Survey on Rain Prediction using Machine Learning

Vijithra Nair[1], Megha Mathew[2], Sweta Bhattacharjee[3], Arashdip Singh[4], Prof. Payel Thakur[5]

[1,2,3,4]UG Student, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India
[5] Assistant Professor, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India

---

Abstract - *As agriculture being the key point of survival, Rainfall is the important source for its cultivation. Rainfall prediction has always been a major problem as prediction of rainfall gives awareness to people and to know in advance about rain so as to take necessary precautions to protect their crops from rain. A particular dataset is taken from Kaggle community and this project predicts whether it will rain tomorrow or not by using the rainfall in dataset. CatBoost model is implemented in this project as it is an open sourced machine learning algorithm, and features great quality without the parameter tuning, categorical feature support, improved accuracy and fast prediction. CatBoost model is a gradient boosting toolkit and two critical algorithms classical and innovative are introduced to create a fight in prediction shift present in currently existing implementations of gradient boosting algorithms. CatBoost performed very well giving an AUC (Area under curve) score 0.8 and ROC (Receiver operating characteristic curve) score as 89. ROC is called as an evaluating curve whereas AUC presents a degree or measure of separability as the model is skilled enough to distinguish between classes. An Exploratory data analysis is done to examine data distribution, outliers and provides tools for visualizing and understanding the data through graphical representation. A dashboard is implemented to showcase the information that is represented in datasets i.e. any changes in the data will result in different types of graphs. A linear SVC (Support vector classifier) provides a best fit hyperplane that divides the data and feeds some features to the classifier to detect what the predicted class is and results in desired output.*

***Key Words*: ARIMA, CatBoost, Random Forest, Rainfall prediction, XgBoost**

1. INTRODUCTION

The world's welfare is agriculture. The achievement of agriculture is dependent on rainfall. It also helps with water resources. Rainfall information in the past helps farmers better manage their crops, leading to economic growth in the country. Prediction of precipitation is beneficial to prevent flooding that saves people's lives and property. Fluctuation in the timing of precipitation and its amount makes forecasting of rainfall a problem for meteorological scientists. To overcome these problems, a machine learning technology is used which is predictive analysis that is a branch of data mining which predicts the future probabilities and trends.

Prediction is the phenomenon of knowing what may happen to a system in the near future. Since rainfall is the major causes of calamities like floods and typhoons, predicting the occurrence of rainfall will help us to be prepared for these calamities. The basic procedures involved are first identifying an initial model, second repeatedly changing the model by removing a predictor variable based on a criteria and then terminating the process when a model which fits the data well. Here, various rainfall prediction projects were developed using multiple linear regression and other models.

This proposed method uses Australian meteorological dataset to predict the rain fall. Usually machine learning algorithms are classified into two major categories i.e. unsupervised learning and supervised learning. All of the clustering algorithms come under the supervised machine learning. Even though many models have developed, it is necessary for doing research using machine learning algorithms to get accurate predictions. The error free prediction provides better planning in the agriculture and other industries, Henceforth we have used the CatBoost model for faster and accurate prediction

2. LITERATURE SURVEY

A. CatBoost Model

According to Shihab Ahmad Shahria this model follows a technique that is additionally divided into four parts: (a) data pre-processing that involves the gathering of pollutants and other alternative meteorological data alongside with correction of missing values; (b) analysis associated with relationship between meteorological parameters or variables which of the pollutants; (c) feature importance involves the screening of meteorological parameters and conjoints the pollutants in air right before the operation on the models; (d) the application of models like ARIMA-ANN, ARIMA-SVM, PCR, DT and CatBoost.[1]

B. XgBoost Model

According to Nikhil Tiwari, Generally, XGBoost is fast when compared to other implementations of gradient boosting. Szilard Pafka performed few benchmarks comparing the performance of XGBoost to different gradient boosting techniques and bagged decision trees. XGBoost focuses on structured and tabular datasets on classification and regression based on its predictive modelling problems. The result is the algorithm for competition winners on the Kaggle competitive data science platform. Gradient boosting is an approach where new models are created that predict the residuals or errors of previous models and then summed up together to make the final prediction. It is called so because it uses a gradient descent algorithm to minimize the loss when new models are added. This approach supports both regression as well as classification predictive modelling problems.[2]

