UMBC

Analysis and Prediction of Rain fall

By
Sivaji Ganesh Konka

UMBC

Contents!

UMBC

01

Introduction & Abstract

Introduction:

Rainfall prediction is an important part of weather forecasting. It involves using multiple methods and strategies to forecast the amount of precipitation that will fall in a given location.

Also, Rainfall prediction accuracy has improved significantly over the years as a result of technological improvements and the development of sophisticated models and algorithms.

This information is essential for various sectors like agriculture, water management, disaster management, and more.

Abstract

Rainfall forecast is influenced by several factors, including atmospheric pressure, humidity, wind patterns, and temperature.

In general, data must be collected from a variety of sources, including meteorological stations, satellites, and radar systems, in order to predict rainfall. then, the data is analyzed using statistical models and machine learning algorithms to identify patterns and trends that can be used to forecast future rainfall events.

In this Project, we are trying to develop a predictive model on the Rainfall Dataset in order to forecast whether or not it will rain in Australia tomorrow. The collection of data contains around ten years of daily weather measurements from numerous places throughout Australia.

UMBC

02

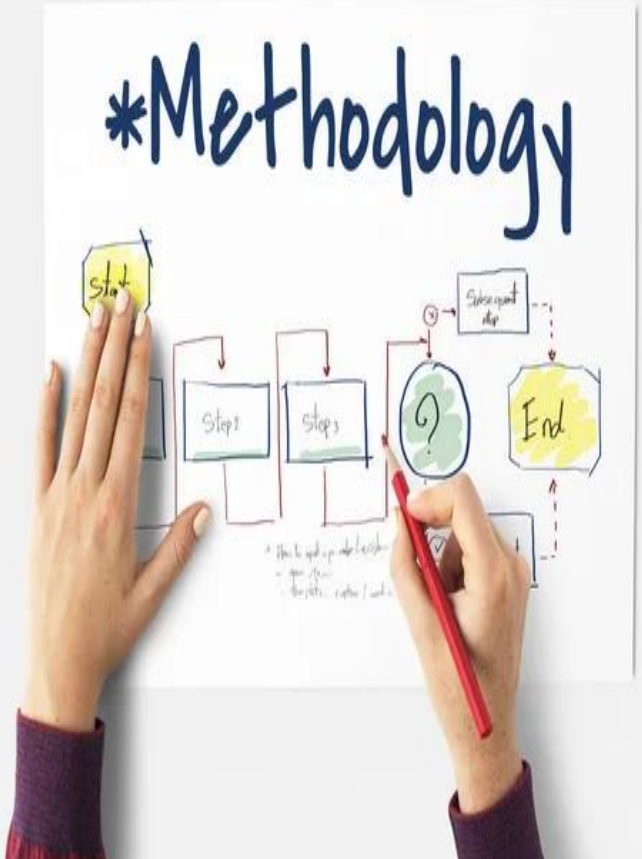
Data Source & Methodology

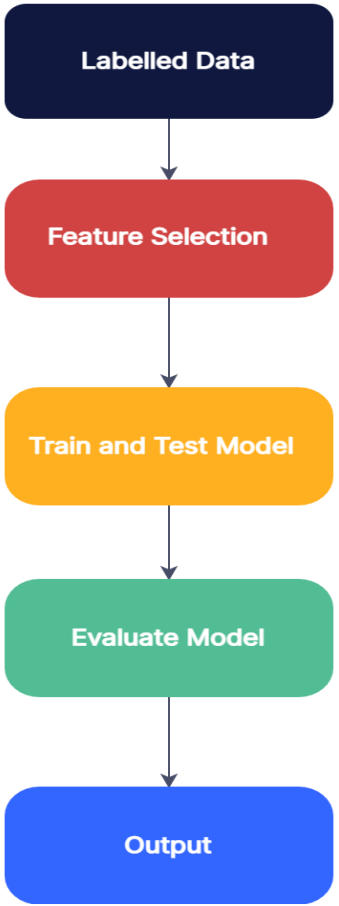
Data Source:

