Machine Learning Project On

Rainfall Prediction Using Supervised Learning

Project By :

Mangesh Raju Bondre

Guided By :

Mayur Salunkhe

Table of Contents

Introduction:

Predicting rainfall isn't something entirely new in itself. The date back rainfall prediction to the beginning of 1860s more or less. However, there is no denying that regardless of the methods implemented, time and resources invested, and feats achieved; predicting the rainfall with high accuracy if not certainty, can still prove ever useful for mankind as a whole, if not for a specific individual. In this project, implement Data Analytics to the best of our limits on various factors mentioned above and draw conclusions out of it. This conclusion will in turn enable us to make a probable or certain prediction regarding the upcoming weather conditions.

Accurate rainfall prediction is a challenging task because of the complex physical processes involved. This complexity is compounded in Australia as the climate can be highly variable. Predicting rainfall accurately aids in effectively planning, managing, and implementing water resources systems and can assist with more sustainable operations. Science and Technology play an essential role in predicting the climate of any country via rainfall prediction. This Project proposes a rainfall prediction for Indian dataset. Intention of this paper is to make weather predictions based upon already collected data regarding various factors like Rainfall, Sunshine, Wind Gust Direction, Humidity, Pressure, Cloud, Rain Today, etc.

Problem Statement:

Climate is an important aspect of human life. So, the Prediction should accurate as much as possible. In this project I have tried to deal with the prediction of the rainfall which is also a major aspect of human life and which provide the major resource of human life which is Fresh Water. Fresh water is always a crucial resource of human survival – not only for the drinking purposes but also for farming, washing and many other purposes**.**

Aim:

To determine the rainfall for effective use of water resources, crop productivity and pre planning of water structures.

Objective:

- Rainfall Prediction Model has a main objective in prediction of the occurrence of rain in a specific division in advance by using various Machine Learning technique and find out which Model is best for rainfall prediction.
- Finding out which features are most important while building model.
- Performing data pre-processing and exploratory data analysis, SMOTE Analysis on the data set.
- Evaluating the models on the test set using the metrics such as Confusion Matrix.

Statistical Tools and Software

Exploratory Data Analysis: Box plot, Heatmap, SMOTE

ML models:

Logistic Regression
Decision Tree Classifier
Random Forest Classifier
K Nearest Neighbour
XGBoost

Software:

Python
Flask
Jupyter Notebook
VS code
Google Chrome

.

Methodology & Data Description Data Collection

Data Sources:

The Dataset used for this project comprises historical Rainfall data, including details about various factors like Temperature, Date, Location, Pressure, Cloud, Wind Directions, Rain Today, etc.

The Dataset was obtained from google, the link for the same is as below: https://drive.google.com/file/d/1x8u9Pvz_DB5CxOpUXOYMRVT7QUhxL7Xd/view?usp=sharing

Data Description:

Min Temp	Minimum Temperature of that day in Degree Celsius.	Numerical
Max Temp	Maximum Temperature of that day in Degree Celsius.	Numerical
Rainfall	Amount of Rainfall in Millimetres.	Numerical
Evaporation	Amount of Evaporation in millimetres per unit time.	Numerical
Sunshine	Density of Sunlight in Watts per square meter.	Numerical
WindGustSpeed	Sudden speed of wind from a particular direction	Categorical
WindSpeed9am	Speed of wind at 9am	Numerical
WindSpeed3pm	Speed of wind at 3pm	Numerical
Humidity9am	Amount of Moisture in air at 9am.	Numerical
Humidity3pm	Amount of Moisture in air at 3pm.	Numerical
Pressure9am	Air Pressure at 9am.	Numerical
Pressure3pm	Air Pressure at 3pm.	Numerical
Cloud9am	Measure of Cloud in okta at 9am.	Numerical
Cloud3pm	Measure of Cloud in okta at 3pm.	Numerical
Temp9am	Temperature at 9am.	Numerical
Temp3pm	Temperature at 3pm.	Numerical
Rain Today	Occurrence of rainfall for today.	Categorical
Rain Tomorrow	Occurrence of rainfall for Tomorrow.	Categorical
year	Year you are trying to predict the rainfall	Numerical
month	Month you are trying to predict the rainfall	Numerical
day	Day you are trying to predict the rainfall	Numerical
Location_cl	Name of the Location	Categorical
WindGustDir_cl	From which direction there is a sudden speed of wind.	Categorical
WindDir9am_cl	Direction of wind at 9am	Categorical
WindDir3m_cl	Direction of wind at 3pm	Categorical

