RAINFALL PREDICTION USING MACHINE LEARNING

Siddharth Linga
Dept of Computer Science
Texas A&M University-Corpus Christi
slinga1@islander.tamucc.edu

Tadikonda Samanthaka Manognya
Dept of Computer Science
Texas A&M University-Corpus Christi
mtadikonda@islander.tamucc.edu

Venkata Hitesh Kumar Akunuri
Dept of Computer Science
Texas A&M University-Corpus Christi
vakunuri@islander.tamucc.edu

Devi Sri Prabhas Nama
Dept of Computer Science
Texas A&M University-Corpus Christi
pnama@islander.tamucc.edu

Abstract

Rainfall prediction remains a serious concern and has attracted the attention of governments, industries, risk management entities, as well as the scientific community. Rainfall is a climatic factor that affects many human activities like agricultural production, construction, power generation, forestry and tourism, among others. To this extent, rainfall prediction is essential since this variable is the one with the highest correlation with adverse natural events such as landslides, flooding, mass movements and avalanches. These incidents have affected society for years. Therefore, having an appropriate approach for rainfall prediction makes it possible to take preventive and mitigation measures for these natural phenomena. To solve this uncertainty, we used various machine learning techniques and models to make accurate and timely predictions. The Proposed method aims to provide end to end machine learning life cycle right from Data pre-processing to implementing models to evaluating them. Data Preprocessing steps include imputing missing values, feature transformation, encoding categorical features, feature scaling and feature selection. We implemented models such as Logistic Regression, Decision Tree, Random Forest, XGBoost, Light .

CHAPTER 1

INTRODUCTION

Rainfall prediction remains a serious concern and has attracted the attention of governments, industries, risk management entities, as well as the scientific community. Rainfall is a climatic factor that affects many human activities like agricultural production, construction, power generation, forestry and tourism, among others. To this extent, rainfall prediction is essential since this variable is the one with the highest correlation with adverse natural events such as landslides, flooding, mass movements and avalanches. These incidents have affected society for years.

Therefore, having an appropriate approach for rainfall prediction makes it possible to take preventive and mitigation measures for these natural phenomena. To solve this uncertainty, we used various machine learning techniques and models to make accurate and timely predictions. The Proposed method aims to provide end to end machine learning life cycle right from Data pre-processing to implementing models to evaluating them. Data Preprocessing steps include imputing missing values, feature transformation, encoding categorical features, feature scaling and feature selection. We implemented models such as Logistic Regression, Decision Tree, Random Forest, XGBoost, Light GBM.

1.1 Motivation

Predicting rainfall is an application of science and technology for predicting the amount of rain over an area. The most important thing is to accurately determine the rainfall for active use of rainfall for water resources, crops, pre-planning of water resources and for agricultural purposes. In earlier rainfall information benefits the farmers for better managing their crops and properties from heavy rainfall. The farmers better manage to increase the economic growth of the country by efficient rainfall information. Prediction of precipitation is necessary to save the life of people's and properties from flooding. Prediction of rainfall helps people in coastal areas by preventing the floods.

1.2 Problem Statement

Prediction of Rainfall is well known to be hard. There are several reasons for this, ranging from the heterogeneity and complexity for considering complex features like Climatic Conditions such as Temperature ,Humidity, Wind speed etc to the adoption of undisciplined weather conditions offering dubious report is hard according to the statistical report. In such a setting Machine Learning algorithms are particularly popular. As opposed to any other techniques which require storing of data in database and retaining them makes harder, machine learning methods operate them at ease. Though this limited perspective might miss important

insights, it has the key advantage of offering an accuracy percentages which abstracts from the complexity of normal approaches and offers a uniform interface to the widest possible range of predicting the rainfall . The key concept lies in finding the best accurate algorithm that fit to the model and predict the rainfall by training the model with the most accurate algorithm.

1.3 Project Objectives

The aim of the study is the prediction of the rainfall using Australian dataset by comparing machine learning methodologies such as Random Forest, Decision Tree Classifier, XGBoost, Logistic Regression, Light GBM. The extraction procedures/algorithms will produce the output by classification of the data. The data gets trained with most accurate algorithm and predict rainfall more correctly and with perfect figures.

The accurate and exact predictions will help in developing more appropriate strategies for agriculture and water reserves and will also be informed to the users if it will rain the next day or not. Accuracy of rainfall forecasting has great importance for countries like India whose economy is largely dependent on agriculture. Thus ,the objective of this survey is to make the prediction of rainfall more accurate in the recent future and also to predict the rainfall values for non linear data.

1.4 Organization of Project

The reminder of this document is first providing the full description of the project. It lists all the functions performed by the system. And it also concerns the details of the system functions and actions of each function which was performed by the system.

Chapter 1: Describes about the Introduction

Chapter 2: Describes about the Literature survey

Chapter 3: Describes about the Software and Hardware Specifications of the project

Chapter 4: Describes about the Proposed Design of the Project

Chapter 5: Describes about the Implementation and Testing of the Project

Chapter 6: Provides the Conclusion and Future Enhancement of the project

CHAPTER 2

LITERATURE SURVEY

2.1 Existing Work

2.1.1 Rainfall prediction using Extreme Gradient Boosting

In this paper the technique used is Extreme Gradient Boosting (XGBoost) it is used an ensemble learning method. It is relying on the results of a single machine learning model such as J48 which is not enough. Although J48 is good enough and had been used in several problems such as wildfire modelling and rain modelling. The Ensemble learning combines multiple learners to get a more powerful prediction. The results also show that the factors that most influence rainfall are the average humidity and the minimum temperature.

2.1.2 Long-Term Rainfall Forecast Model Based on the TabNet and LightGbm Algorithm

In this paper the technique used is TabNet and LightGbm Algorithm.TabNet is a mechanism in which it makes possible to explain how the model arrives at it's predictions and helps it learn more accurate models. using feature fusion, mining the potential value of each feature to improve the upper limit of the model, using the Borderline SMOTE algorithm to improve the imbalance of the data set. Adversarial verification is used to improve the distribution difference between the training set and the test set.Borderline SMOTE algorithm is improved the imbalance of the data set. This result proves the reliability of using the hybrid model of TabNet and LightGbm to predict rainfall.

2.1.3 Prediction of Rainfall Using Machine Learning Techniques

In this paper the techniques used are Lasso Regression,Multiple Linear Regression and Support Vector Regression.Regression Analysis is deals with the dependence of one variable on one or more other variables which helps in prediction. SVR is a valuable and adaptable strategy, helping the client to manage the impediments relating to distributional properties of fundamental factors, geometry of the information and the normal issue of model over fitting. SVR is better than MLR as an expectation strategy. MLR can't catch the non-linearity in a data set and SVR winds up helpful in such circumstances.

2.1.4 Efficient Rainfall Prediction and Analysis using Machine Learning Techniques

In this paper the techniques used are Random Forest and Logistic Regression. Both algorithms performed well depending on their technique and evaluated with great accuracy, high speed in less time. Results have shown that both algorithms performed well.The accuracy for the logistic regression algorithm is slightly more efficient than the random forest algorithm.

