Weather Prediction Using Machine Learning

Ajay Rawat[1], Adarsh Kumar[2], Manjinder Singh [3]
[1,2],Research Scholar, School of Computer science and engineering, Lovely Professional University ,
[3]Associate Professor, School of Computer science and engineering, Lovely Professional University

ajay.11909456@lpu.in[1], adarsh.11909469@lpu.in[2], manjinder.29540@lpu.co.in[3]

Abstract: **Climate plays an important role in many important sectors such as security, agriculture, electricity, utilities, shipping, aviation, agriculture and forestry. Today's climate change is more dramatic, so a more effective model must be found for such unpredictable weather. Appropriate weather systems will help farmers plan appropriate agricultural practices to protect crops. These estimates affect the country's economy and the lives of its people. Looking at today's results, it has been shown that tools based on machine learning (ML) techniques are more effective in predicting weather conditions than traditional models based on weather. The aim of our work is to find an effective weather model. This project includes algorithms like- catboost, xgboost, random forest, knn, svm, naïve bayes, support vector classifier etc. We use distribution indicators - the roc curve and the auc score - to compare different models and choose the more precise model for climate change.**

Keywords: **Catboost, Random Forest, Logistic Regression, Extreme Gradient Boosting, SVM, K-Nearest Neighbors, Naïve Bayesroc curve and auc score.**

I. INTRODUCTION

Weather forecasting prediction involves the development and dissemination of information about future weather based on the collection and analysis of weather observations. Weather forecasting is the forecast of weather results obtained from numerical solution of equations governing weather and changing conditions. Nowcasting is a short-lived form of weather broadcasting, i.e., current weather and forecast up to two hours ahead. Forecasts longer than the next two weeks are called long-term forecasts or deferred climates.

In this paper, we use seven models for predicting the next day's weather that rainfall happens or not by taking extensive historical weather data from Kaggle. We take this data and use it to develop and find a simple learning model that can predict the weather for the next few days. These models can work with low-cost, low-cost equipment while providing fast and reliable predictions that we can use in our daily lives. Short-term weather forecasting using machine learning, capable of running on low-cost machines, is an important part of this document. Will use an extensive historical data from Kaggle. Comprehensive evaluation of the proposed method and comparison of various machine learning models for future climate prediction. For result evaluation and comparison of different model's roc curve and auc score is used.

The flow of the paper can be outlined as follows- In unit 2 provides an extensive discussion of the proposed methodology, beginning with the original data set and explaining the seven machine learning models and tools utilized in this research. The findings and analysis are depicted in unit 3 and 4, respectively, with conclusion appearing in unit 5.

II. RELATED WORKS

[1] Shivam Tandon, Pawan Kumar Singh, Abhishek Patel - In this study a technology is delivered using gadget gaining knowledge of strategies to offer weather forecasts. sensible models may be created the use of machine getting to know technologies which can be a whole lot less difficult than exsisting models. They can be run on nearly any system, along with mobile devices. additionally, they plan to accumulate weather information from various components of a town using low-price internet of things (IoT) gadgets, such as temperature and humidity sensors. The wide variety of local capabilities inside the education dataset can be increased by way of the use of exceptional sensors. Our prediction models will increase even further because of this data, in an effort to be combined with statistics from climate

stations.

[2] Suvendra Kumar Jayasingh, Jibendru Kumar Mantri, Sipali Pradhan – In this study they estimate the weather events the usage of a gadget studying model that takes under consideration the unique climate parameters. on this paper, they offered distinct machine gaining knowledge of fashions which can be used for prediction of climate with a great deal less difficult and easier way than the physical fashions. Evaluation of the model's accuracy showed that the device recognizing the model outperformed the standard model. This model uses data collected from previous sources with the highest accuracy as high as 81.67%.

[3] G. Hemalatha, K. Srinivasa Rao, D. Arun Kumar - In this study, FCNN model is proposed for forecasting weather data. The conceptual version of FCNN can be efficient and comprehensive for capturing the nonlinearity of potential concepts in the data. The model is better than OA, UA, PA, KC models. This model achieved 87.83% OA when tested with the IMD dataset. Additionally, the model can be extended to classes of larger datasets.

