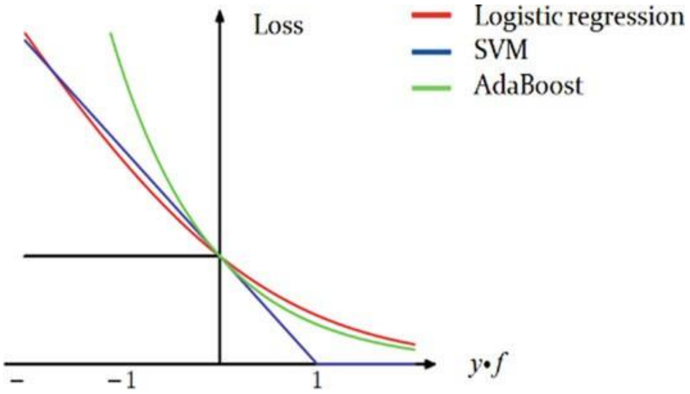
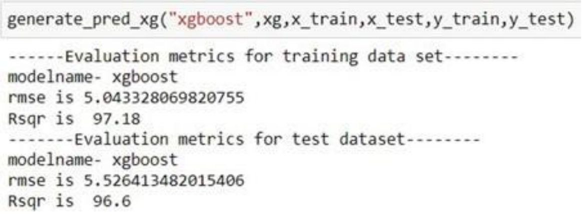


Fig. 7 Adaboost

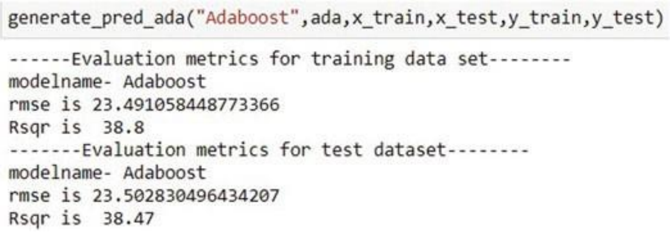


```
generate_pred_xg("xgboost",xg,x_train,x_test,y_train,y_test)

------Evaluation metrics for training data set--------
modelname- xgboost
rmse is 5.043328069820755
Rsqr is  97.18
-------Evaluation metrics for test dataset--------
modelname- xgboost
rmse is 5.526413482015406
Rsqr is  96.6
```

Adaboost

Showing genuine Freund and Robert Schapire developed Adaboost (Adaptive Boosting), a statistical categorization method for which they were given the Mathe- maticians Award in 2003. It can be combined with a number of learning algorithms to improve results. The output of other learning algorithms (“learning and self”) is combined into a weighted sum that reflects the proper outcome of the boosted classifier as shown in Fig. 7.



```
generate_pred_ada("Adaboost",ada,x_train,x_test,y_train,y_test)

------Evaluation metrics for training data set--------
modelname- Adaboost
rmse is 23.491058448773366
Rsqr is  38.8
-------Evaluation metrics for test dataset--------
modelname- Adaboost
rmse is 23.502830496434207
Rsqr is  38.47
```

Gradient Boosting

A prominent boosting technique is gradient boosting. Each predictor in gradient boosting corrects the mistake of its preceding as shown in Fig. 8. With exception of

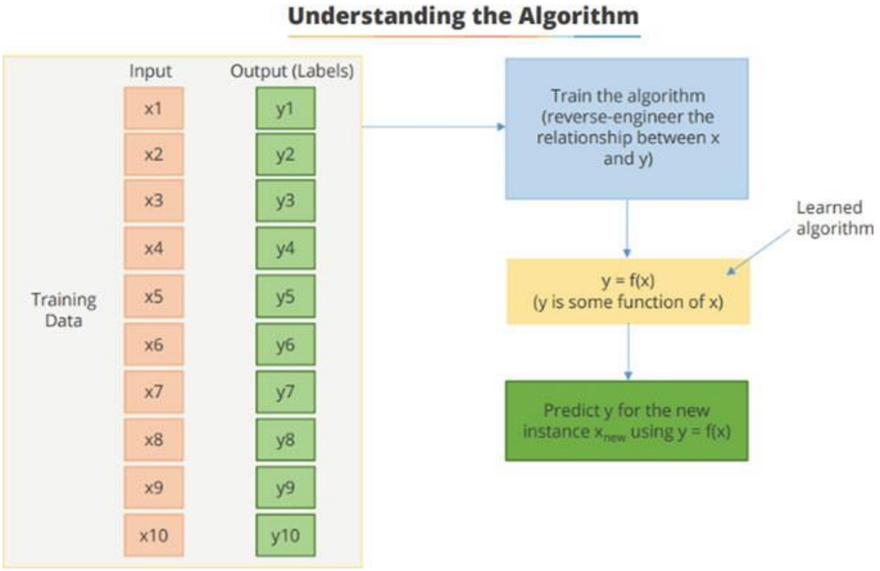
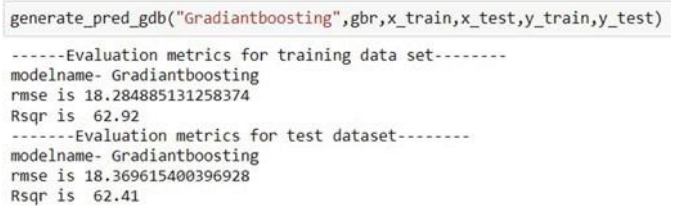