2.1.5 Prediction of Rainfall using Logistic Regression

In this paper the technique used is Logistic Regression. Logistic regression allows one to predict a discrete outcome. The outliers in logistic regression can severely affect the fitting of the model as their omission can turn a poorly fitted model to be a model of good fit.The results show that Logistic Regression model can predict the rainfall accurately. The table-2.1 compares the Existing Systems with the proposed System.

2.2 Limitations of Existing Work

Rainfall prediction using Extreme Gradient Boosting paper used Extreme Gradient Boosting method. It has a capability to estimate the rainfall with training RMSE of 2.7mm. This is limited to only certain areas which may not be as useful as it claims to be. It also does not consider all features responsible for rainfall, which can also lead to overfitting.

The paper Long-Term Rainfall Forecast Model Based on The TabNet and LightGbm Algorithm haven't even explored the data for many years, i.e. it did not consider larger dataset. In this paper, Forecasting the rainfall was based on TabNet and LightGbm Algorithm. They found better accuracy using the LightGbm, that too of 91% where it can predict the area's rainfall very well. But then, it is limited to only a single area Beijing, China.

Prediction of Rainfall Using Machine Learning Techniques paper mainly deals with only regression models (i.e. Multiple Linear Regression, Support Vector Regression). The dataset was accurate and considered 19 attributes and also collected the data for individual months. This paper concluded that SVM is the best model for predicting rainfall. But, SVM couldn't deal with large datasets and can't perform well when the dataset has more target classes.

Efficient Rainfall Prediction and Analysis using Machine Learning Techniques paper uses Random Forest and Logistic Regression as the algorithms for predicting rainfall. This paper takes into account of almost all the features that are responsible for predicting the rainfall, but outliers weren't taken into account while pre-processing the data which will lead to loss of data and prediction can't be accurate.

Prediction of Rainfall Using Logistic Regression paper uses Logistic Regression as the algorithm for predicting rainfall. This project studied various analytic methods for dealing with spurious observations and dealt them with clustering techniques.This paper considered only few parameters and also limited in predicting rainfall in single area.

CHAPTER 4

PROPOSED SYSTEM DESIGN

4.1 Proposed methods

Proposed System Predicts the Rainfall Using The Higher accurate Algorithm. Proposed System compares 5 algorithms namely Random Forest , Decision Tree, Logistic Regression , XG Boost, Light GBM and finds out the higher accuracy algorithm based on the accuracy score. And as a Limitation, Parameters /Features Considered in Existing System are less .Proposed System considers many features as parameters in predicting the rainfall Such as(Location , Temperature , Humidity, Pressure).

The Fig.4.1 diagram shows the methodology of the project

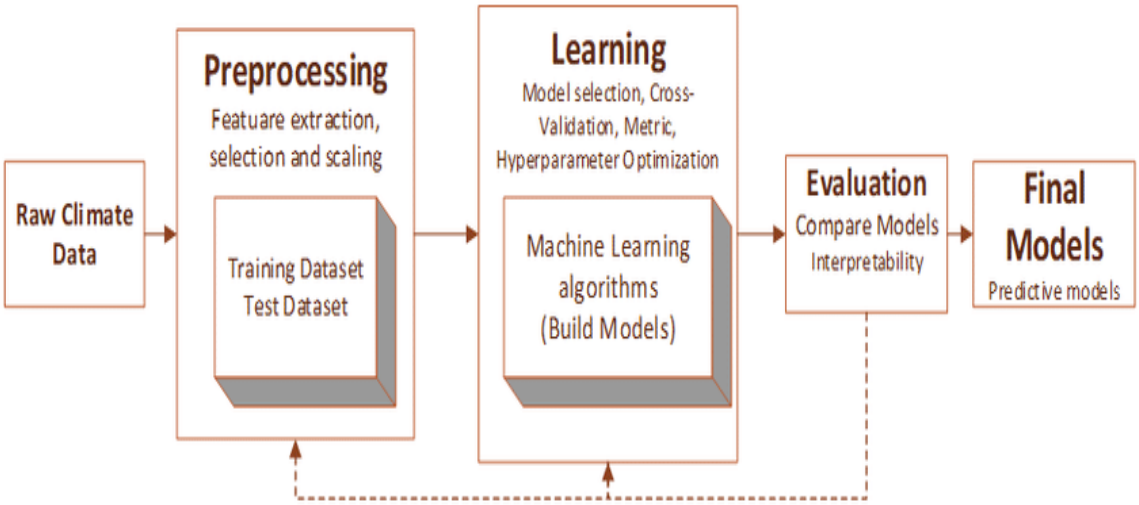


Fig. 4.1 Methodology Diagram

Software And Hardware Specifications

Hardware Requirements

Operating System	:	Windows 7 or above
Processor	:	15 INTEL
Hard disk	:	500GB or above
RAM	:	8GB or above

Software Requirements

Programming Language	:	Python
IDE	:	Jupyter Notebook , Spyder
UML Design	:	Star uml

4.2 System Architecture

System architecture depicts the flow of the entire model. This project at first collects the data, pre-process it by various methods and remove null values, and then testing, training is done later algorithms are applied to the model. Based on the accuracy produced by algorithms, the best accurate algorithm is picked.

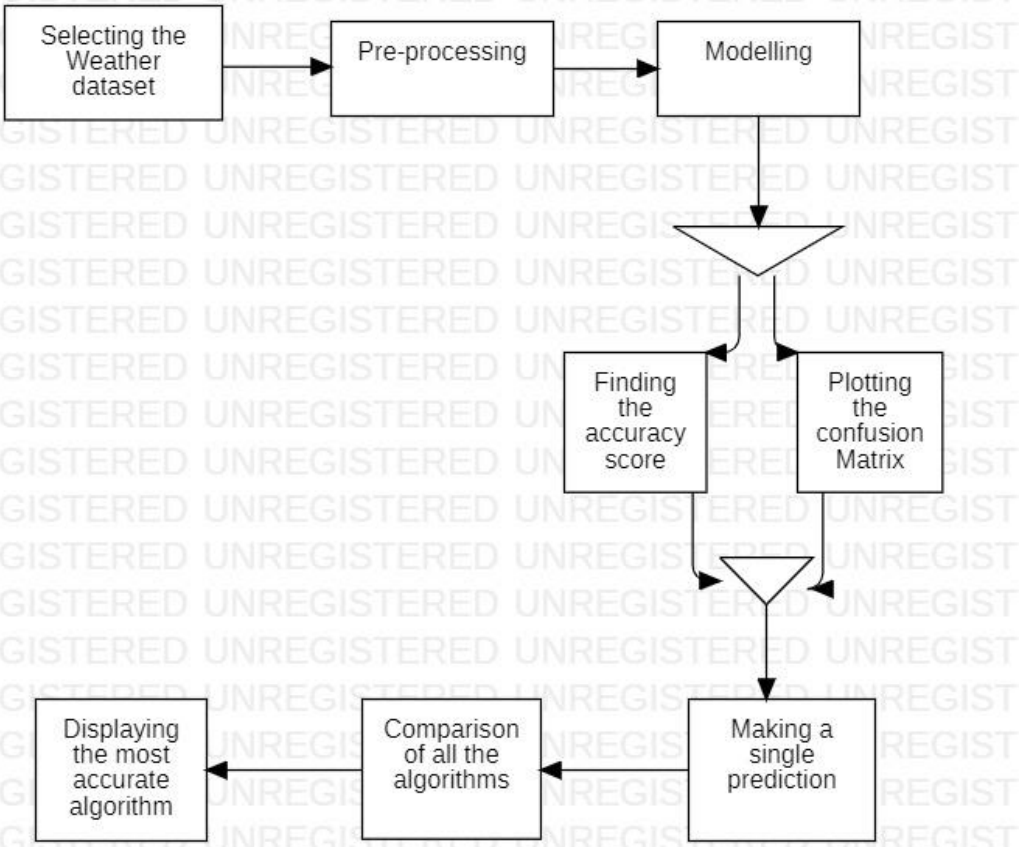


Fig. 4.2 System Architecture

After predicting the Highest Accuracy, The model gets trained based on the more accurate algorithm and developed a Web App using Stream Lit where user can enter the features such as area, temperature, humidity.. and find out whether there will be a rain the next day.