[4] Prathap N Kashyap, Preetham M S, Rohan Venkatesh Nayak, Uday B, Radhika T V – a web application evolved predicts the climate based totally at the machine learning techniques. system mastering techniques use few features In any machine learning related version it is hard to discover the proper set of rules for the right hassle. The climate prediction utility can be improvised by using the use of different gadget learning algorithms too. in this software best few capabilities are predicted. Other functions also can be added with the aid of education the model and integrating it with the present day application. The utility also can be provided with an option to input the desired capabilities from the user instead of an API call.

[5] Rajasekaran Meenal, Kiruthic Kailash, Prawin Angel Michael, Jeyaraj Jency Joseph, Francis Thomas Josh, Ekambaram Rajasekaran - In this research, weather parameters such as precipitation, wind speed, solar radiation and relative humidity are estimated using horizontal and tree model together with gadget learning algorithms. In addition, it is recommended to use the temperature model based on the visible model, which is Hargreaves and Samani and Bristow and Campbell versions for solar power of many places in India. Validation of the empirical and gadget working model using closed test results were obtained from IMD, Pune. They found that model ML-based results were better in evaluating the empirical model with a comparative "R" value of 0.9259 and an MSE of 0.1397. Therefore, with better weather and less work, research work came to a good halt.

[6] Amit Kumar Agarwal, Manish Shrimali, Sukanya Saxena, Ankur Sirohi, Anmol Jain - In the research paper, they found that the relationship between heat and cold in different places affects the weather, causing heavy rains, floods and weather disturbances.

[7] Rudransh Kush, Shubham Gautam, Sai Suvam Patnaik, Tanisha Chaudhary – In this project, natural climatic change isn't plenty dangerous for the earth but whilst a few outside pressure is imposed then this pressure affects the balance of the weather cycle and misbalances the natural glide which results in a negative effect over the biotic additives surviving on the earth and the abiotic factors which facilitates in reshaping the earth's floor and wishes to be taken well care off else will smash them absolutely. Predicting the climate beforehand by means of statistically research and analysis will assist in overcoming a lot of these issues and might force the respective corporations to make right regulations in accordance through studying the present and beyond information. Due to such high research approximately surroundings, recognition may be spread amongst human beings regarding reducing surroundings pollutants and the way can they contribute towards creating a easy and inexperienced surrounding. Concluding the factors, this

research will without a doubt make a superb impact and will honestly convey the natural climatic trade waft lower back to tune with the aid of teaching and motivating humans what to do and what now not to do which could have an effect on the mother Nature.

[8] Nazim Osman Bushara and Ajith Abraham – In this work 14 base algorithms taken into consideration Date, minimum Temperature, Humidity and Wind route as predictors for the rainfall, and people algorithms have followed furnished check set as check choice. The Correlation coefficient of all base classifiers is greater than zero.8. So, on this study they both taken into consideration simplest seven predictors for rainfall prediction and use some greater climate factors along with atmosphere strain, sea floor temperature, and many others, so they will acquire extra correct prediction. additionally, if Ensemble strategies had been carried out the effects may be advanced.

[9] Sana Khan, Mani Priya Mishra, Rukaiya Khatoon, Ritika Singh - In this paper, they worked with mixture of Naïve Bayes algorithm to are expecting climate circumstance. The steady records i.e., time-collection statistics is assembled associated evaluation is carried out on this dataset utilising an interface named weather Prediction system, evolved utilising Java using Eclipse tools. The destiny work of this venture is to encompass loads of characteristics of weather condition to are expecting and to paintings with different category set of rules to turn out to be plenty of accurate in prediction.

III.METHODOLOGIES:

A. *DATA COLLECTION:*

The data we got from Kaggle to build the model consists of nearly 10 years of daily weather observations from various weather stations in Australia. RainTomorrow is a target to predict. That means - did it rain, did it rain or didn't it rain

the next day? a column, it's indicated that day there is 1mm or more rain. This Dataset (14.09 MB) has extensive 21 parameters for prediction, like humidity, cloud, pressure, wind, temperature etc. at different time of day.

B. DATA PREPROCESSING:

- No value from the random assignment model to capture the variance.
- Position, wind direction, etc. categorical values are handled using target guided encoding.
- Use IQR and box plots to handle outliers.
- Unbalanced data was processed using SMOTE.