Model Deployment:

Frontend Development:

The developed machine learning model for predicting Rainfall has been integrated into a user-friendly web application. The frontend of the application has been developed using HTML, CSS, and JavaScript to provide an interactive interface for users to input climate or weather specific data and receive prediction.

User Interface Design:

The user interface has been designed to be intuitive and easy to use. Users are prompted to input relevant Climate details, such as temperature, Humidity, pressure, cloud, rain today, rainfall, sunshine, etc.

Integration with Flask:

The backend functionality is facilitated by Flask, a micro web framework for Python. Flask acts as the middleware, processing user inputs, utilizing the trained model, and returning the prediction that whether it will rain or not to the frontend in real-time.

Model Accessibility:

The trained machine learning model is serialized using Python's pickle module and saved as a .pkl file. This serialized model file is loaded within the Flask application to ensure seamless access and utilization for predictions.

Scalability and Maintenance:

Considerations for scalability and future maintenance have been accounted for in the application's design. The modular structure of the codebase allows for straightforward updates and enhancements to accommodate potential model improvements or additional features.

Machine learning algorithms:

Logistic Regression:

logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Logistic regression is used for both classification and regression problems. It is used to estimate the probability of an instance if it belongs to one class or another. If it belongs to a particular class, it is called a positive class labeled as 1, and if it does not belong to a particular class, it is called a negative class labeled as 0.

Decision Trees:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Random Forest:

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

XG Boost

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage is with decision Trees

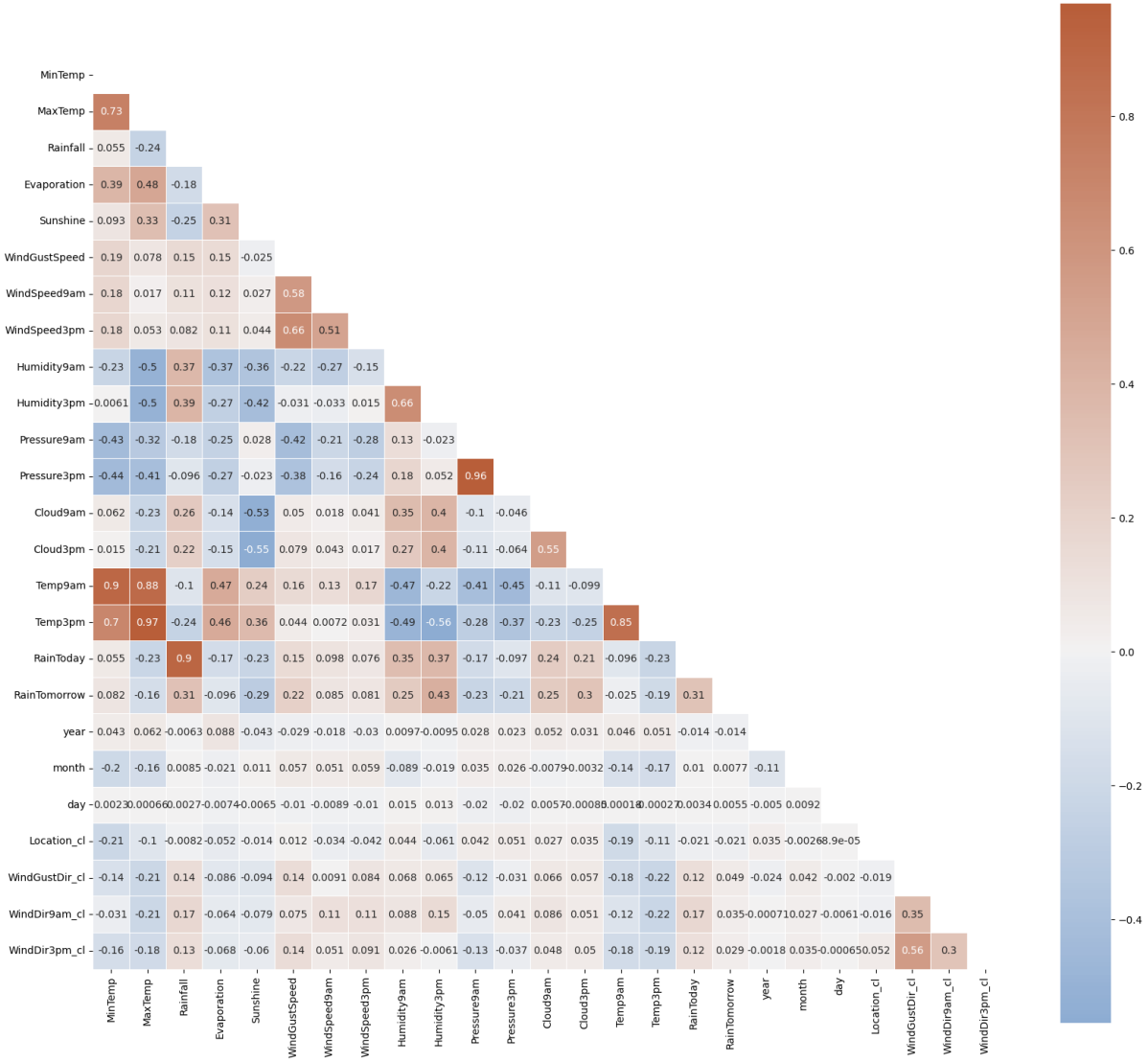
The K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is used for classification and regression tasks. While training the KNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance. Next, the algorithm identifies the K nearest neighbors to the input data point based on their distances. the case of classification, the algorithm assigns the most common class label among the K neighbors as the predicted label for the input data point.

Interpretation:

Exploratory data analysis

Heatmap:



Interpretation:

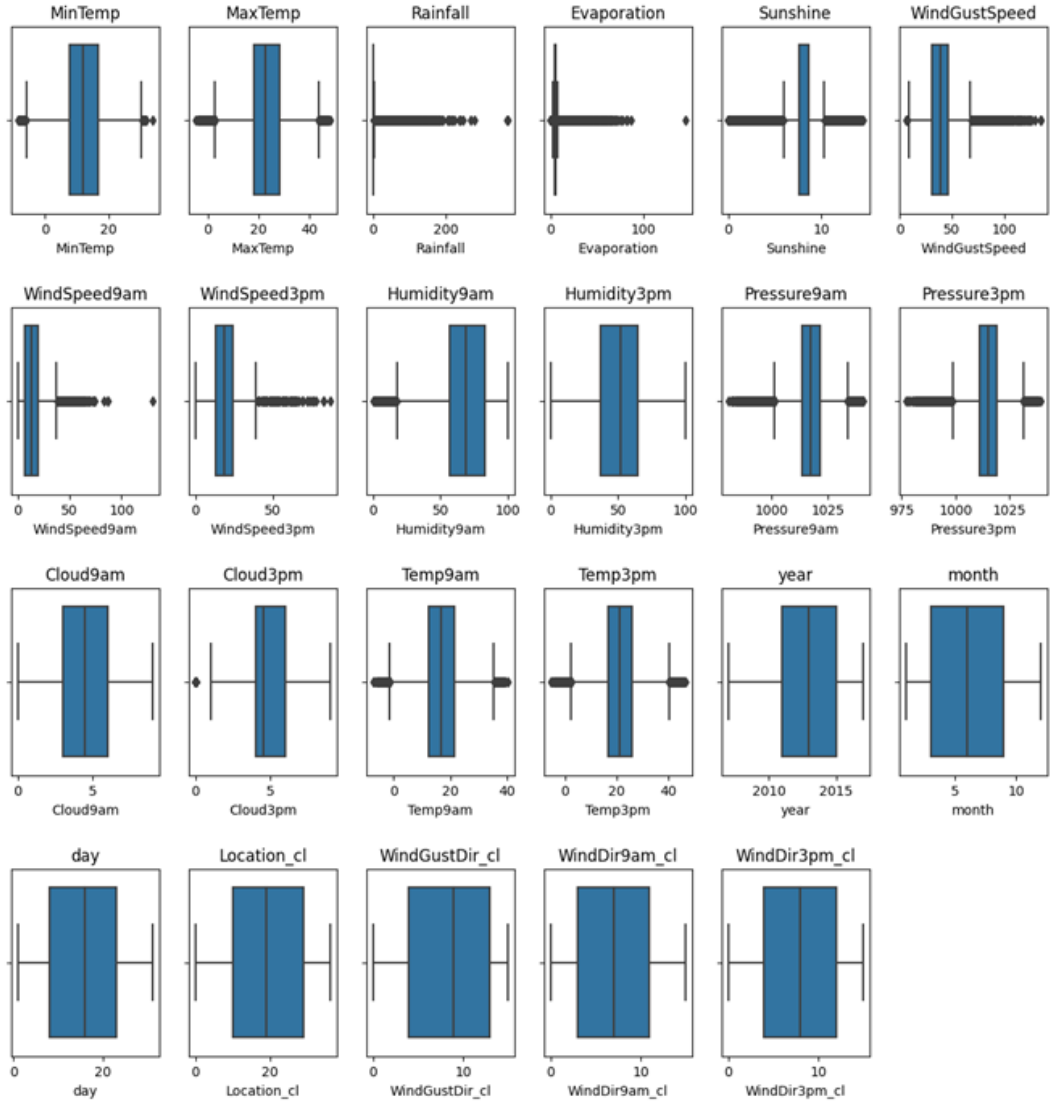
1.Heatmap is used to find out the relationship between two continuous variables**.**

2.Darker brown shades show highest positive correlations whereas darker blue shade shows negative correlation.

3.RAinfall, Humidity3pm, Cloud3pm, Rain Today shows Highest Positive Correlation with our Target Variable whereas Sunshine and pressure shows the highest negative correlation.

4.Temp9am has the highest positive correlation with Min Temp while Temp3pm and Humidity3pm shows the highest negative correlation with each other.

Outliers Detection using Boxplot:

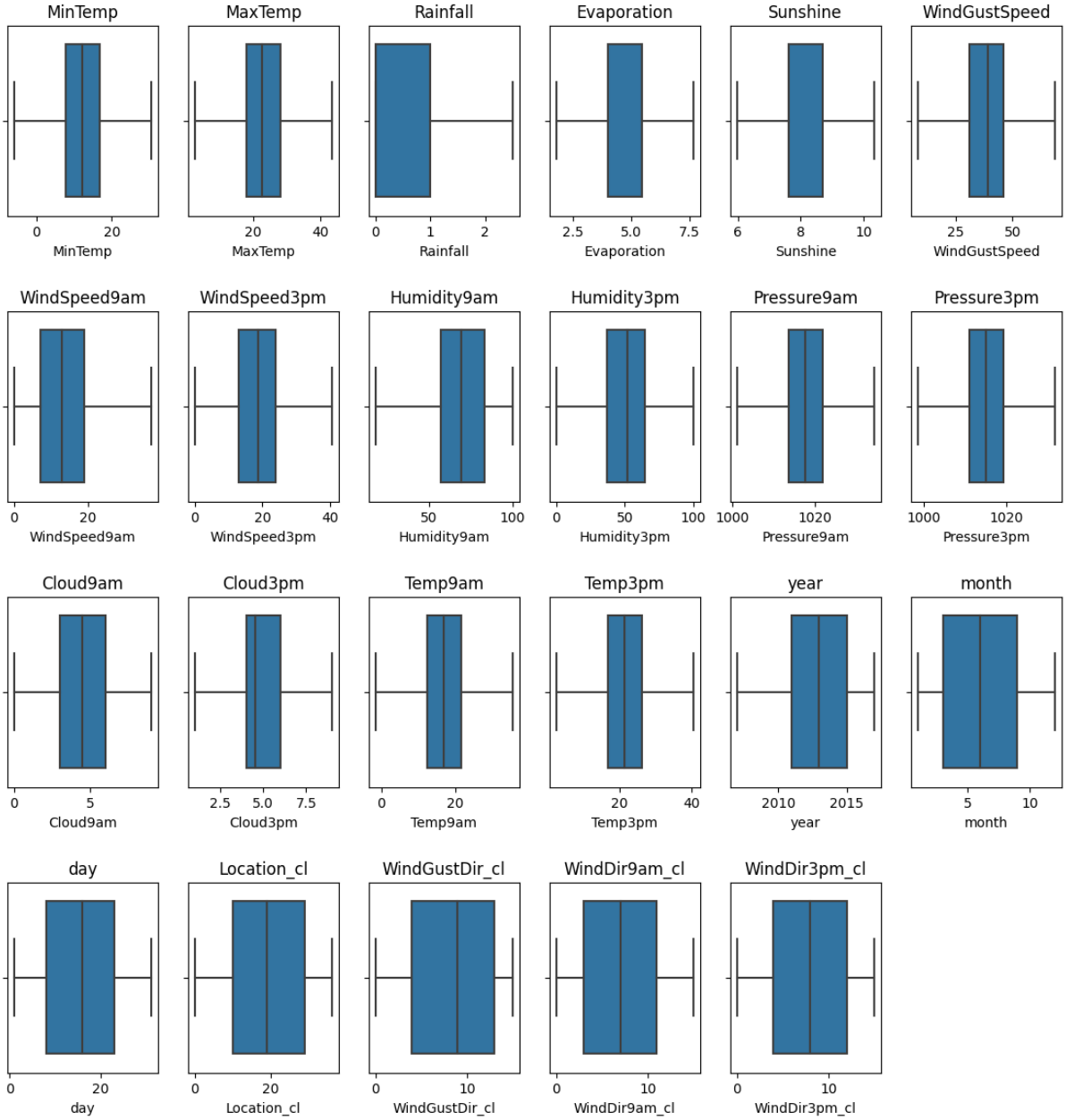


Interpretation:

1. Here we have used the box plot to find out the outliers from the datasets.

2. Here we can see that MinTemp,MaxTemp,Rainfall,Evaporation,Sunshine,Wind GustSpeed,WindSpeed9am,WindSpeed3pm,Humidity9am,Humidity3pm,Pressure9 am,Pressure3pm,Cloud9am,Cloud3pm,Temp9am,Temp3pm are the columns with outliers

Handling Outliers:



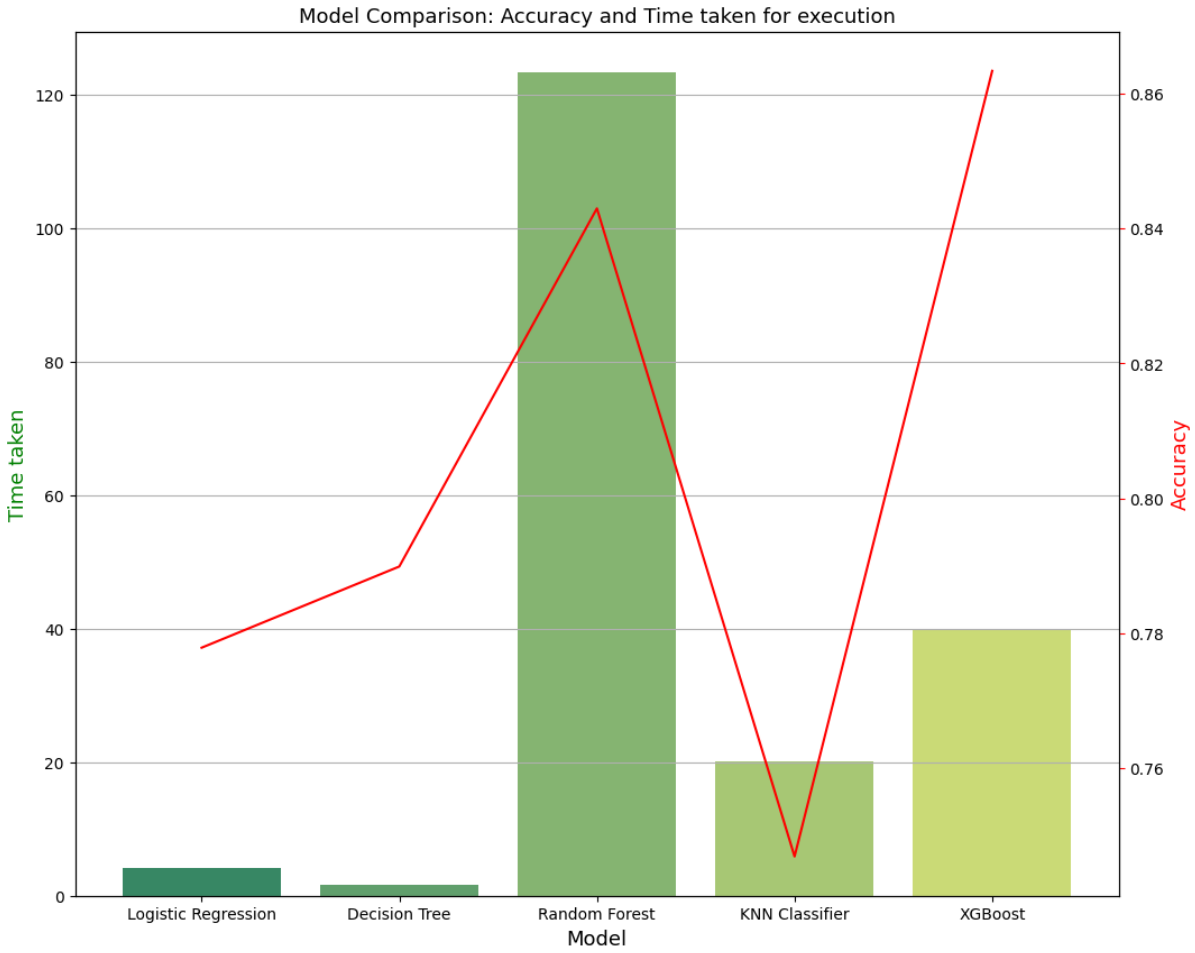
Interpretation:

1.Here we have implemented the IQR method for handling outliers.

2.Hence, we can see that the outliers have been handled successfully for the features shown in the above graph.

3.IQR uses method of identifying outliers to set up a “fence” outside of Q1 and Q3.