4.3 UML Diagrams

4.3.1 Use Case Diagram

Use case diagram can summarize the details of your system's users (also known as actors) and their interactions with the system. To build one, you'll use a set of specialized symbols and connectors. An effective use case diagram can help your team discuss and represent:

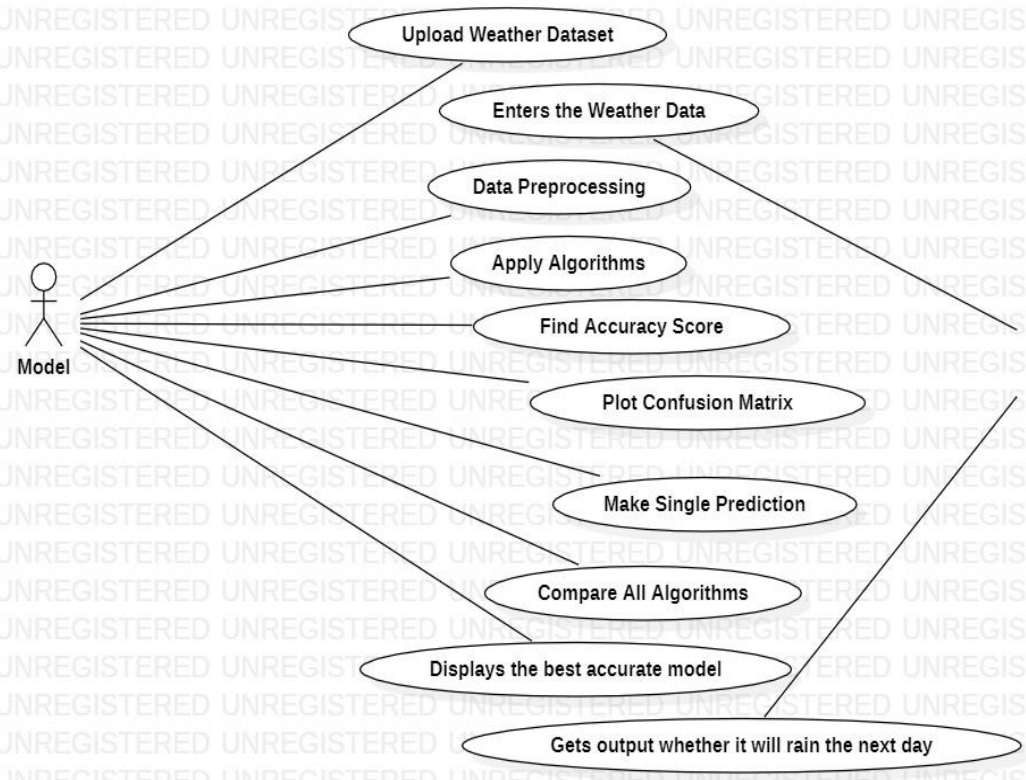


Fig. 4.3 Use Case Diagram

CHAPTER 5

IMPLEMENTATION AND TESTING

5.1 Implementation modules

The primary objective of this project is to analyse and study various regression models that can overcome the existing model constraints and predict the higher accurate algorithm for the model. It was expanded to create a user-friendly web app for prediction. Innovative techniques of ML are used for predicting the rainfall in terms of accuracy and other visualization methods. This methodology contains many steps for predicting the rainfall. The steps are mainly divided as parts and the pipeline is as follows.

1.Exploratory Data Analysis

i. Data Exploration
ii. Visualisation

2. Data Pre-processing

i. Class Imbalance
ii. Feature Selection
iii. Missing Values
iv. Encoding Categorical Values
v. Feature Scaling

3. Modelling

i. Building
ii. Training
iii. Testing

5.1.1 Exploratory data Analysis

Data Collection:

Data is collected and observed based on the availability of weather forecast in different areas of Australia, Project used binary classification of dataset collected by the Kaggle website.It almost contains 1 lakh records and 23 features. Available and presence of this data is better for conducting the research.

Table 5.1 The numerical features that have been considered in the data set.

ATTRIBUTE NAMES	OBJECT TYPE	UNITS
Date	predictor	DD/MM/YYYY
Min Temperature	predictor	Degree Celsius
Max Temperature	predictor	Degree Celsius
Rainfall	predictor	In mm
Evaporation	predictor	In mm
Sunshine	predictor	In mm
WindSpeed9am	predictor	Meters per Second
Wind Speed3pm	predictor	Meters per Second
Humidity9 am	predictor	%
Humidity3pm	predictor	%
Pressure9am	predictor	Millimetres of mercury
Pressure3pm	predictor	Millimetres of mercury
Cloud9am	predictor	Precipitation amount
Cloud3pm	predictor	Precipitation amount
Temp 9 am	predictor	Degree Celsius
.Temp 3 pm	predictor	Degree Celsius

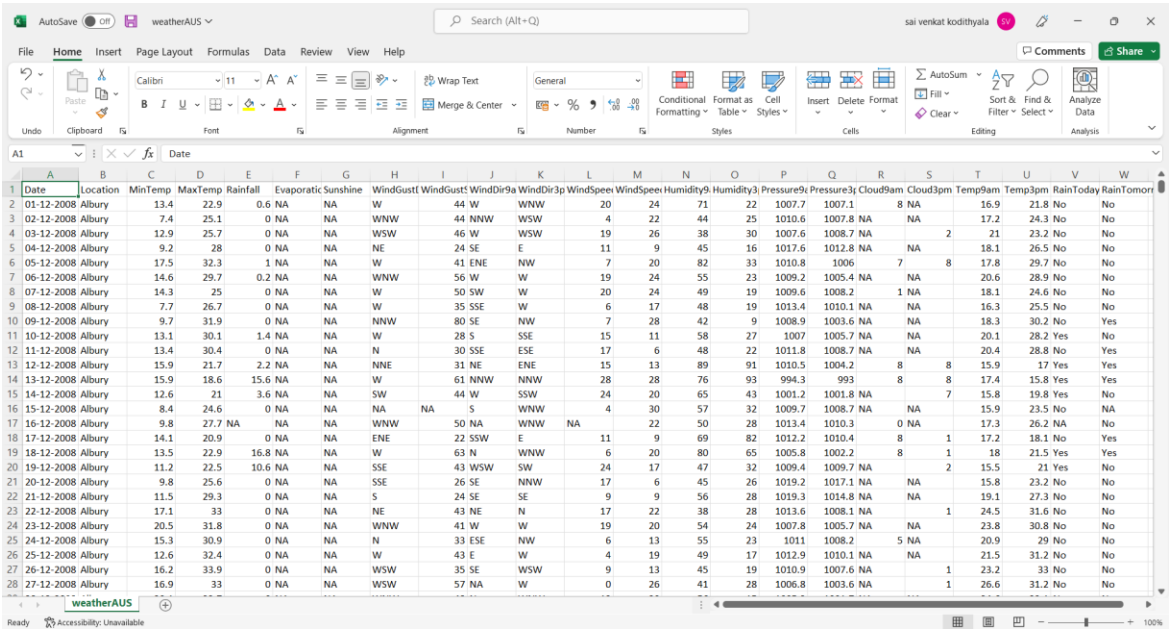
Wind Gust Speed	predictor	Meters per second

Table 5.2 categorical features that have been considered in dataset.