C. MODELS:

Different types of models are:

- *CATBOOST*

CatBoost is a gradient boosting machine learning library designed for high-performance and easy-to-use machine learning tasks. It was developed by the Russian company Yandex and is open-source software. It has the ability to work with categorical features without the need for pre-processing or one-hot encoding, which can save time and reduce the risk of overfitting. It also includes a variety of advanced features, such as automatic handling of missing values and built-in cross-validation.

CatBoost supports both classification and regression tasks and can handle large datasets with millions of rows and tens of thousands of columns. It also provides a range of hyperparameters that can be tuned to optimize performance and avoid overfitting. Overall, CatBoost is a powerful tool for machine learning tasks, especially those with complex, high-dimensional datasets containing categorical features.

decision trees and gradient boosting to improve accuracy and speed. It is particularly effective for large, high-dimensional datasets and is known for its efficiency and scalability. It can handle missing data and works well with a variety of input formats, including numeric, categorical, and text data. Features of XGBoost is its ability to handle overfitting through regularization techniques, such as L1 and L2 regularization, and early stopping.

XGBoost has been extensively used for many kind of applications, including fraud detection, recommendation systems, and natural language processing. It is considered one of the top machine learning libraries in the industry and has won numerous

- *RANDOM FOREST*

Random Forest is a model that merging multiple decision trees to build more accurate and robust model. In a Random Forest, multiple decision trees are created using subsets of the original dataset, with each tree making predictions independently. The final prediction is then determined by aggregating the predictions of all trees. It has ability to manage many features and avoid overfitting. The algorithm achieves this by randomly selecting a subset of features for each tree, reducing the connection between the trees, and increasing the diversity of the model.

It can be used for both classification and regression task and has several hyperparameters that can be tuned to optimize performance.

- LOGISTIC REGRESSION

Logistic regression is an example of a model often used in binary distribution problems where the purpose is to predict whether an event has occurred or not. In logistic regression, the input variables are used to calculate the weighted sum, and then the output is mapped to a probability value between 0 and 1 by a logistic function. This probability value is used to determine the final binary distribution. Models can be easily identified, and the coefficients of the different inputs can be used to determine the strength and direction of their relationship to the output.

Logistic regression can be modified to handle multiclass classification problems using a technique called multinomial logistic regression or softmax regression.

- XGBOOST

It's stand for "Extreme Gradient Boosting," is an open-source library. It is a modified version of the gradient boosting method that uses a merging of

competitions on Kaggle and other data science platforms.

- *SVM*

Support Vector Machines is a supervised model that is extensively used for classification and regression tasks. This allows SVM to be highly effective in cases where there is a clear separation between the classes. It can handle both linear and nonlinear classification problems by using different types of kernels to transform the input data into a higher no of features space where the classes can be more easily separated.

SVM also can handle large datasets and is less susceptible to overfitting than some other machine learning algorithms. It can be very expensive task,

especially when working with large datasets or complex kernels.

- *KNN*

K-Nearest Neighbour is a non-parametric model that can predicts the class of a new observation by finding the nearest K training example in the specified space and taking the majority class among these neighbours.

It is often used in applications where the decision boundary is highly irregular or nonlinear. It can handle both continuous and categorical input variables However, KNN not recommended with large dataset because it's too computationally expensive.

- *NAIVE BAYES*

Naive Bayes is an algorithm uses the training data to calculate the conditional probability of each input variable given each class variable. This allows it to estimate the probability of a new observation belonging to each class based on its input variables.

It can work effectively well with a high number of features in datasets. It can handle both continuous and categorical input variables and is relatively insensitive to irrelevant features. One of the drawbacks of Naive Bayes is the assumption of conditional independence, which may not hold true in some cases. However, the model can still perform well in practice, especially when the training data is large.