- Here the data is collected from **Australian Government Bureau of Metrology.**
- Data source: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.
- Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml
- Data consist of 145460 rows and 23 columns
- Copyright Commonwealth of Australia 2010, Bureau of Meteorology

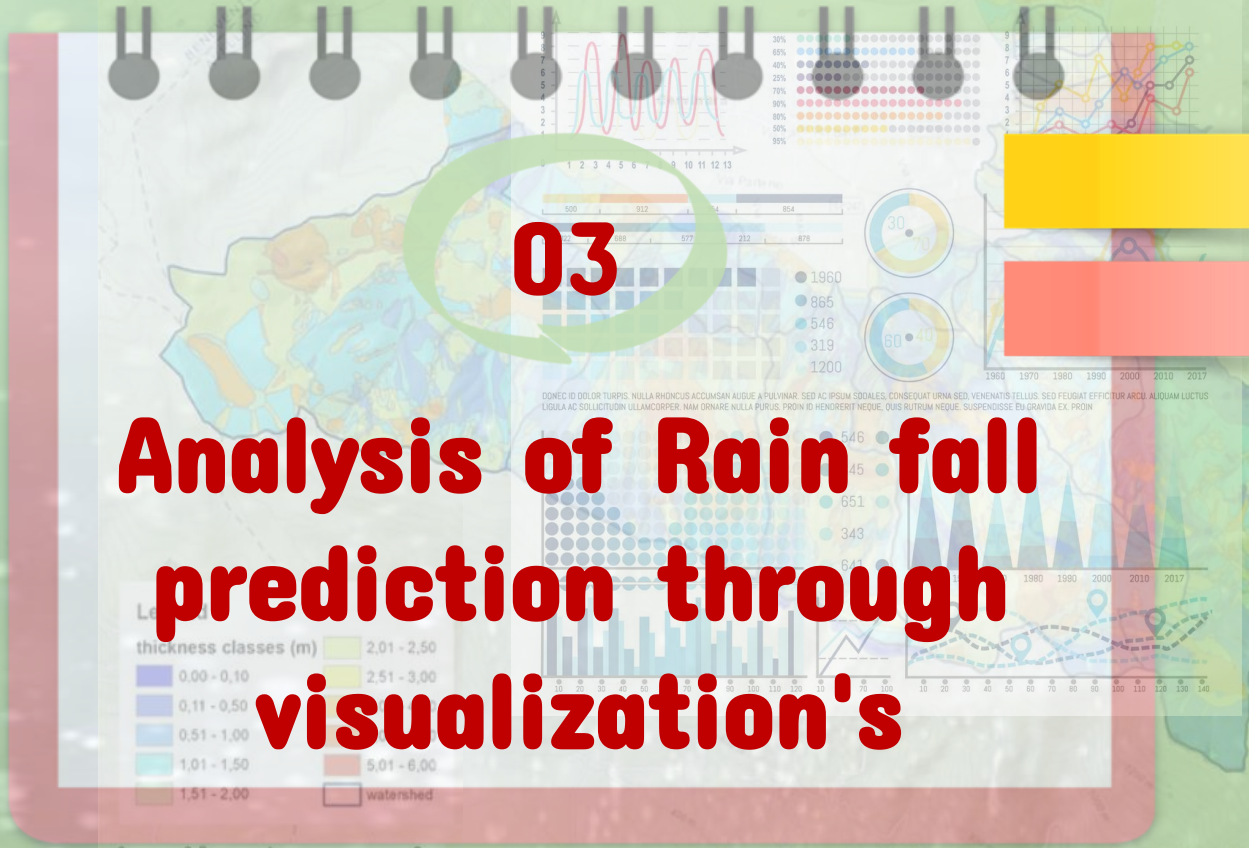
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	71.0	22.0
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	44.0	25.0
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	38.0	30.0

Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
1007.7	1007.1	8.0	NaN	16.9	21.8	No	No
1010.6	1007.8	NaN	NaN	17.2	24.3	No	No
1007.6	1008.7	NaN	2.0	21.0	23.2	No	No