C. Random Forest

Urmay shah's Random forest is a tree based model, it is a collection of many tree models. Different tuning parameters are used for tuning the model. In random forest one of the parameters shows exactly how many trees should be more used to get the accurate results. Random forest works really well with high variant and low biasing models. It is observed that after 250 number trees error rate doesn't change. Hence 250 number of trees are restricted to in the forest. Random forest method is excellent for light rain predictions as it gives the best accuracy. It also performs well for the no rain, moderate rain, and for light rain.[3]

D. ARIMA Model

CMAK Zeelan Basha states ARIMA MODEL (AutoRegressive Integrated Moving Average) is used for time series prediction and analysis and forecasting. It contains four methods and is proposed by Box and Jenkins. The following are the four steps used in the ARIMA model.

Stage-1: Identification of a series of responses is done in the first stage which is used in calculating the time series and autocorrelations using statement IDENTIFY

Stage-2: In this stage Estimation of the previously identified variables is done and also the parameters are estimated using the statement ESTIMATE.

Stage-3: Diagnostics correction of the above gathered variables and parameters are all implemented in this stage. Stage4: In this stage the predicting values of time series are forecasted which are future values, using the ARIMA model using the statement FORECAST. The parameters used in this model are p,d,q which describes 'p' as the number of lag observations, 'q' as the degree of differencing and' as the moving average order.[4]

2.1 SUMMARY OF RELATED WORK

The summary of methods used in literature is given in Table 1.

Table -1: Sample Table format

Table 1 Summary of literature survey

Literature	ARIMA	CatBoost	Hybrid
Shihab Ahmad Shahria et al. 2021[1]	Yes	Yes	Yes
Nikhil Tiwari et al. 2020 [2]	No	Yes	No
Urmay Shah et al. 2018 [3]	Yes	No	No
CMAK Zeelan Basha et al. 2020 [4]	Yes	No	Yes

The overview of comparison of different parameters is given in Table 2.

Table 2 Summary of literature survey

Literature	Performance Parameters
Shihab Ahmad Shahria et al. 2021[1]	CatBoost,ARIMA-ANN, ARIMA-SVM, DT
Nikhil Tiwari et al. 2020 [2]	Neural Networks, Support Vector Regressor (SVR), Elastic Net , Ridge Regression , Lasso Regression , Linear Regression , XGBoost , Random Forest , Bagging Regressor, Gradient Boosting Regressor
Urmay Shah et al. 2018 [3]	ARIMA(Auto-Regressive Integrate d Moving Average), SVM(Simple Moving Average Model), Decision Tree, Holt Winter, , Random Forest, Neural Network method ,Seasonal Naive method
CMAK Zeelan Basha et al. 2020 [4]	ARIMA Model(Auto-Regressive Integrate d Moving Average), Artificial Neural Network, Logistic Regression, Support Vector Machine and Self Organizing Map

3. PROPOSED WORK

The Proposed system consists of using the CatBoost Algorithm. It performs very well giving an AUC (Area under curve) score 0.8 and ROC(Receiver operating characteristic curve) score as 89.ROC is evaluating curve and AUC presents degree or measure of separability as this model is capable of distinguishing between classes. An Exploratory data analysis is done to examine data distribution, outliers and provides tools or visualizing and understanding the data through graphical representation.

3.1 SYSTEM ARCHITECTURE

The system architecture is given in Figure 1. Each block is described in this Section.