Understanding the Algorithm
Input
Output (Labels)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10
y1
y2
y3
y4
y5
y6
y7
y8
y9
y10
Training Data
Train the algorithm (reverse-engineer the relationship between x and y)
Learned algorithm
y = f(x) (y is some function of x)
Predict y for the new instance x_new using y = f(x)

Fig. 8 Gradient boosting

Adaboost, the training instance parameters are not adjusted; instead, each predictor is trained using the immediate predecessor residual as labels.

CART is the foundation learner in a method called Gradient Boosted Trees (Classification and Regression Trees).



```
generate_pred_gdb("Gradiantboosting",gbr,x_train,x_test,y_train,y_test)

------Evaluation metrics for training data set--------
modelname- Gradiantboosting
rmse is 18.284885131258374
Rsqr is  62.92
-------Evaluation metrics for test dataset--------
modelname- Gradiantboosting
rmse is 18.369615400396928
Rsqr is  62.41
```

5 Data Analysis

Features description

The weather data is collected for prediction analysis for 21 years (1996–2017) and stored in a dataset. There are 7 variables and 7750 instances in the dataset used as shown in Table 1.

Table 1 Dataset used in the experiment

Temp		Dew	Humidity	Pressure	Visibility	Wind	Event
0	28	24	76	1002	5	11	No rain
1	29	26	85	1003	5	8	No rain
2	32	26	78	1004	5	11	No rain
3	31	26	81	1003	4	13	No rain
4	31	26	86	1001	4	10	Rain, thunderstorm

Table 2 Accuracy of different models

S. No.	Model name	Accuracy
1	Random forest	79.52
2	Decision tree	71.23
3	Support vector machine	59.33
4	KNN	77.86
5	Adaboost	71.43
6	Xgboost	79.94
7	Gradient boosting	81.67
8	Naïve Bayes	73.09
9	Logistic regression	78.14

6 Result Analysis

The different machine learning models were trained with the weather data of 21 years and the model was used to predict the test data. Hence, the performance of all the models based on their accuracy is summarized in Table 2. The graphical comparison of accuracy of these models is shown in Fig. 9.

The different machine learning models were trained with the weather data of 21 years and the model was used to predict the test data. Hence, the performance of all the models based on their $f1$ score is summarized in Table 3 and the graphical comparison of $f1$ score of these models is shown in Fig. 10.

7 Conclusions and Future Work

We can estimate the weather events using a machine learning model that takes into account the different weather parameters. In this paper, we presented different machine learning models which can be used for prediction of weather with much simpler and easier way than the physical models. The accuracy evaluation of the models shows that the machine learning models perform better than the traditional

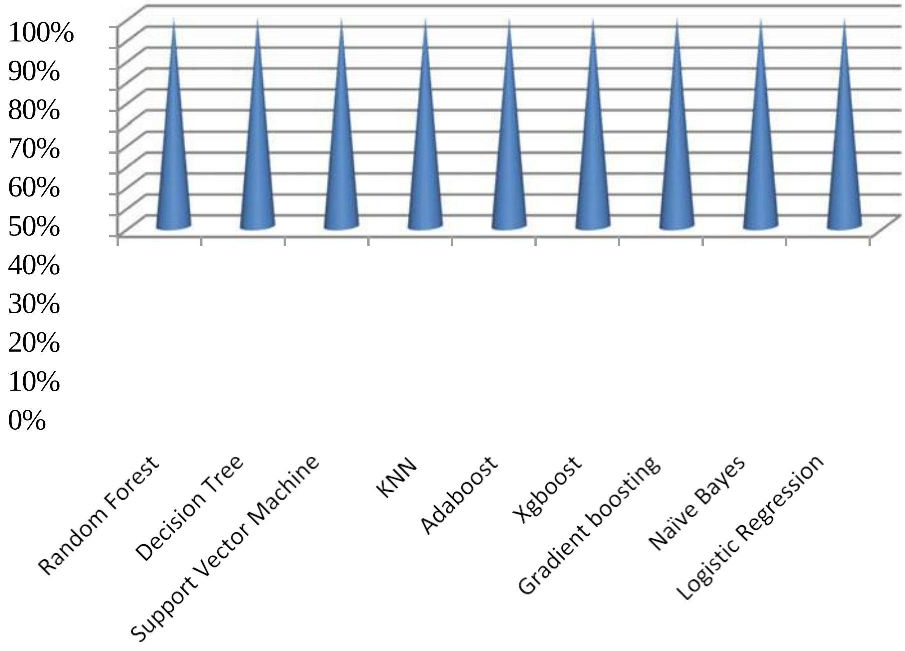


Fig. 9 The graphical comparison of accuracy of different models

Table 3 *F*1 score of different models

S. No.	Model name	*F*1 score
1	Random forest	0.81
2	Decision tree	0.75
3	Support vector machine	0.74
4	KNN	0.80
5	Adaboost	0.74
6	Xgboost	0.80
7	Gradient boosting	0.83
8	Naïve Bayes	0.79
9	Logistic regression	0.80