*Methodology
Start
Step2
Step3
End


Labelled Data
Feature Selection
Train and Test Model
Evaluate Model
Output

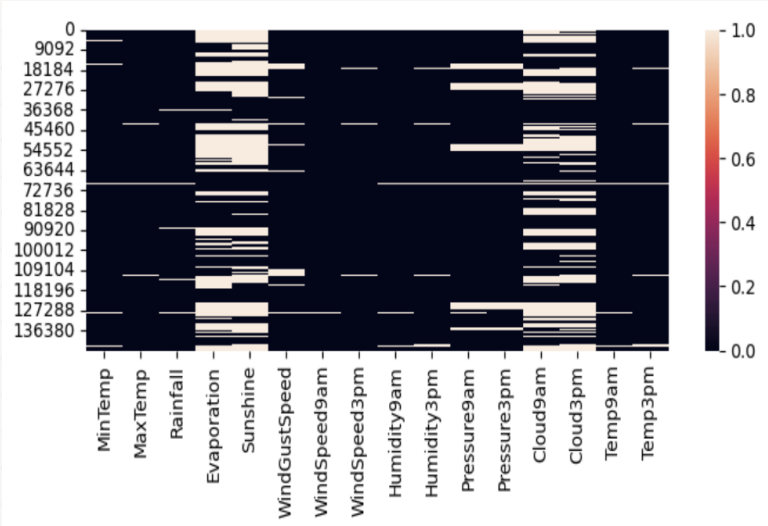
UMBC



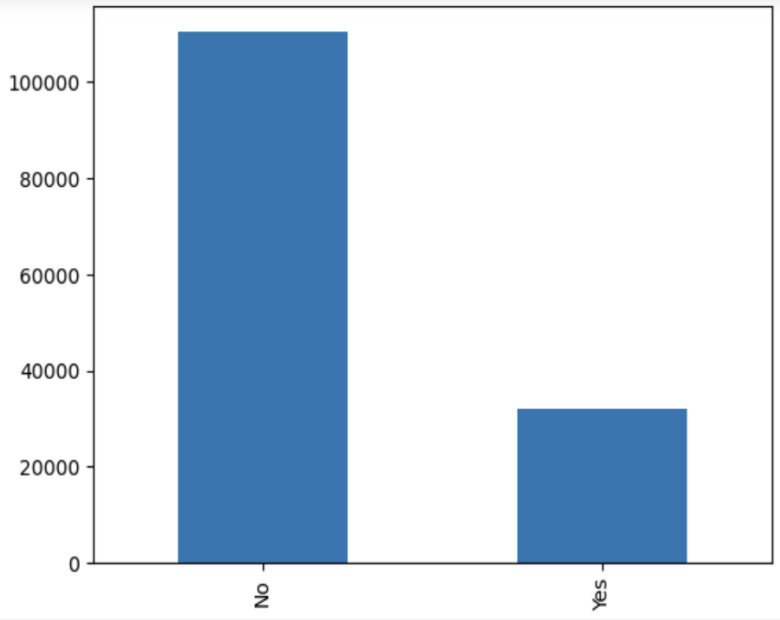
03

Analysis of Rain fall prediction through visualization's

- ✓ Plotting Distribution of null values in the data set using pictorial representation.
- ✓ The columns: Evaporation, Sunshine,Cloud9am,cloud 3pm contribute highest number of null values in this data set.