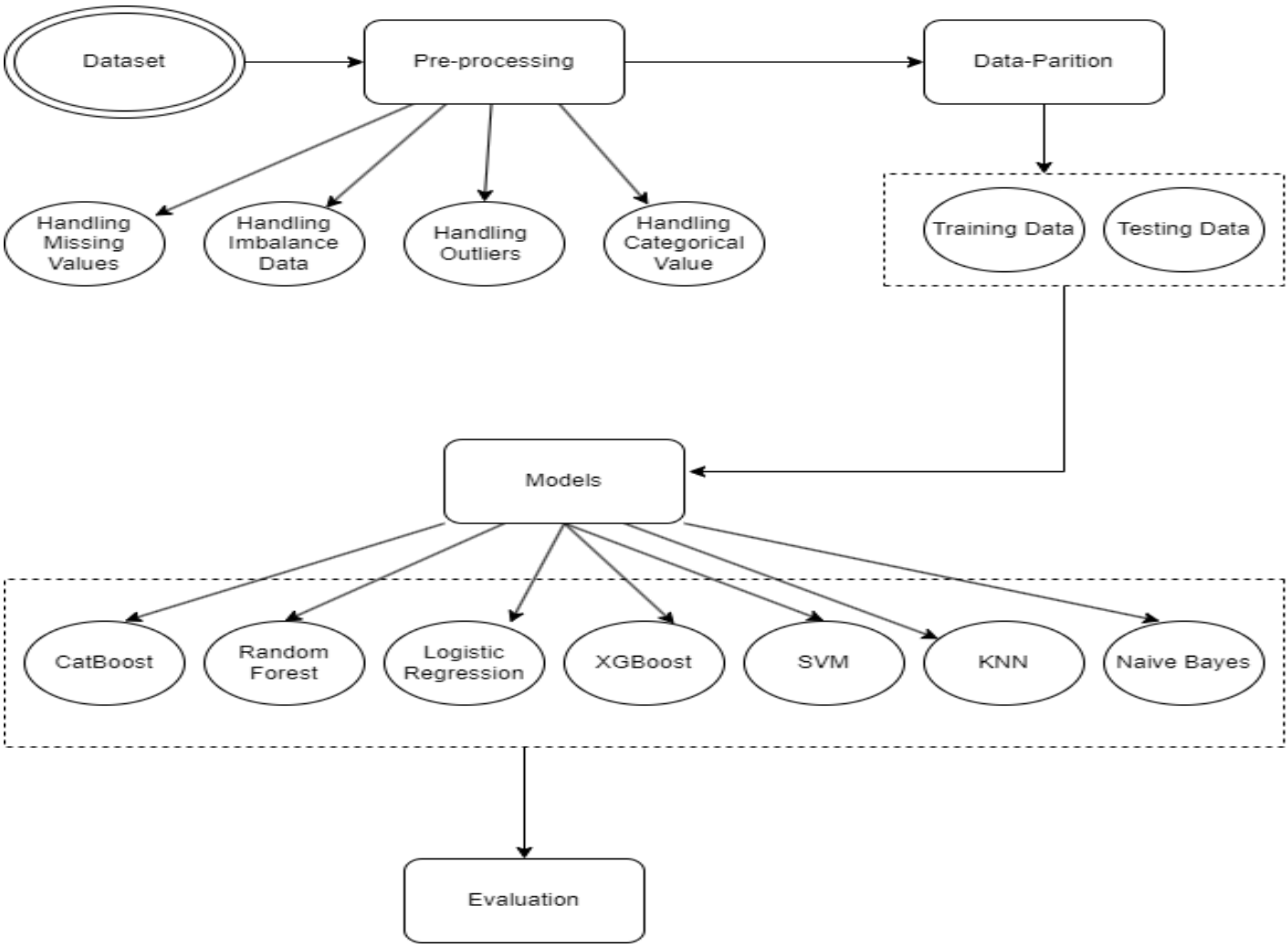

Fig 1. Flowchart of System

IV. EXPERIMENTAL RESULTS:

A. RESULT EVALUATION:

- *Accuracy*

Accuracy is a measure used to evaluate the performance of classification models. It measures the percentage of correct classified samples from all samples in the data.

- *AUC*

AUC stands for "Area Under the Curve." It is threshold independent, used to make the final classification decision, making it useful when comparing models with different thresholds.

- *ROC AUC Score*

ROC AUC Score is based on its Receiver Operating Characteristic (ROC) curve. The ROCAUC score is simply the AUC score calculated from the ROC curve. It is popular because it is a single number that summarizes the model's overall performance, and it is independent of the classification threshold used to make the final prediction.