models. These models made use of the dataset collected from predefined recourses in which the maximum accuracy is observed upto 81.67%. In future, is is planned to use the different IoT devices to collect the accurate data so that the data set to be used in the model will be more exact and accordingly the performance of the model will be more correct.

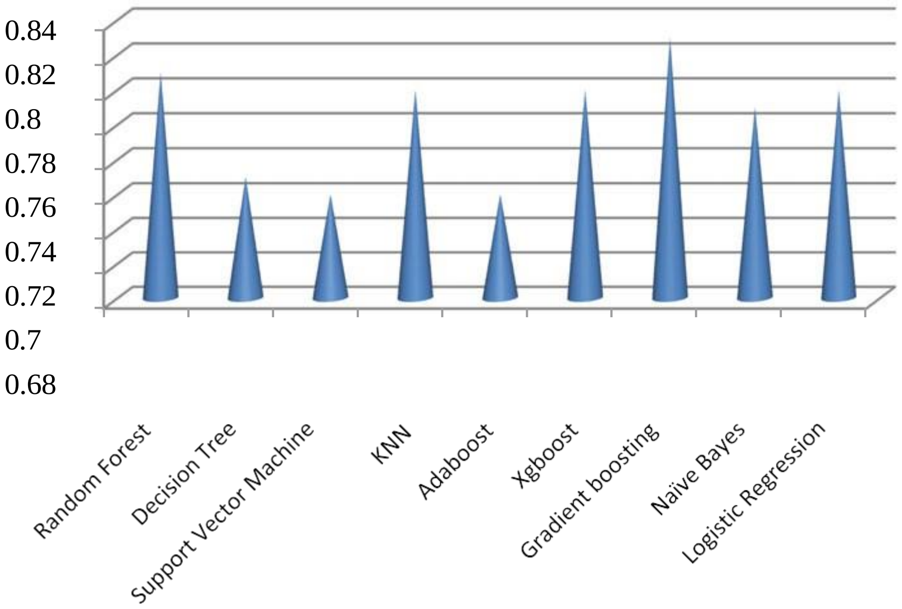


Fig. 10 The graphical comparison of *F*1 score of different models

References

1. Khajure S, Mohod SW (2016) Future weather forecasting using soft computing techniques. Proc Comput Sci 78:402–407. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2016.02.081
2. Bhardwaj R, Duhoon V (2018) Weather forecasting using soft computing techniques. In: Inter- national conference on computing, power and communication technologies (GUCON), pp 1111–1115. https://doi.org/10.1109/GUCON.2018.8675088
3. Vathsala H, Koolagudi SG (2021) Neuro-fuzzy model for quantified rainfall prediction using data mining and soft computing approaches. IETE J Res. https://doi.org/10.1080/03772063. 2021.1912648
4. Balogh B, Saint-Martin D, Ribes A (2021) A toy model to investigate stability of AI-based dynamical systems. Geophys Res Lett 48(8). https://doi.org/10.1029/2020GL092133
5. Jayasingh SK, Gountia D, Samal N, Chinara PK (2021) A novel approach for data classification using neural network. IETE J Res. https://doi.org/10.1080/03772063.2021.1986152
6. Schultz MG, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen LH, Mozaffari A, Stadtler S (2021) Can deep learning beat numerical weather prediction. Phil Trans R Soc A 379:20200097. https://doi.org/10.1098/rsta.2020.0097
7. Sharma A, Agarwal S (2012) Temperature prediction using wavelet neural network. Res J Inf Technol 4:22–30. https://doi.org/10.3923/rjit.2012.22.30. https://scialert.net/abstract/?doi=rjit. 2012.22.30

8. Lin S, Wang G, Zhang S, Li J (2006) Time series prediction based on support vector regres- sion. Inf Technol J 5:353–357. https://doi.org/10.3923/itj.2006.353.357. https://scialert.net/abs tract/?doi=itj.2006.353.357
9. Lee J, Lee J (2016) Constructing efficient regional hazardous weather prediction models through big data analysis. IJFIS 16:1–12. https://doi.org/10.5391/IJFIS.2016.16.1.1
10. Jayasingh SK, Mantri JK, Pradhan S (2021) Weather prediction using hybrid soft computing models. In: Udgata SK, Sethi S, Srirama SN (eds) Intelligent systems. Lecture notes in networks and systems, vol 185. Springer, Singapore. https://doi.org/10.1007/978-981-33-6081-5_4