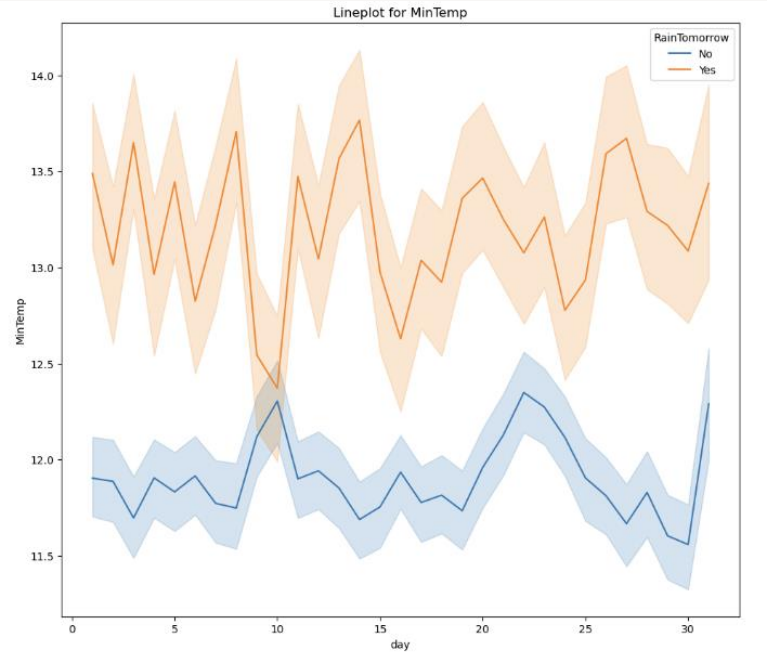


- ✓ Plotting the instances of the target variable, that is rain tomorrow.
- ✓ As we can see we have more number of “Nos” in the given data

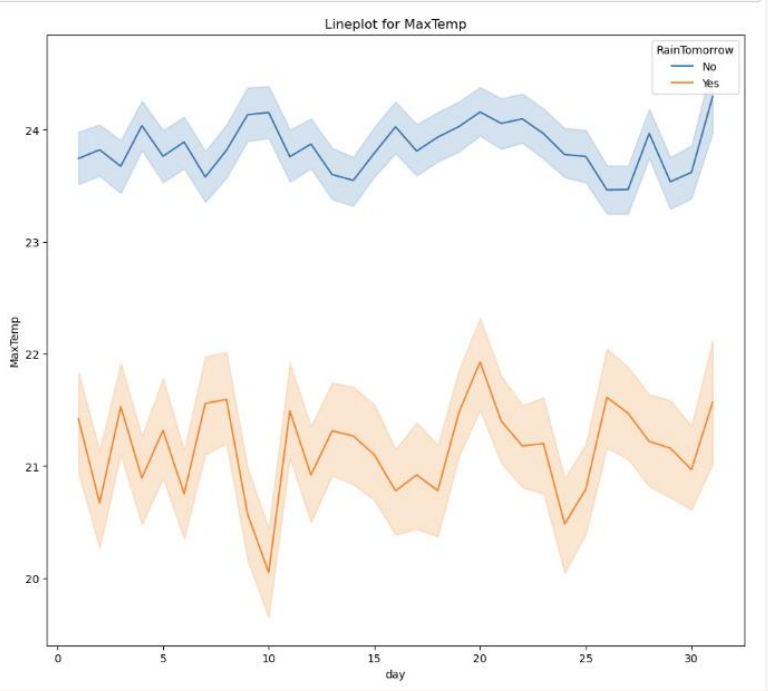



```
No      110316
Yes      31877
Name: RainTomorrow, dtype: int64
```

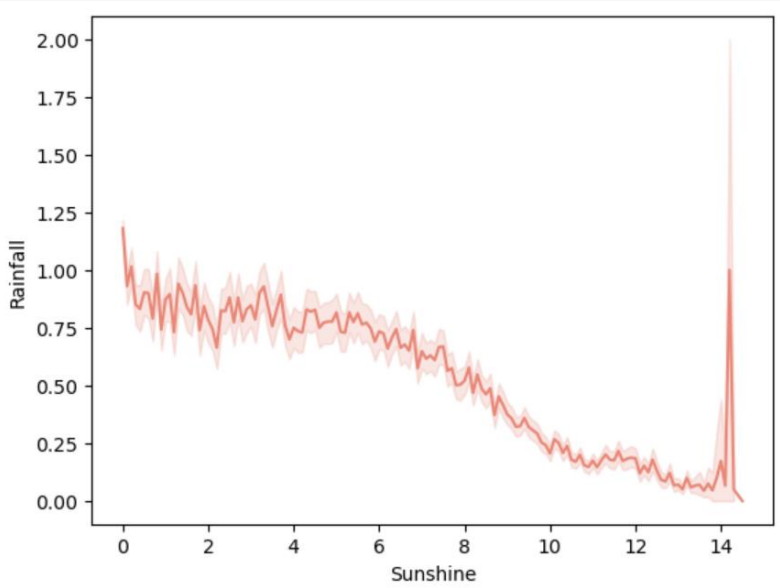
- ✓ Here we can see the correlation between the target variable and minimum temperature
- ✓ The graph depicts that when there is minimum temperature, we can expect the rain fall as per the data.