Results:



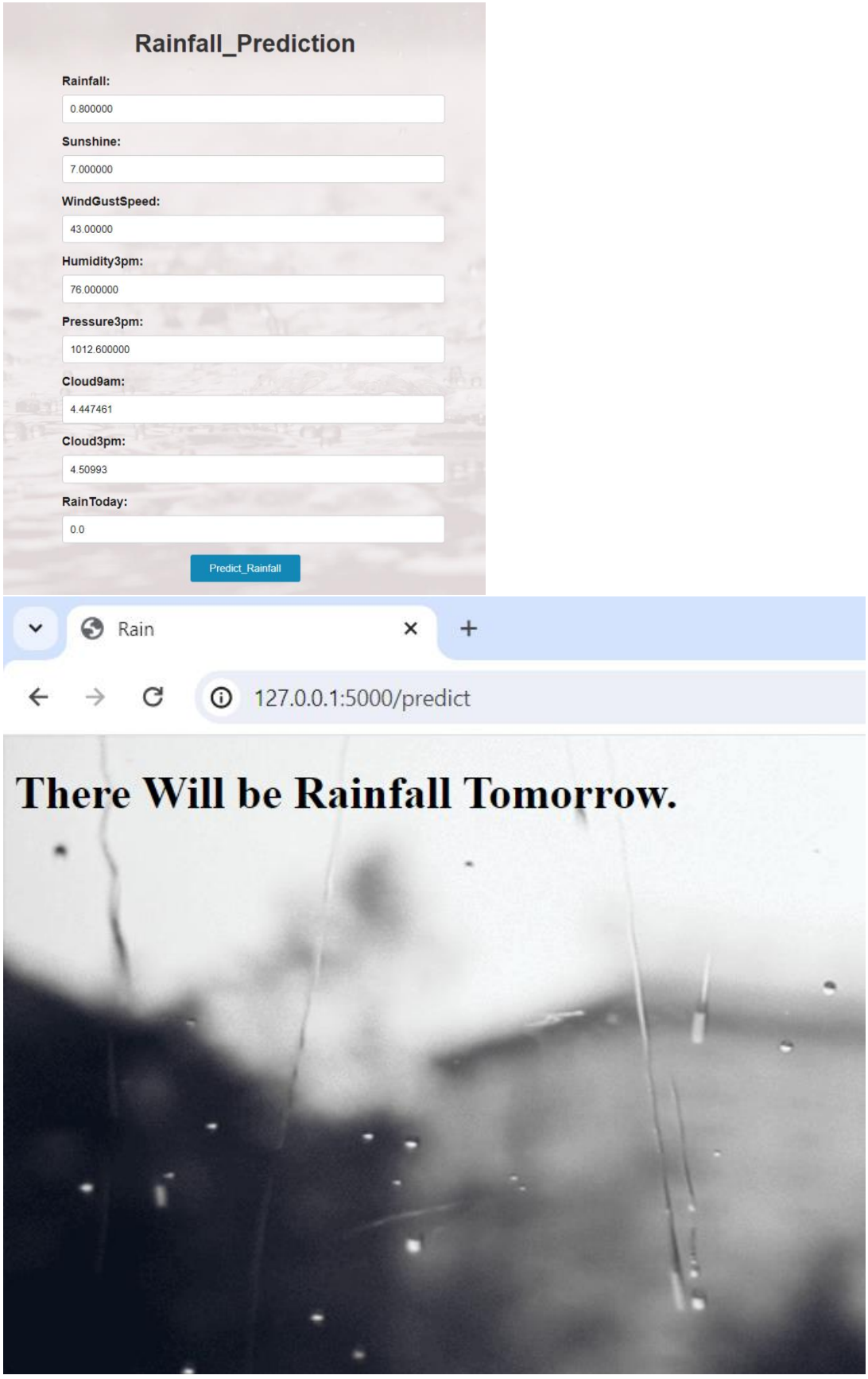
Interpretation:

1.The best performing model is the hyperparameter-tuned XGBoost model with an accuracy of approximately 86.5% also the time I took to build XGBoost model is also highest among all models followed by random forest, then decision tree the n Logistic Regression and at last it is KNN Classifier.