MODELS	ACCURACY	ROC AUC SCORE	AUC
CATBOOST	*0.86*	*0.75*	*0.89*
RANDOM FOREST	*0.84*	*0.75*	*0.87*
LOGISTIC REGRESSION	*0.76*	*0.76*	*0.85*
EXTREME GRADIENT BOOSTING	*0.85*	*0.74*	*0.88*
SVM	*0.77*	*0.76*	*0.85*
K-NEAREST NEIGHBORS	*0.75*	*0.74*	*0.79*
NAÏVE BAYES	*0.74*	*0.74*	*0.82*

Table 1. Accuracy Matrix

B. CONFUSION MATRIX

It's a table used to evaluate the performance of a classification algorithm. These criteria can provide insights into how well the model is performing and can be used to optimize the model for better results.

- True Positive (TP) is truly positive & forecast as positive.
- False Negative (FN) is truly positive & forecast as negative.
- False Positive (FP) is truly negative & forecast as positive.
- True Negative (TN) is truly negative & forecast as negative.

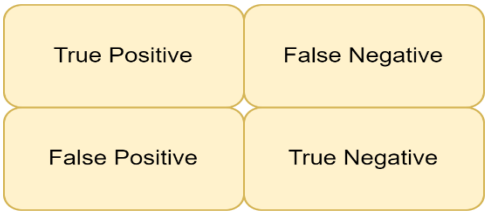

Fig 2. Confusion matrix

Precision = **TP/ (TP+FP)**
Sensitivity = **TP/(TP+FN)**
Specificity = **TN/(TN+FN)**
Accuracy = **(TP + TN)/(P+N)**

MODELS	PRECISION	RECALL	F1-SCORE
CATBOOST	*0.88*	*0.95*	*0.91*
RANDOM FOREST	*0.89*	*0.91*	*0.90*
LOGISTIC REGRESSION	*0.92*	*0.77*	*0.84*
EXTREME GRADIENT BOOSTING	*0.88*	*0.94*	*0.91*
SVM	*0.92*	*0.78*	*0.85*
K-NEAREST NEIGHBORS	*0.91*	*0.76*	*0.83*
NAÏVE BAYES	*0.91*	*0.75*	*0.82*

Table 2. Confusion Matrix

C. TOOLS & TIME CONSUMED:

The proposed system is trained on AMD A8-6410 APU (2 GHz), 4GB GPU and 8GB RAM. Google. Colab is used to execute code. The execution time consist of all models was approx. 64 minutes.

V. CONCLUSION & FUTURE WORK

This study develops a method for classifying whether rain happens or not. Out of all seven models catboost, xgboost and random forest were top 3 with highest accuracy of catboost model as 86% whereas other two with accuracy of 85% and 84% respectively.

We compared seven machine learning models of weather prediction for significance of agricultural activities and precaution from disaster. In future this weather prediction project can be better performed for prediction mistakes, such as rainfall and produce more accurate predictions using hyperparameter optimization for better accuracy.

REFERENCES

[1] “Shivam Tandon, Pawan Kumar Singh, Abhishek Patel” - Weather Prediction Using Machine Learning, 2017.

[2] “Suvendra Kumar Jayasingh, Jibendru Kumar Mantri, Sipali Pradhan” - Smart Weather Prediction Using Machine Learning, 2022.

[3] “G. Hemalatha, K. Srinivasa Rao, D. Arun Kumar” - Weather Prediction using Advanced Machine Learning Techniques, AMSE 2021.

[4] “Prathap N Kashyap, Preetham M S, Rohan Venkatesh Nayak, Uday B, Radhika T V” - Smart Weather Prediction Technique Using Machine Learning, 2020.

[5] “Rajasekaran Meenal, Kiruthic Kailash, Prawin Angel Michael, Jeyaraj Jency Joseph, Francis Thomas Josh, Ekambaram Rajasekaran” - Machine learning based smart weather prediction, 2022.

[6] “Amit Kumar Agarwal, Manish Shrimali, Sukanya Saxena, Ankur Sirohi, Anmol Jain” - Forecasting using Machine Learning, 2019.

[7] “Rudransh Kush, Shubham Gautam, Sai Suvam Patnaik, Tanisha Chaudhary” - Climate Monitoring and Prediction using Supervised Machine Learning, 2022.

[8] “Nazim Osman Bushara and Ajith Abraham” - Weather Forecasting in Sudan Using Machine Learning Schemes, 2014.

[9] “Sana Khan, Mani Priya Mishra, Rukaiya Khatoon, Ritika Singh” - Weather Prediction Using Machine Learning, 2022.