RAINFALL PREDICTION using AI/ML

Project Report
Presented by

Team SRT

Contents:

Abstract:

Traditionally, rainfall prediction was based on physical models that considered the interactions of various atmospheric and oceanic factors. However, these models were limited by the availability of data, computational power, and the complexity of the physical processes involved. In recent years, the use of AI/ML techniques has gained popularity in the field of rainfall prediction due to their ability to handle large datasets and extract complex patterns. By analyzing large datasets of historical weather data, these models can extract patterns and trends that enable more precise forecasting of rainfall. This concludes that the use of AI/ML in rainfall prediction can significantly aid in better planning and decision-making, ultimately contributing to the sustainable management of resources.

Acknowledgement:

We are successfully completing this project, many people helped us. We would like to thank all those who are related to this project.

For all the efforts behind the project work, we first & foremost would like to express our sincere appreciation to the staff of Department of Electronics and Telecommunication Engineering under Dr A.V. Nandedkar Sir, for their extended help & suggestions at every stage of this project. It is with a great sense of gratitude that we acknowledge the support, time to time suggestions and highly indebted to our guide. Finally, we pay our sincere thanks to all those who indirectly and directly helped us towards the successful completion of this project report.

Suchit Vikas Gaidhane 2020BEC032

Rutik Jayram Torambe 2020BEC014

Tejas Vijay Warade 2020BEC006

Introduction:

Rainfall prediction is a crucial area of research in the field of meteorology and agriculture. Accurate prediction of rainfall patterns can aid in better planning and decision-making for a wide range of activities, such as crop cultivation, disaster management, and water resource management. With the advent of advanced technologies such as Artificial Intelligence (AI) and Machine Learning (ML), it is now possible to develop more precise and reliable models for rainfall prediction. These techniques utilize large datasets of historical weather data, which are processed and analyzed to extract patterns and trends that can be used to make predictions about future rainfall. This approach has the potential to significantly improve the accuracy and timeliness of rainfall forecasts, leading to more effective planning and management of various activities. This project aims to explore the use of AI/ML techniques in rainfall prediction and its potential benefits.

Benefits of AI/ML-based Rainfall Prediction:

AI/ML-based rainfall prediction models offer several benefits over traditional methods, such as:

- ✓ **Increased accuracy: AI/ML-based models can extract patterns and trends that may not be apparent using traditional methods, leading to more accurate predictions.**

- ✓ **Timely forecasting: AI/ML-based models can process large datasets quickly, enabling timely forecasting of rainfall patterns.**

- ✓ **Scalability: AI/ML-based models can handle large datasets, making them scalable for different regions and time periods.**

- ✓ **Cost-effective: AI/ML-based models require minimal physical infrastructure and can be implemented using cloud-based services, making them cost-effective.**

Creation of ML Model:

1.Defining the problem and collecting data:

The "Weather in Australia" dataset available on Kaggle contains historical weather data from various weather stations across Australia. The dataset comprises daily weather observations recorded from 2007 to 2017, including variables such as temperature, rainfall, wind speed and direction, humidity, and air pressure.

The dataset contains a total of 1,45,460 records with 24 features, including the target variable "RainTomorrow", which indicates whether it rained the next day or not. This binary variable is used for predicting rainfall using machine learning models.

The features included in the dataset are as follows:

1. Date: The date of the observation.
2. Location: The location of the weather station where the observation was recorded.
3. MinTemp: The minimum temperature recorded on the day in degrees Celsius.
4. MaxTemp: The maximum temperature recorded on the day in degrees Celsius.
5. Rainfall: The amount of rainfall recorded in millimeters on the day.
6. Evaporation: The amount of evaporation recorded in millimeters on the day.
7. Sunshine: The amount of sunshine recorded in hours on the day.
8. WindGustDir: The direction of the strongest gust of wind on the day.

9. WindGustSpeed: The speed of the strongest gust of wind on the day in kilometers per hour.
10. WindDir9am: The direction of the wind at 9 am on the day.
11. WindDir3pm: The direction of the wind at 3 pm on the day.
12. WindSpeed9am: The speed of the wind at 9 am on the day in kilometers per hour.
13. WindSpeed3pm: The speed of the wind at 3 pm on the day in kilometers per hour.
14. Humidity9am: The humidity recorded at 9 am on the day.
15. Humidity3pm: The humidity recorded at 3 pm on the day.
16. Pressure9am: The air pressure recorded at 9 am on the day in hectopascals.
17. Pressure3pm: The air pressure recorded at 3 pm on the day in hectopascals.
18. Cloud9am: The fraction of sky obscured by cloud at 9 am on the day.
19. Cloud3pm: The fraction of sky obscured by cloud at 3 pm on the day.
20. Temp9am: The temperature recorded at 9 am on the day in degrees Celsius.
21. Temp3pm: The temperature recorded at 3 pm on the day in degrees Celsius.
22. RainToday: A binary variable indicating whether it rained on the day or not.
23. RainTomorrow: A binary variable indicating whether it will rain the next day or not (the target variable).
24. Dataset: A variable indicating whether the observation is from the "train" or "test" set.