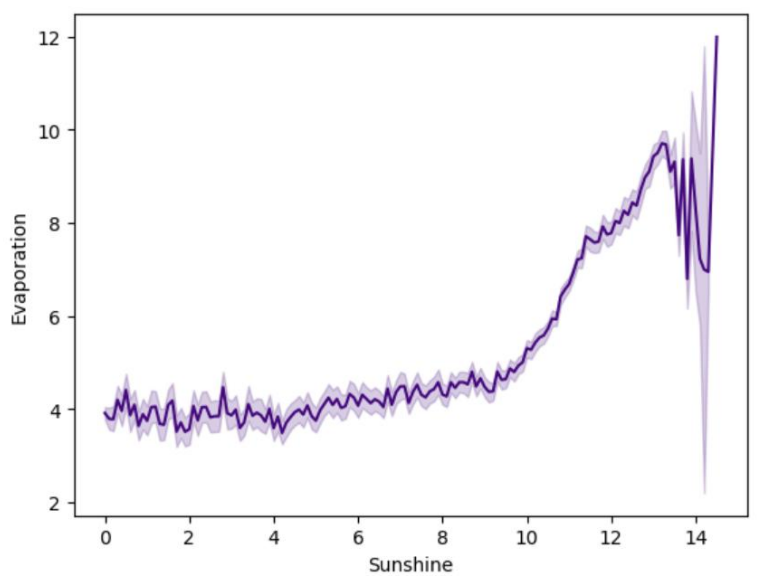
- ✓ Here we can see the correlation between the target variable and maximum temperature
- ✓ The graph depicts that when there is maximum temperature, we cannot expect the rain fall as per the data.

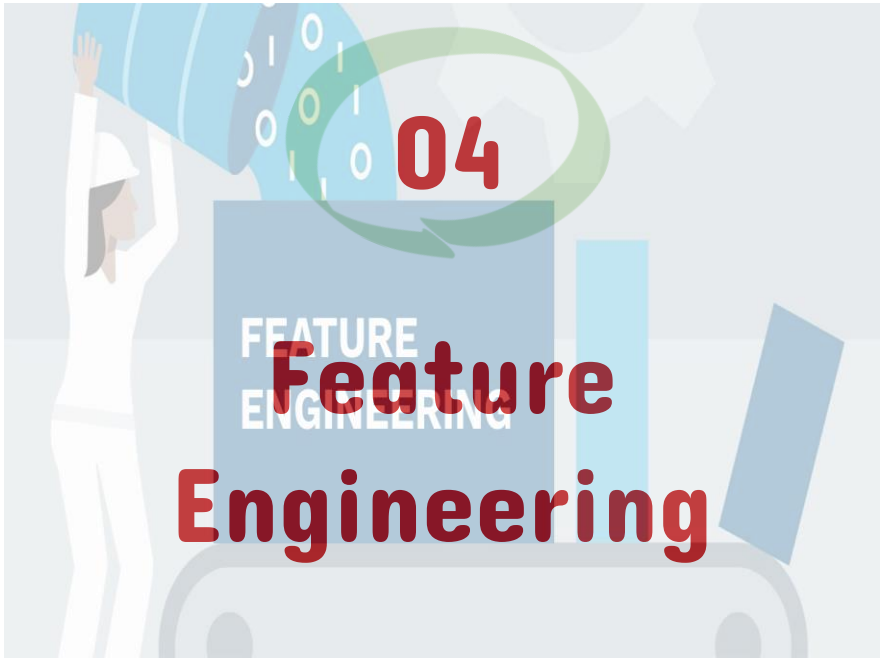


- ✓ Plotting Here we can see the correlation between rainfall and sunshine.
- ✓ The graph depits that when there is rainfall the sunshine gradually decreases.