Location	predictor	Area
Wind Gust Dir	predictor	Direction name
Wind Dir 9 am	predictor	Direction name
Wind Dir 3 pm	predictor	Direction name
Rain Today	predictor	Categorical value
Rain Tomorrow	Target/Response	Categorical value

The above dataset contains float and string values. A dataset is further divided into training and testing dataset:80% of it as training, and 20% as testing. To get the value for the target this work chooses 22 columns for regression models and 1 column is for the target variables.

The Fig.5.1 diagram is the sample screenshot of the dataset that is considered in the project.



Date	Location	MinTemp	MaxTemp	Rainfall	Evaporatic	Sunshine	WindGustI	WindGustS	WindDir9a	WindDir3p	WindSpeed	WindSpeed	Humidity9	Humidity3	Pressure9	Pressure3	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorr
01-12-2008	Albury	13.4	22.9	0.6	NA	NA	W	44	W	WNW	20	24	71	22	1007.7	1007.1	8	NA	16.9	21.8	No	No
02-12-2008	Albury	7.4	25.1	0	NA	NA	WNW	44	NNW	WSW	4	22	44	25	1010.6	1007.8	NA	NA	17.2	24.3	No	No
03-12-2008	Albury	12.9	25.7	0	NA	NA	WSW	46	W	WSW	19	26	38	30	1007.6	1008.7	NA	2	21	23.2	No	No
04-12-2008	Albury	9.2	28	0	NA	NA	NE	24	SE	E	11	9	45	16	1017.6	1012.8	NA	NA	18.1	26.5	No	No
05-12-2008	Albury	17.5	32.3	1	NA	NA	W	41	ENE	NW	7	20	82	33	1010.8	1006	7	8	17.8	29.7	No	No
06-12-2008	Albury	14.6	29.7	0.2	NA	NA	WNW	56	W	W	19	24	55	23	1009.2	1005.4	NA	NA	20.6	28.9	No	No
07-12-2008	Albury	14.3	25	0	NA	NA	W	50	SW	W	20	24	49	19	1009.6	1008.2	1	NA	18.1	24.6	No	No
08-12-2008	Albury	7.7	26.7	0	NA	NA	W	35	SSE	W	6	17	48	19	1013.4	1010.1	NA	NA	16.3	25.5	No	No
09-12-2008	Albury	9.7	31.9	0	NA	NA	NNW	80	SE	NW	7	28	42	9	1008.9	1003.6	NA	NA	18.3	30.2	No	Yes
10-12-2008	Albury	13.1	30.1	1.4	NA	NA	W	28	S	SSE	15	11	58	27	1007	1005.7	NA	NA	20.1	28.2	Yes	No
11-12-2008	Albury	13.4	30.4	0	NA	NA	N	30	SSE	ESE	17	6	48	22	1011.8	1008.7	NA	NA	20.4	28.8	No	Yes
12-12-2008	Albury	15.9	21.7	2.2	NA	NA	NNE	31	NE	ENE	15	13	89	91	1010.5	1004.2	8	8	15.9	17	Yes	Yes
13-12-2008	Albury	15.9	18.6	15.6	NA	NA	W	61	NNW	NNW	28	28	76	93	994.3	993	8	8	17.4	15.8	Yes	Yes
14-12-2008	Albury	12.6	21	3.6	NA	NA	SW	44	W	SSW	24	20	65	43	1001.2	1001.8	NA	7	15.8	19.8	Yes	No
15-12-2008	Albury	8.4	24.6	0	NA	NA	NA	NA	S	WNW	4	30	57	32	1009.7	1008.7	NA	NA	15.9	23.5	No	NA
16-12-2008	Albury	9.8	27.7	NA	NA	NA	WNW	50	NA	WNW	NA	22	50	28	1013.4	1010.3	0	NA	17.3	26.2	NA	No
17-12-2008	Albury	14.1	20.9	0	NA	NA	ENE	22	SSW	E	11	9	69	82	1012.2	1010.4	8	1	17.2	18.1	No	Yes
18-12-2008	Albury	13.5	22.9	16.8	NA	NA	W	63	N	WNW	6	20	80	65	1005.8	1002.2	8	1	18	21.5	Yes	Yes
19-12-2008	Albury	11.2	22.5	10.6	NA	NA	SSE	43	WSW	SW	24	17	47	32	1009.4	1009.7	NA	2	15.5	21	Yes	No
20-12-2008	Albury	9.8	25.6	0	NA	NA	SSE	26	SE	NNW	17	6	45	26	1019.2	1017.1	NA	NA	15.8	23.2	No	No
21-12-2008	Albury	11.5	29.3	0	NA	NA	S	24	SE	SE	9	9	56	28	1019.3	1014.8	NA	NA	19.1	27.3	No	No
22-12-2008	Albury	17.1	33	0	NA	NA	NE	43	NE	N	17	22	38	28	1013.6	1008.1	NA	1	24.5	31.6	No	No
23-12-2008	Albury	20.5	31.8	0	NA	NA	WNW	41	W	W	19	20	54	24	1007.8	1005.7	NA	NA	23.8	30.8	No	No
24-12-2008	Albury	15.3	30.9	0	NA	NA	N	33	ESE	NW	6	13	55	23	1011	1008.2	5	NA	20.9	29	No	No
25-12-2008	Albury	12.6	32.4	0	NA	NA	W	43	E	W	4	19	49	17	1012.9	1010.1	NA	NA	21.5	31.2	No	No
26-12-2008	Albury	16.2	33.9	0	NA	NA	WSW	35	SE	WSW	9	13	45	19	1010.9	1007.6	NA	1	23.2	33	No	No
27-12-2008	Albury	16.9	33	0	NA	NA	WSW	57	NA	W	0	26	41	28	1006.8	1003.6	NA	1	26.6	31.2	No	No

Fig. 5.1 Dataset used in the model

5.1.2 Data pre-processing

For eliminating the inconsistencies or duplicates in data we use pre-processing it features the better compatibility with the model. After collecting of data this step is done, the pre-processing step considers four other steps:

i. **Feature Selection**

Feature selection is the technique to decrease number of input variables when developing a predictive model. In this project feature selection is done based on the exploratory data analysis. This work considered the attribute “Rain Tomorrow” as a target variable (Y) since it need to be predicted. Project mainly considered the data that really needs to it i.e. Evaporation, Sunshine which are predictors (X) have been eliminated because of their high percentage of missing values or outliers.

ii. Handling the Outliers

In order to handle the outliers or noisy data, which make the data vulnerable to fit and result in bad accuracy of the model should be eliminated and this can be done by using the Simple Imputer.

iii. Dealing with the Categorical value

Categorical values are that which have string values or object values as their input. These include Location, WindGustDir, WindDir9am, WindDir3pm, RainToday.As machines do not define their understandability to the Strings and process them. Therefore, it’s time to encode the data with Label Encoder.

iv. Feature Selection

Data that have been considered had diverse range in their size /proportions. Most of the ML algorithms use Euclidean distance in-order to reduce the distance between two data points ,but their was problem lied .The features with higher in their size will weigh much large in the distance calculations than with the lower size data. In order, to compress this effect, we have to make the features come along with the same size. This can be achieved and done in the project by considering Standard Scaler.