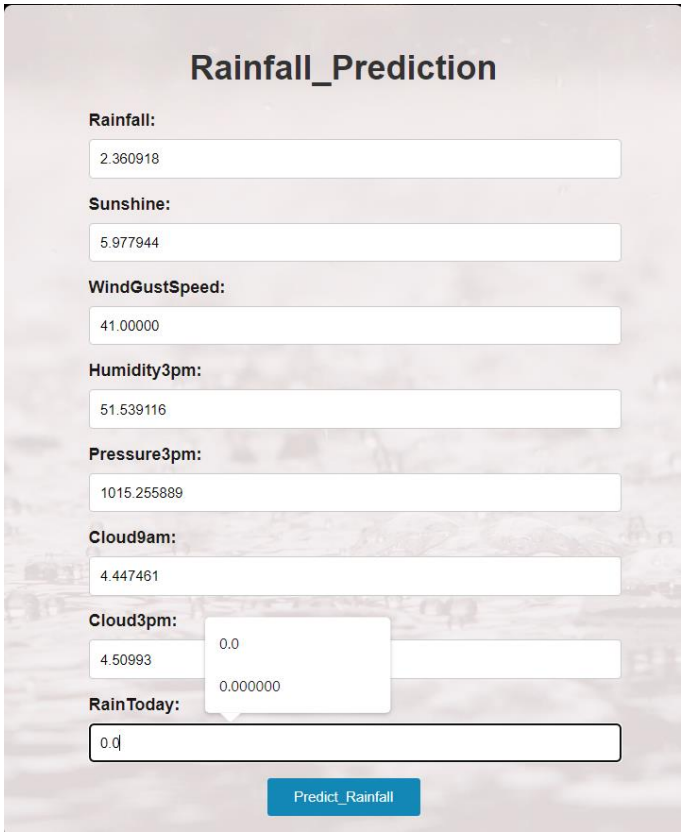
2.Least accuracy is shown by KNN classifier since KNN is also called as a lazy learner because it stores entire training dataset as a reference and it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.

3.Trainibg time is lowest for Decision Tree followed by Logistic Regression followed by KNN then Random Forest and Lastly XGBoost.

Working of Model: Predicting for Rainfall



Working of Model: Predicting for No – Rainfall



Conclusions:

- The proposed methodology used data from previous observations of Rainfall, Sunshine, Wind Direction, Humidity, Pressure, Cloud, Rain Today etc. to predict upcoming weather conditions.
- First, the methodology got rid of all the Nan values and apparently "cleaned" the data.
- Furthermore, visualized the data to find the correlation between different data fields. Then we implemented the divided the dataset into training and testing data.
- Performed the smote analysis.
- Finally, predicted the future values by training the data set and also predict one major condition, that is, rainfall.

Limitations:

- The data sample is limited to monthly statistics only and does not provide the daily output predictions.
- The climatic change and the global warming effect may impact the accuracy of the expected output.
- The locations for the data processing used in this study are geographically different and distanced that could also impact the correlation efficient that will measure the performance.

Future Scope:

- Further hyperparameter tuning
- Engineering new features such as trailing amounts of rain or sunshine
- Collecting additional data from nearby countries and attempting to predict the amount of rainfall
- Techniques like principal component analysis can also be used for dimensionality reduction.
- A richer data set with more features can be used for more accurate results.
- Data set with data of employees of different regions can be help full to understand turnover trends of different regions in the world

References:

Vikas Kumar, Vishal Kumar Yadav, Er. Sandeep Dubey https://www.ijraset.com/research-paper/rainfall-prediction- using-ml [3] Nikhil Oswal

https://www.researchgate.net/publication/336914968_Predicting_Rainfall_using_Machine_Learning_Techniques

Moulana Mohammed, Roshitha Kolapalli, +1 author S. Maturi https://www.semanticscholar.org/paper/Prediction-Of-Rainfall-Using-Machine-LearningMohammed- Kolapalli/9bf0a85ff6323559663cb45c89ae1e51684401dc