- ✓ we can see the correlation between Evaporation and sunshine.
- ✓ The graph depicts that when there is gradual increase in sunshine the amount of evaporation also increases.





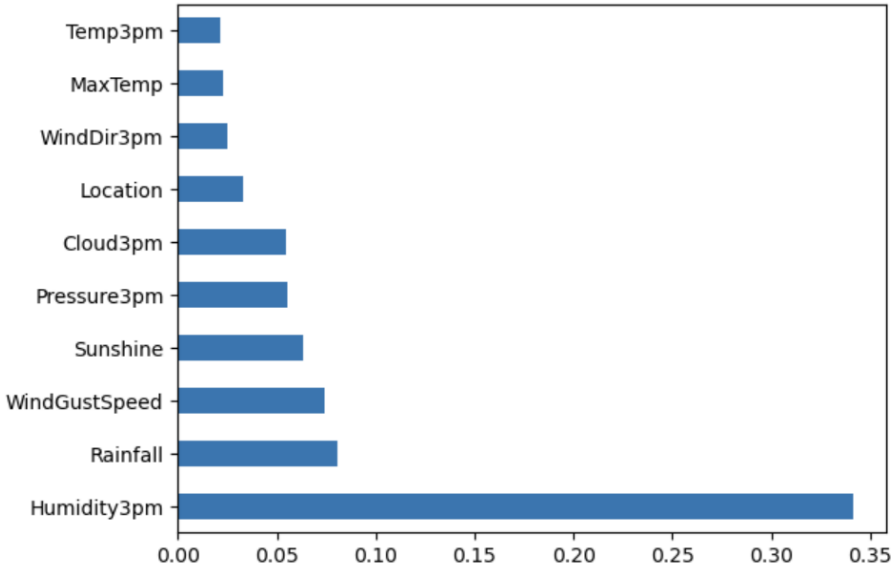
04

Feature Engineering

Feature Selection in Machine learning:

- We have approached this method of feature selection using XGB Classifier

- Here are the important features that have more impact on the target variable.



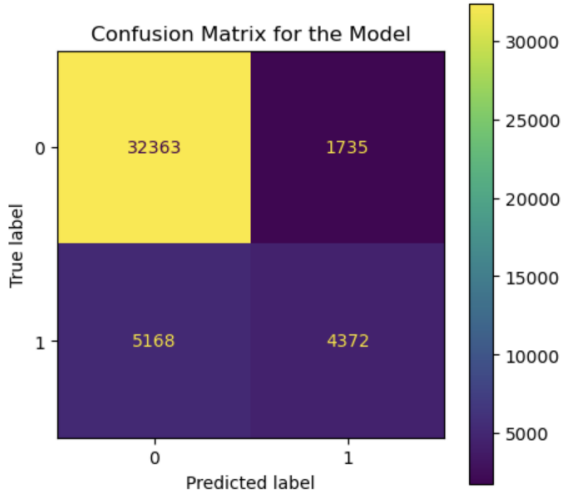
05

Model Prediction Algorithms

Logistic Regression

- ✓ Logistic regression is a machine learning approach that is used for binary classification. It is not, as the name implies, a set of rules for regression problems with non-stop results. Instead, logistic regression is the preferred method for binary classification. This gives a discrete binary value between 0 and 1. Logistic regression estimates probabilities using its core logistic feature to assess the relationship between dependent variables and one or more independent variables. These probabilities are then transformed into binary numbers for an estimation.
- ✓ Its range is fixed to be between zero and one. Additionally, R performs well when there is a large gap between the groups. here we have gained accuracy for LR is 84%.