5.1.3 Data Modelling

In this step the project first divided the data into training and testing based on the pareto standard principle (80-20) rule ,then we implement the machine learning models .This step is more important and crucial step in supervising the model .This step makes independent variable to predict the data based on the input variable(i.e maintaining the regression).The prediction algorithms mainly deal with the accuracy. This proposed method implements 5 machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Light GBM, XGBOOST. In order to control the efficiency and accuracy this step is mandatory and must be done. A proper study and review by some parameters on these 5 algorithms lead to find the better accuracy algorithm.The next section briefly illustrates about the features of every technique.

5.2 Algorithms

5.2.1 Logistic Regression Classifier

Logistic Regression Classifier is a supervised learning technique, which deals with the prediction of the target variable by considering the predictor values. This prediction or the outcome will be a categorical value such as yes/no,0 or 1, true or false etc. Here ,it represent s yes/no .In this we fit the s shaped logistic function with the maximum values ranging from(0 , 1) .Log odds are the one which refers to the likelihood of the independent variable “Rain Tomorrow”. As dependent variables are discrete, so we use a sigmoid function in order to predict it.

Mathematical steps that get logistic Regression equations are:

1) Equation of the straight line.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

2) As y (independent variable) can only range between 0 and 1 so divide with the above equation by (1-y).

$$\frac{y}{1-y}; 0 \; for \; y = 0, and \; infinity \; for \; y = 1$$

3) But we need range between -[infinity] to +[infinity], algorithm of the equation is as follows:

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

This is the final equation to predict which is sigmoid function

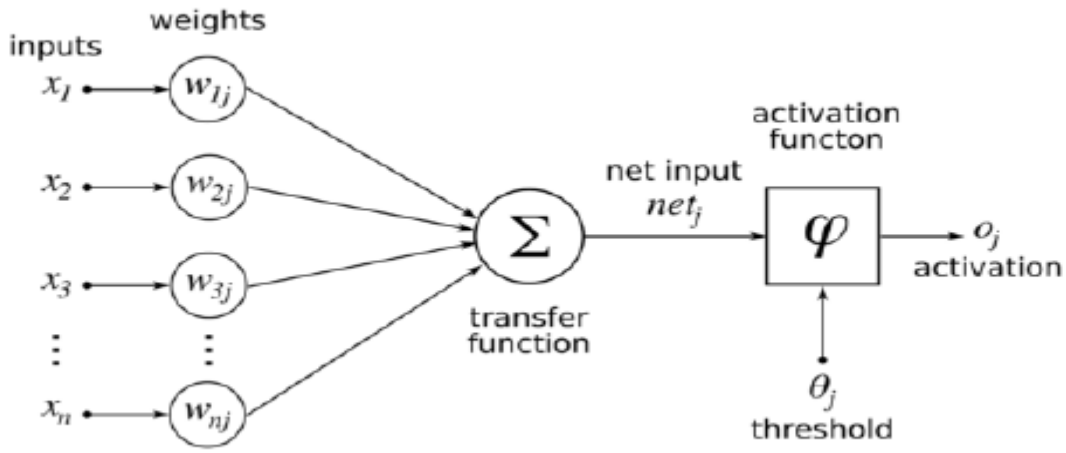


Fig. 5.3 Logistic Regression Classifier

5.2.2 Decision Tree Classifier

This classifier is used for both regression and classification problems. It always depends upon the condition of the data, simply as an “ifelse” statement .It always execute the if statement when the condition is true otherwise ,it will always outputs when decisions are to be made .The procedure for this algorithm follows,

Step-1: Collecting the weather dataset where the target variable need to be predicted by the input /predictor variables

Step-2: Splitting the dataset into training and testing is done here.

Step-3: Measuring the IG during the Training process

Step-4: Train the model continuously until it gets completed.

Step-5: After this process ,leaf nodes are created which represents the classifier predictions.

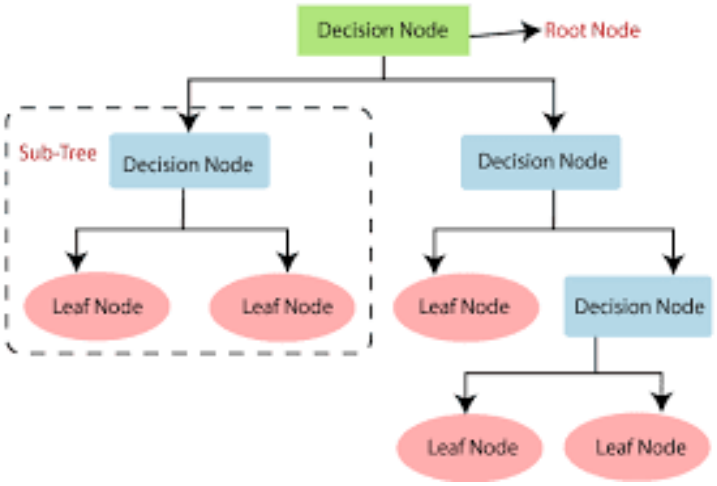


Fig. 5.4 Decision Tree Classifer

5.2.3 Random Forest Classifier

It is a supervised learning technique which do not consider parameters, It is mainly a set of single decision trees for best accuracy. The main con to random forest classifier is, it can able to serialize the ML model with less parameters and make the model more accurate. The process includes following steps:

Step-1: Select ‘x’ no of random samples from the weather data set.

Step-2: For every random sample, individual decision trees are formed.

Step-3: For every tree generated there will be a predicted output.

Step-4: Finally,the voting will be performed for each decision tree, and the majority of the voting in the step-3 is considered for final forecasting.