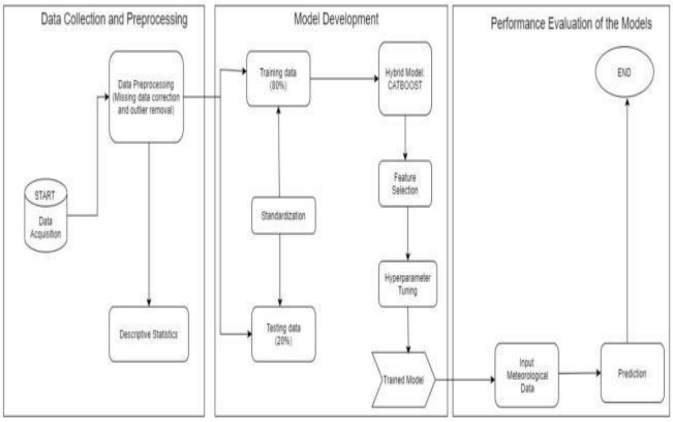


Fig. 1 Proposed system architecture

A. Input Block Description:

1. Prediction: It takes in various input parameters from the user so as to predict whether a certain day would be sunny or rainy.

2. Accuracy: It takes in many input parameters such as temperature, evaporation, wind speed, pressure, humidity and location so as to get the best accurate results. Thus the aim is to provide with accurate result in order to get correct prediction of weather for future, so that in critical conditions people can be aware of upcoming natural calamities.

3. Quick Result: It guarantees the quick result due to the use of advanced technology which was not in the earlier manual way of calculating all the parameters by hand.

B. Block 1 Description:

1. Data Acquisition: Collecting data for training the Machine Learning model is the basic step in the machine learning pipeline. These predictions can only be as good as depending on the data on which they have been trained.

2. Data pre-processing: Real-world raw data and images are often incomplete, inconsistent and lacking in certain behaviors or trends. They may also contain many errors. So once the errors are collected, they are pre-processed into a format such that the machine learning algorithm can be used for the model.

3. Descriptive Statistics: It provides in detail the summarizing information about the whole characteristics and the distribution of values in more than one dataset. Whereas classical descriptive statistics allow analysts to have a overall look at the central tendency and the degree of dispersion values that are present in datasets. They are useful in understanding of data distribution and in comparing data distributions.

B. Block 2 Description (Model Deployment):

The second block is the model development. Prediction of everyday rainfall is important for flood forecasting, reservoir operation, and other hydrological applications. The artificial intelligence algorithm is used for stochastic forecasting rainfall which is not good enough of simulating unseen extreme rainfall events which become common due to climate change. Here, the date is trained to follow a certain trend at about 80% and tested about 20% which makes it a complete of 100% following basic standardization. The trained data acts as an input to the hybrid model used which is CatBoost which undergoes feature selection i.e. sorting out different types based of fixed criteria. Then the trained data goes into hyper parameter tuning. In order to select a model, one should know the most suitable hyper parameters to progress. The proposed method can automatically optimize the hyper parameters of CatBoost for precipitation modelling and prediction. Finally, the trained data gets converted into trained model to be sent for performance evaluation.

C. Block 3 Description (Performance evaluation of the models):

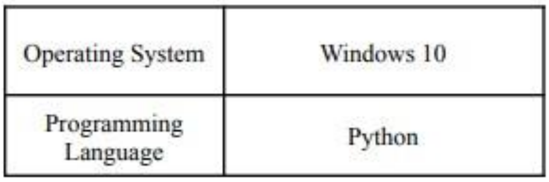
The third block is based on the performance evaluation of the models. Model evaluation projects to estimate the generalization accuracy of a model on future (unseen/out of-sample) data. Methods for evaluating the performance of a model are divided into 2 categories: holdout and Cross-validation. Both methods use a test set (i.e. data which is not seen by the model) to evaluate model performance. It is not advisable to use the data we used to build the model to evaluate it. This is because the model will remember the whole training set, and will predict the correct label for any point in the training set and this is called as overfitting. This involves randomly dividing datasets into three subsets. As training set is a subset of the dataset which is used to build predictive models. Validation set is also a subset of the dataset which assess the performance of the model built in the training phase itself. One of the features it provides is giving a test platform for fine tuning a model's parameters and from it selecting the most appropriate performing model. Every modelling algorithms doesn't necessarily need a validation set. Test set/ unseen data is a subset of the dataset that is used to likely assess the further future performance of the model.