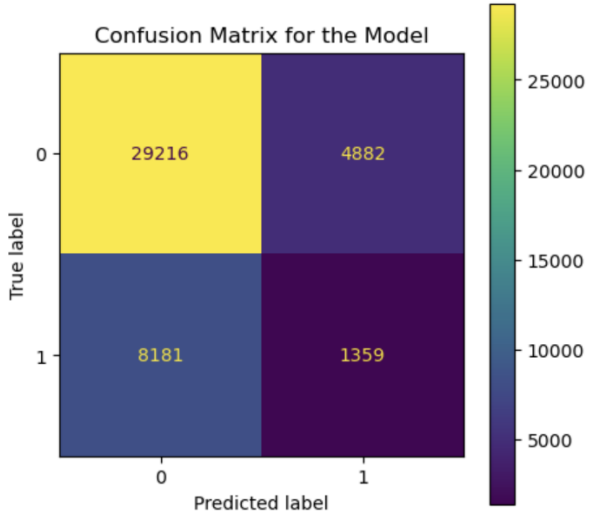
	precision	recall	f1-score	support
0	0.86	0.95	0.90	34098
1	0.72	0.46	0.56	9540
accuracy			0.84	43638
macro avg	0.79	0.70	0.73	43638
weighted avg	0.83	0.84	0.83	43638



Random Forest

- ✓ Random forest is an algorithm used in behavior analysis that is built on modeling assumptions and decision trees. It consists of several decision trees representing a different example of the classification of data input into a random forest.
- ✓ Here we achieved Accuracy of 86% for Random Forest.

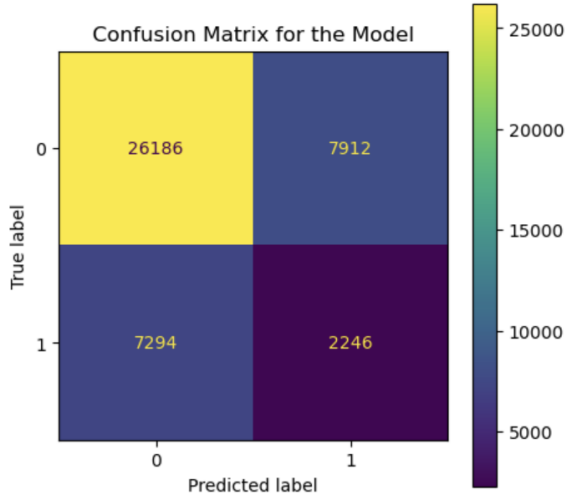
	precision	recall	f1-score	support
0	0.87	0.96	0.91	34075
1	0.77	0.50	0.61	9563
accuracy			0.86	43638
macro avg	0.82	0.73	0.76	43638
weighted avg	0.85	0.86	0.85	43638



Decision Tree

- ✓ Decision trees are models that begin with root split into branches at nodes representing the predictor variables using the rules.
- ✓ Each node Is then divided into multiple nodes. The goal is to use the decision tree to divide the data space into dense and thin regions.
- ✓ As the number of splits increases the time required to build the tree also increases. And here we achieved Accuracy of 78% for Decision tree .

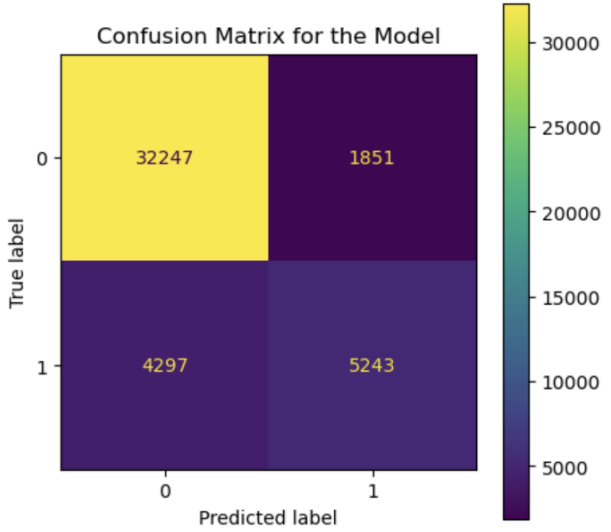
	precision	recall	f1-score	support
0	0.87	0.85	0.86	34075
1	0.51	0.54	0.52	9563
accuracy			0.78	43638
macro avg	0.69	0.70	0.69	43638
weighted avg	0.79	0.78	0.79	43638