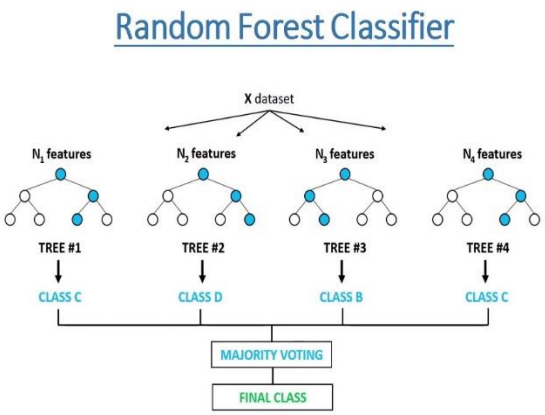


Fig. 5.5 RandomForest Classifier

5.2.4 Boosting Algorithms

Boosting Algorithm is nothing but Ensemble method .This is a powerful technique which selects the exact classification or regression model from different number of incorrect classifications.

i. Light Gradient Boosted Machine (LGBM)Classifier

This Algorithm was considered because of its exceptional performance and the perfectness in solving the dataset problems. It is based on the open source framework based on the generation of decision trees and always deals with efficacy and efficiency to work .This algorithm use low memory and capable of handling the large-scale data. It always deals with the properties like Histogram Splitting, Bagging, Goss etc. It mainly applies leaf-wise tree growth, significant for classifying and ranking. It is designed mainly for the purpose of hybrid model. LGBM becomes the best choice when memory requirement, speed, and arithmetic memory requirement are considered. It accelerates the training process, improves efficiency, optimizes memory, does efficient ICU utilization, and enhances accuracy.

ii. Extreme Gradient Boosting (XGB) Classifier

XGB is a variation of boosting algorithm that implements an innovative tree search technique. This unique technique is used for generating the forecasting models when regression model comes into play. It is mainly designed for processing large datasets and offers an efficient and recursive solutions for optimizing the new problems.

LGBM and XGB are similar to each other ,but the only feature that lags out is LGBM uses level-wise tree growth .Because of this it is faster than the XGB and also XGB is considered as the memory-intensive. Even with these many pros the XGB outstands its performance and state-of-art feature as a gradient boosting algorithm.

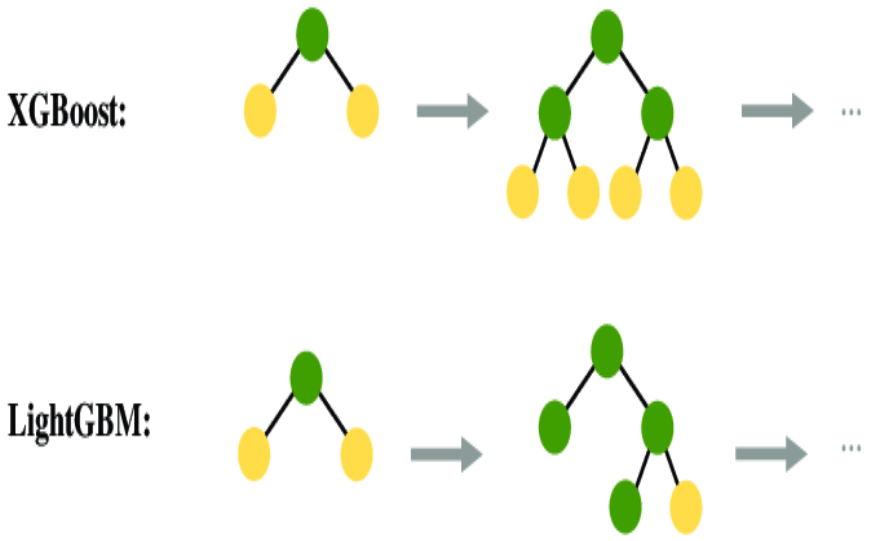


Fig. 5.6 LightGBM and XGBoost Classifiers

5.3 Evaluation metrics

5.3.1 Confusion matrix

This matrix defines the independent variable or target variable based on the predictor values based on the algorithms performed. This can be seen as to how well the classification model is working.

It includes Four Categories:

1.TP: the actual output is YES

2.TN: the actual output is NO

3.FP: the actual output is NO

4.FN: the actual output is YES

5.3.2 Accuracy

It is the ration of correct predictions to the total number of input samples.

Calculation as:

Acc= $(TP + TN)/(TP + FN + TN + FP)$

Where Acc-Accuracy, TP-True positive, TN-True Negative, FN-False Negative, FP-False Positive.

5.3.3 Precision

It displays the ration between the correctly predicted positive results to the total predicted positive results by the classifier

Calculated as:

Precision=(TP) / (TP+FP)

5.3.4 Sensitivity

It displays the ratio between the correctly predicted positive results to the total no of true positive results and false negative results.

Calculated as :

Recall=(TP) / (TP+FN)

5.3.5 F1-score

It is predicted by the harmonic mean of precision and recall and the range of F1-score always varies from 0 to 1

Calculate as:

F1-score= (2) / (1/Recall)*(1/Precision)

5.4 Testing

5.4.1 Types of Testing Techniques:

1) Functional testing:

This type of testing is done against the functional requirements of the project.

Types:

1. Unit testing: Each unit /module of the project is individually tested to check for bugs. If any bugs found by the testing team, it is reported to the developer for fixing.

2. Integration testing: All the units are now integrated as one single unit and checked for bugs. This also checks if all the modules are working properly with each other.

3. System testing: This testing checks for operating system compatibility. It includes both functional and non functional requirements.

4. Sanity testing: It ensures change in the code doesn’t affect the working of the project.

5. Smoke testing: this type of testing is a set of small tests designed for each build.

6. Interface testing: Testing of the interface and its proper functioning.

7. Regression testing: Testing the software repetitively when a new requirement is added, when bug fixed etc.

8. Beta/Acceptance testing: User level testing to obtain user feedback on the product.

2) Non-functional testing

This type of testing is mainly concerned with the non-functional requirements such as performance of the system under various scenarios.

1. Performance testing: Checks for speed, stability and reliability of the software, hardware or even the network of the system under test.

2. Compatibility testing: This type of testing checks for compatibility of the system with different operating systems, different networks etc.

3. Localization testing: This checks for the localized version of the product mainly concerned with UI.

4. Security testing: Checks if the software has vulnerabilities and if any, fix them.

5. Reliability testing: Checks for the reliability of the software.

6.Stress testing: This testing checks the performance of the system when it is exposed to different stress levels.

7.Usability testing: Type of testing checks the easily the software is being used by the customers.

8. Compliance testing: Type of testing to determine the compliance of a system with internal or external standards.

5.4.2 Test Cases

Table 5.3 Test Cases

SNO	Test Case	Expected Result	Observed Result	Result
1	Huge amount of data for training	The model can evaluate large amount of data, i.e model can have and achieve high - accuracy, availability.	The model can predict accurately with large data.	**Pass**
2	Clearing/Removing noisy data, null values	The proposed model can remove the noisy data and effectively manage the null values.	The outliers are removed and the null values are managed.	**Pass**
3	Considerable taking the missing values	This model can handle missing values by replacing them with the mode of the data.	We observed that missing values are replaced with most frequent value.	**Pass**
4	Taking account of string values	It considers String values as an input and this is converted by encoder.	Using encoder techniques, the string values are converted into numerical values	**Pass**
5	Managing large amount of categorical values	Feature Hashing is one of engineering scheme for dealing with large scale categorical features. This technique is done for WinDir3pm feature in our model.	The large amount of categorical values are managed well.	**Pass**
6	Accuracy of the model	This model should give the accuracy of 90% and it should be able to predict accurately	This model is showing accuracy of 85% and predicting accurately.	**Pass**

5.5 Results

The Fig.5.8,Fig.5.9,Fig.5.10,Fig.5.11,Fig.5.12 show us the confusion matrices of algorithms Random Forest, Logistic Regression, Decision Tree, LightGBM and XGBoost respectively. They help us to arrive at an accurate prediction.