2. Data cleaning and preprocessing: **It contains cleaning and preprocessing data to remove missing or irrelevant data, normalize or scale the data, and encode categorical variables.**

- **We imported dataset using pandas dataframes, viewed the dimensions of dataset. Then checked the missing values and filled them using random sample imputation, using sample median, mode, etc.**
- **Then visualize the data, we used seaborn boxplot to view outliers in data and also used distplot which shows distribution plot.**
- **Then we used sklearn labelencoder to encode target labels and normalize data. Also it converted non numerical data to numerical data.**
- **Next we plotted correlation heatmap which shows the whether features within a dataset correlate with each other.**
- **Before data analysis we removed the outliers. Then analysed data to get the best fit line using scipy stats probplot.**

3. Split the data: **We need to split your data into training and testing sets. The training set is used to train your model, while the testing set is used to evaluate the performance of your model. We used sklearn train_test_split to split the given data into train and test data.**

4.Train the data: **We used SMOTE (Synthetic Minority Oversampling Technique) which is a statistical technique for increasing the number of cases in your dataset in a balanced way.**

We selected catboost classifier to create our model because in this case accuracy is high around 86%. We determined accuracy using sklearn metrics accuracy_score, confusion matrix and classification report. Also calculated roc_auc_score which gives area under the ROC curve (receiver operating characteristic curve).

5.Save and deploy the model:

Once the model is trained and evaluated, we need to deploy it in a production environment, where it can be used to make predictions on new data. We saved the model using joblib and deployed model using flask.

Analysis of Model:

Methods and accuracy score:

Sr.no	Method	Accuracy score	ROC AUC score
1	Catboost classifier	0.8640	0.7565
2	Random forest classifier	0.8455	0.7648
3	Logistic regression	0.7743	0.7677
4	Gaussian naïve bayes	0.7538	0.7428
5	K-Neighbor classifier	0.7538	0.7429
6	XGBoost classifier	0.8562	0.7489

How catboost classifier works?

CatBoost is a supervised machine learning method that uses decision trees for classification and regression. As its name suggests, CatBoost has two main features, it works with categorical data (the Cat) and it uses gradient boosting (the Boost). Gradient boosting is a process in which many decision trees are constructed iteratively. Each subsequent tree improves the result of the previous tree, leading to better results. CatBoost improves on the original gradient boost method for a faster implementation.

1.Data Preprocessing: The first step is to preprocess the data. This may include feature selection, data cleaning, and data normalization.

2.Initialization: The CatBoost algorithm initializes the model with a single decision tree that predicts the mean target value of the training data.

3.Gradient Boosting: The CatBoost algorithm uses gradient boosting to improve the accuracy of the model. Gradient boosting is an iterative process that trains multiple decision trees in sequence. The goal is to correct the errors of the previous trees with each new tree.

4.Tree Generation: In each iteration, the CatBoost algorithm generates a new decision tree. It selects the best split points by minimizing the loss function. The loss function is a measure of the difference between the predicted values and the actual values.

5.Shrinkage: The CatBoost algorithm applies shrinkage to the decision trees to prevent overfitting. Shrinkage involves reducing the weight of the new trees added to the model, which makes the model less sensitive to the noise in the training data.

6.Prediction: Once the model is trained, it can be used to make predictions on new data. The CatBoost algorithm uses the majority vote of the decision trees to make the final prediction.

Conclusion:

In conclusion, the use of AI/ML techniques in rainfall prediction offers significant potential benefits for various activities such as agriculture, disaster management, and water resource management. The models developed using AI/ML can handle large datasets, extract complex patterns, and provide accurate and timely predictions. It is evident that AI/ML-based rainfall prediction is a promising area of research that can contribute to the sustainable management of resources.

References:

Dataset: https://www.kaggle.com/datasets/gauravduttakiit/weather-in-aus

Github: https://github.com/Tejas-w01/Rainfall-Prediction