XGB Classifier

- ✓ XGBoost stands for extreme Gradient Boosting and it's an open-source implementation of the gradient boosted trees algorithm. It has been one of the most popular machine learning techniques in Kaggle competitions, due to its prediction power and ease of use. It is a supervised learning algorithm that can be used for regression or classification tasks.
- ✓ we achieved Accuracy of 78% for GB Classifier

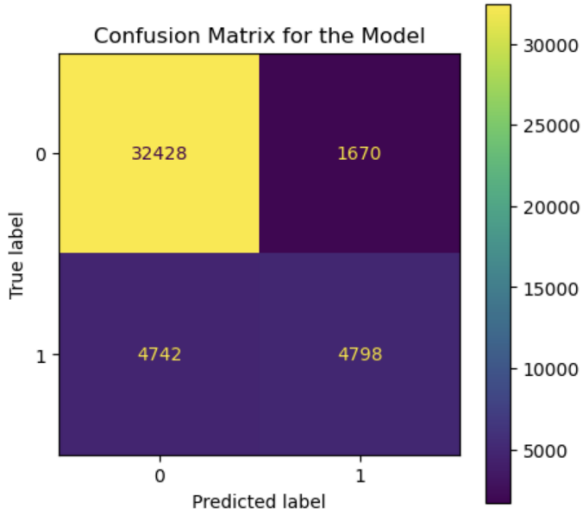
	precision	recall	f1-score	support
0	0.88	0.95	0.91	34098
1	0.74	0.55	0.63	9540
accuracy			0.86	43638
macro avg	0.81	0.75	0.77	43638
weighted avg	0.85	0.86	0.85	43638



GB Classifier

- ✓ Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.
- ✓ When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.
- ✓ A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.
- ✓ we achieved Accuracy of 78% for GB Classifier

	precision	recall	f1-score	support
0	0.87	0.95	0.91	34098
1	0.74	0.50	0.60	9540
accuracy			0.85	43638
macro avg	0.81	0.73	0.75	43638
weighted avg	0.84	0.85	0.84	43638



06

Conclusion

Conclusion:

Based upon the prediction results, we conclude that our prediction is correct, and the factors which we have considered have significant impact in predicting the Rainfall.

07

Future Scope

Future Scope

The proposed model can be inferred to predict the rainfall. If relevant data is available, the model can also be trained to take additional parameters. The model, with minor modifications, can be integrated into simulation software to better predict the rainfall and give some emergency warning about the weather. So that people can take necessary precautions.

08

References

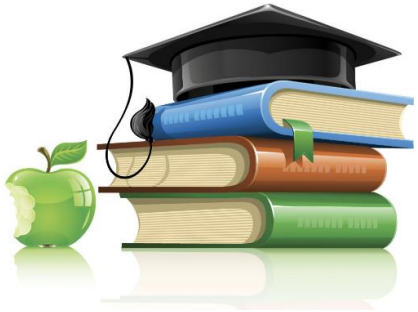
References:

[1]. V. Rao and J. Sachdev, "A machine learning approach to classify news articles based on location", 2017 International Conference on Intelligent Sustainable Systems (ICISS), pp. 863-867, 2017.

[2]. V. Kumar and S. Minz, "Poem classification using machine learning approach", Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), pp. 675-682, December 28–30, 2012.

[3]. B.Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques" in Language Processing (EMNLP), Philadelphia, pp. 79-86, July 2002.

[4]. B. Stehman, "Selecting and interpreting measures of thematic classification accuracy", Remote Sensing of Environment, 1997.




thank you