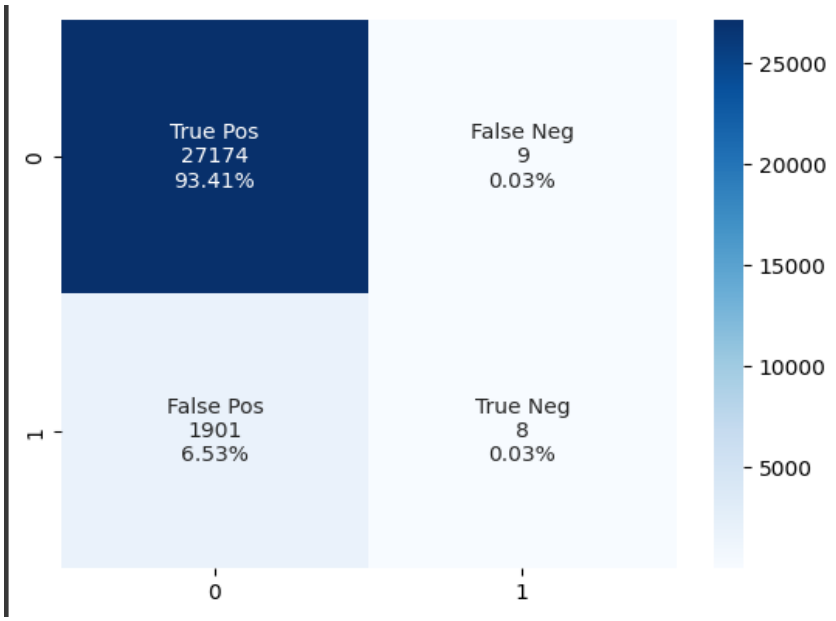


Fig. 5.8 Random Forest Confusion Matrix

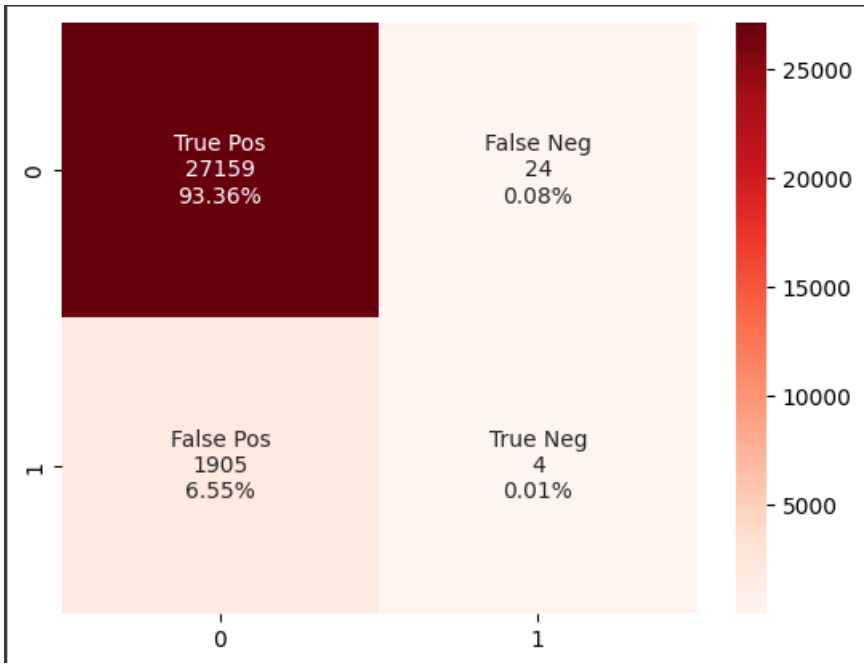


Fig. 5.9 Logistic Regression Confusion Matrix

Fig. 5.10 Decision Tree Confusion Matrix

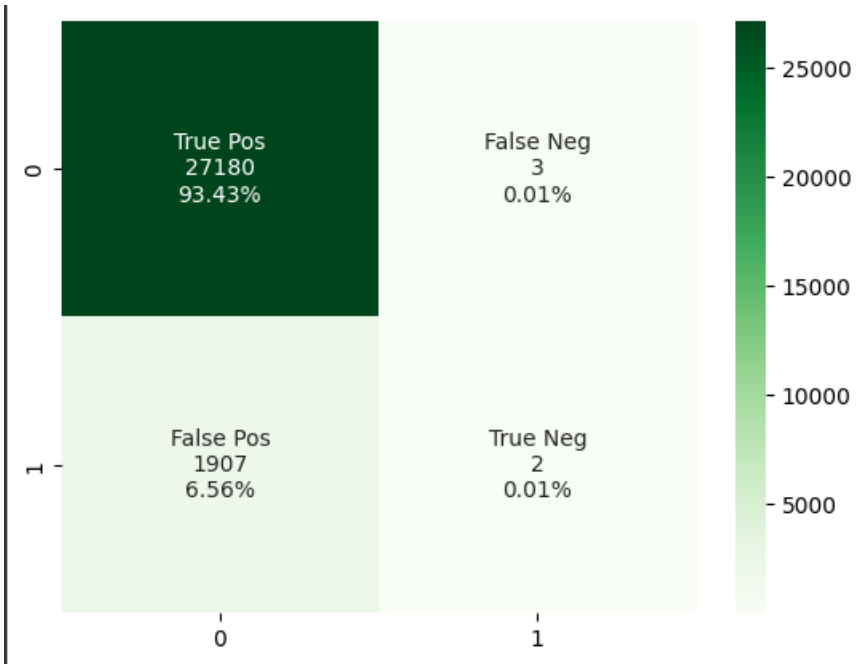


Fig. 5.11 Light GBM Confusion Matrix

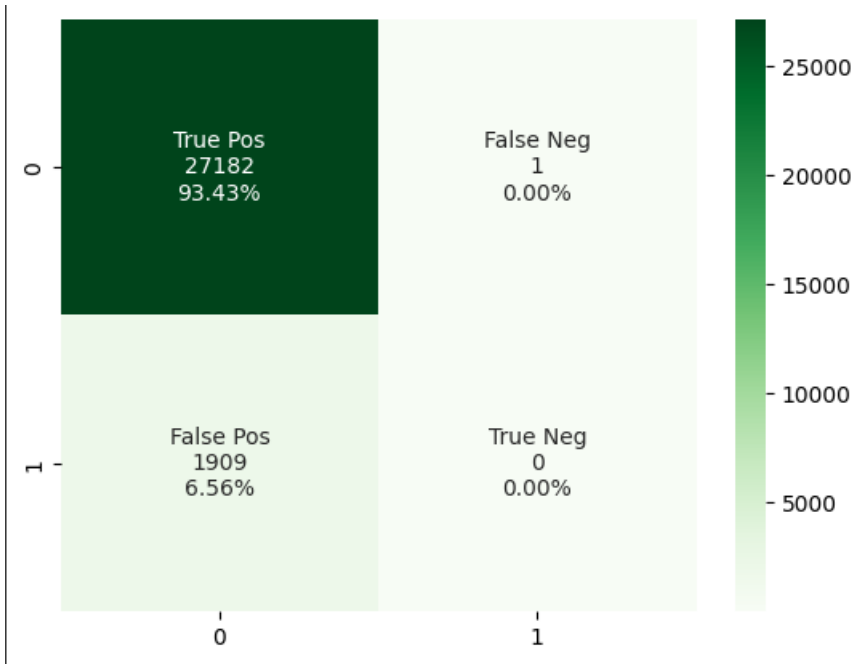


Fig. 5.12 XGBoost Confusion Matrix

Based on the metrics found proposed system the optimal algorithm,

Table 5.4 The Evaluation scores of each algorithm

classifier	F1-score	Precision	Sensitivity	Accuracy
Random Forest	0.60	0.75	0.49	0.934347
Decision Tree	0.52	0.51	0.52	0.877011
Logistic Regression	0.56	0.72	0.46	0.933693
XGBOOST	0.05	0.97	0.03	0.934346
Light GBM	0.61	0.75	0.52	0.934346