Overfitting occurs when a model fits to the training set much better than it fits the test set. In rainfall prediction, the input is meteorological data like wind speed, pressure, temperature, humidity, date and location which then undergoes through CatBoost model predictor where it will give the output being if it will rain the next day or if it will be a sunny day. Thus, ending the execution of the model helps in obtaining a final and accurate result.

3.2 REQUIREMENT ANALYSIS

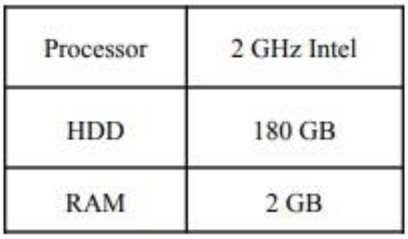
The implementation detail is given in this section.

3.2.1 Software



Operating System	Windows 10
Programming Language	Python

3.2.1 Hardware



Processor	2 GHz Intel
HDD	180 GB
RAM	2 GB

3.3 DATASET AND PARAMETERS

Dataset from Kaggle contains of about 10 years of daily weather observations from various different locations across Australia. Prediction of next-day rain by training classification models on the target variable Rain Tomorrow and various parameters including Date, Location, Mintemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, etc. The missing values are usually handled by random sample imputation to observe the variance categorical Vvlues like location, wind direction are handled by using Target guided Encoding Outliers are handled using IQR and boxplot Imbalanced Dataset was handled using SMOTE (Synthetic Minority Oversampling Technique).

4. CONCLUSION

Rain fall prediction has a very important role in agriculture production. The growth of the agricultural production is based on the amount of rainfall. Hence, it is advisable to predict the rainfall of a season to guide the farmers in agriculture. In this paper, a gradient boost approach namely CatBoost has been proposed to forecast rainfall for a selected location in Australia. The proposed method predicts the rainfall for the Australian dataset using decision tree algorithm (Catboost model) and provides improved results in terms of accuracy and prediction. . The study showed that the Catboost model was performing better in months with high annual averages compared to alternative approaches. In future scope, group of techniques will be used to combine the diversities of the models. Moreover, additional locations and datasets can be incorporated.

ACKNOWLEDGEMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Payel Thakur for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department, Dr. Sharvari Govilkar and our Principal, Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] Yang Liu, Qingzhi Zhao, Wanqiang Yao, Xiongwei Ma, Yibin Yao and Lilong Liu, "Short-term rainfall forecast model based on the improved BP–NN algorithm",2019.Available:https://www.nature.com/articles/s41598-019-56452-5[Submitted on 24 December 2019]

[2] Vishal Morde. "XGBoost Algorithm: Long May SheReign!",2019.Available:https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-longshe-may-rein-edd9f99be63d [Submitted on 8 April 2019]

[3] Ramya Bhaskar Sundaram, "An End-to-End Guide to Understand the Math behind XGBoost", 2018.Available:https://www.analyticsvidhya.com/blog/2018/09/an-endto-end-guide-to-understand-the-math-behind-xgboost/[Submitted on 6 September 2018]

[4] Available:https://www.kaggle.com/prashant111/catboost-classifier-in-python

[5] Aman Kharwal, "Rainfall Prediction with Machine Learning",2020.Available:https://thecleverprogrammer.com/2020/09/11/rainfallprediction-with-machine-learning/ [Submitted on 11 September 2020]

[6] Anamika Jha "CatBoost – A new game of Machine Learning:", 2020. Available:https://affine.ai/catboost-a-new-game-of-machinelearning/ [Submitted on 21 October 2020]

BIOGRAPHIES

Vijithra Nair

Megha Mathew

Sweta Bhattacharjee

Arashdip Singh