After finding the accuracy of algorithms, the better accuracy is found based on the accuracy score. The LGBM and Random Forest Algorithms show better accuracy and used to predict Rainfall for the next day.

The Fig. 5.13 is the barplot comparing the accuracy of prediction of rainfall using different algorithms. This helps us to find out which is the most accurate algorithm.

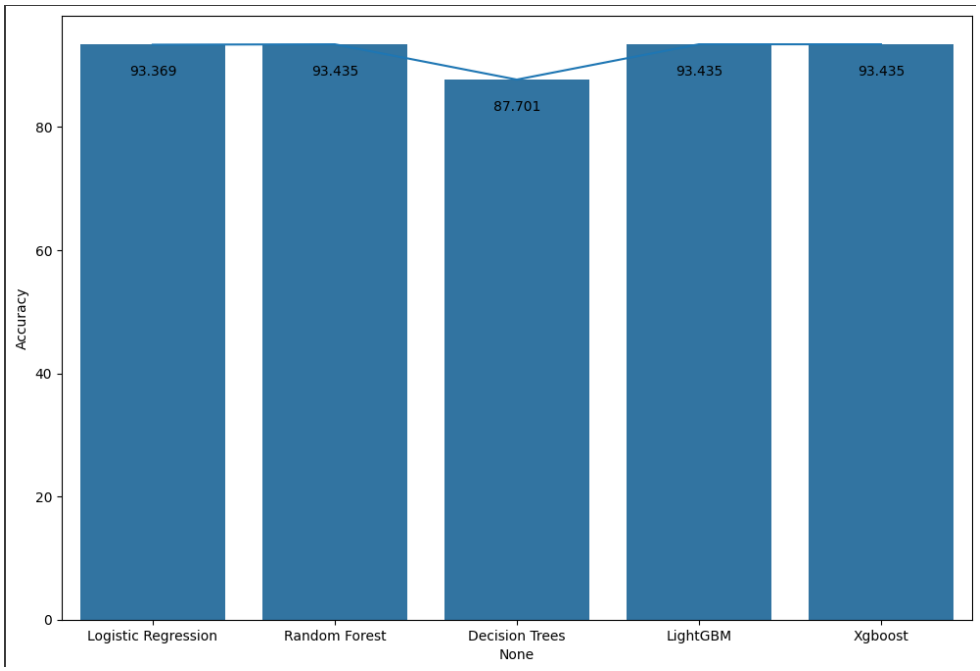


Fig. 5.13 Barplot representing the accuracy scores of algorithms

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 Conclusion

This Proposed system, we explored and applied several pre-processing steps and learned there impact on the overall performance of our classifiers. We also carried a comparative study of all the classifiers with different input data and observed how the input data can affect the model predictions.

We can conclude that Australian weather is uncertain and there is no such correlation among rainfall and the respective region and time. We figured certain patterns and relationships among data which helped in determining important features.

We tried Implementing the model with the most accurate Algorithm and thus succeeded and trained the model, now user can also find whether there will be a rain the next day in Australia by using Desktop Application ,This was done using Stream Lit.

6.2 Future Enhancement

As we have a huge amount of data, we can apply Deep Learning models such as -Multilayer Perceptron, Convolutional Neural Network, and others. It would be great to -perform a comparative study between the Machine learning classifiers and Deep learning model.

Predicting the rainfall of a specific geographic location would be a challenge. Improvising the prediction model to predict the weather conditions and even predicting the loses of rainfall. Coping with the changing parameter values and making the code compatible for the changes in the parameter values.

REFERENCES

[1]. M T Anwar, E Winarno ,W Hadikurniawati and M Novita(2021) Rainfall prediction using Extreme Gradient Boosting, Journal of Physics: Conference Series, Series no: 1869 012078

[2]. Jianzhuo Yan, Tianyu Xu, Yongchuan Yu, Hongxia Xu (2020) Long-Term Rainfall Forecast Model Based on The TabNet and LightGbm Algorithm,Europe PMC, PPR: PPR243083

[3]. Moulana Mohammed, Roshitha Kolapalli, Niharika Golla, Siva Sai Maturi(2020) Prediction Of Rainfall Using Machine Learning Techniques, VOLUME 9, ISSUE 01, ISSN 2277-8616

[4]. Gowtham Sethupathi.M , Yenugudhati Sai Ganesh , Mohammad Mansoor Alic(2021) Efficient Rainfall Prediction and Analysis using Machine Learning Technique,IJRC Vol.12 No.6 , 3467-3474

[5]. A.H.M. Rahmatullah Imon,ManosC Roy,S.K.Bhattaacharjee (2012) Prediction of Rainfall Using Logistic Regression, PJSOR, Vol. 8, No. 3, pages 655-667.

[6]. Verma A P and Chakraborty B S (2020)Performance Estimation of ARIMA Model for Orographic Rainfall Region ,URSI Regional Conference on Radio Science (URSIRCRS) pp 1–4

[7]. Pham B T, Le L M, Le T-T, Bui K-T T, Le V M, Ly H-B and Prakash I(2020) Development of advanced artificial intelligence models for daily rainfall prediction,Atmos. Res. 237 104845

[8]. https://www.kaggle.com/

[9]. Chen T, He T, Benesty M, Khotilovich V and Tang Y(2015)Xgboost: extreme gradient boosting R Packag. version 0.4-2 1–4

[10]. Geetha, A., and G. M. Nasira. (2014)Data mining for meteorological applications: Decision trees for modeling rainfall prediction. IEEE International Conference on Computational Intelligence and Computing Research. IEEE, 2014

[11]. Thirumalai, Chandrasegar, et al. (2017)Heuristic prediction of rainfall using machine learning techniques." International Conference on Trends in Electronics and Informatics (ICEI). IEEE, 2017.

[12]. Jain R and Nayyar A (2018) Predicting employee attrition using xgboost machine learning approach 2018 International Conference on System Modeling & Advancement in Research Trends (SMART) pp 113–20

[13]. Pan B(2018)Application of XGBoost algorithm in hourly PM2. 5 concentration prediction IOP Conference Series: Earth and Environmental Science vol 113 p 12127

[14]. Dhaliwal S S, Nahid A-A and Abbas R (2018)Effective intrusion detection system using XGBoost Information 9 149