Rainfall prediction using machine learning

Kurra Lavanya
700728999
dept.Computer Science
University of Central Missouri
lxk89990@ucmo.edu

Ramya Edpuganti
700733458
dept.Computer Science
University of Central Missouri
Rxe34580@ucmo.edu

Srikavya Gottumukkala
700742824
dept.Computer Science
University of Central Missouri
Sxg28240@ucmo.edu

Jitheandra Subramaniam
700742721
dept.Computer Science
University of Central Missouri
jxs27210@ucmo.edu

Abstract—Due to large amounts of greenhouse gasses being released into the air there is an increasing trend of Global warming.One of the major causes of nonseasonal rains is Global warming.With the nonseasonal rains main industries that will be affected are agricultural.Along with the agricultural industry the economy and life of the people also will be at the stake.Predicting the rainfall at a particular location is depends on the atmospheric parameters specific to that area.In general the meteorology department predicts the rainfall using different methods. These methods are not accessible to other people. While predicting they consider atmospheric parameters into consideration.Based on the past and present data they observe the patterns and make an educated guess to predict the rain.But there is quantitative method to measure the probability of the guess.Machine learning offers forecasting models with the given data.The advantage of using machine learning models gives us quantitative analysis.We can predict the accuracy of the classes and different visualizations to filter the outliers in the data and data cleaning.In this paper we classify or forecast the rainfall using meteorological parameters using machine learning techniques.Our project is a binary classification problem.The features of the project include ensemble models and SMOTE oversampling methods. SMOTE balances the imbalance classes for better prediction results.Machine learning Decision tree, Gradient Boosting, Ada Boosting and K-Means clustering algorithms are used to predict next-day rain by training classification models on the target variable Rain Tomorrow. In this paper we have used weather AUS dataset which is collected from kaggle repository.

Index Terms—Ada Boosting, ANOVA(Analysis of Variance), atmospheric parameters, Binary classification, Gradient Boosting, SMOTE(Synthetic Minority Over-sampling Technique), feature selection,

I. Introduction

Many countries which are majorly dependent on agriculture and poor road systems this project acts as an alert to avoid the loss of crops and human lives.According to the study due to nonseasonal rains created havoc in many places leaving the countries in huge economical loss. Heavy rainfalls create havoc in many places costing human lives and crops damage are main affected sectors. People will be alert in the season of monsoon but nonseasonal rains will damage the crops of summer-sown for example rice, soybean, cotton, pulses and vegetables are the staple foods in many areas. If the rainfall is heavy before the harvesting it causes huge damage to the crops. With the damage of crops will lead to food inflation that in turn increases the prices of the food.Uttar Pradesh is highest rice producing state in India. In the year of October 2022 Uttar Pradesh state witnessed a rainfall 5 times more than regular rainfall. It damaged the rice crop. Ina positive perspective predicting the rainfall in particular locations helps the farmers with the decision making like planning the crops and how to make the yield efficient. Knowing the rainfall details on or before makes the process easy for a farmer. This positive perspective is applied to the other fields as well not only the agricultural sector. For example knowing the rainfall rate on or before saves the lives of the people because every time around 10 people loose life in rain related accidents in the country. The other reason for knowing the rainfall or predicting the rainfall helps the government plan the resources before the disasters because most of the areas in India are flood prone areas.This damage is recurrent. Out of 329 million hectares of area flood prone area is 40 that is 12 percent of the area. The estimated damage between the years 1996 to 2005 was Rs. 4745 crores and the average for the last 53 years is Rs. 1805 crore. The main reasons for this are global warming and increased urbanisation. Statistics show that average of with 75 lakh hectares of land is affected by floods every year and the cost of lives are 1600, damaged crops and houses and public utilities.The frequency of the floods is 5 times in every year. In the year of 1977 people lost their lives maximum of 11,316 which is the highest number. Rains cause damage to both public and private properties. With the data from meteorological department we have found that floods affected the areas which are less prone to floods. Predicting the rains in advance makes the flood management tasks effective. With the inability to carry the sludge makes the rivers to cause floods and most of the floods caused by the rivers which are Indian originated and flowing into other countries it adds to the other dimension of the problem and making it complex

Working with the forecasting data like rainfall prediction comes with many challenges. Data is diverse and the finding the patterns is very difficult. In this paper data used is weather AUS dataset. Dataset contains meteorologic parameters like wind speed, wind direction. To find the patterns in the data, visualization tools plotly is used which gives the aesthetic

Github link : https://github.com/Jitheandra/Finalproject

appearance of the visualizations compared to seaborn and matplotlib.

Predicting the weather related factors force us to look at the data for patterns because predictions depend on various parameters. For example rainfall depends on humidity, pressure and so many factors in general. But we can not train the model with all the parameter as it increases the training time in reality. So to solve this problem machine learning offers a technique called feature selection. There are different types of feature selection methods depending on the input-output and supervised and unsupervised learning methods.Under supervised learning methods we have 3 methods intrinsic, wrapper and filter methods. In the filter method features are eliminated using correlation matrix i.e. input that is related with the output. In the wrapper method we take the subset of the features train the model based on the output and performance of the model we add or subtract the features this process is continued until the best accuracy is achieved. In in the intrinsic method we combine the both the methods filter and wrapper method. Feature selection methods also categorized based on input output types. If the input is numerical and the output is also numerical then we can use Pearson's correlation coefficient and Spearman's rank coefficient. If the input is numerical and the output is categorical we use ANOVA feature selection and Kendall's rank coefficient. If the input is categorical and the output is numerical we use ANOVA correlation coefficient non linear and Kendall's rank coefficient non linear. If both input and output are categorical we use Chi-squared test and mutual information.

In general all the data is not balanced. But dealing with the imbalanced data poses few problems to the accuracy and other parameters. If the data is imbalanced we should not use accuracy parameter as performance measure because it will not give the whole picture of the performance of the models in general. We can deal the imbalanced data in two ways one is exploring the various performance metrics and the other technique is over sampling or under sampling the data. In the first approach we can consider the precision, recall and F1score parameters instead of the default parameter accuracy. In the second approach we can use over sampling which increases the samples of non dominant classes or under sampling which reduces the dominant samples. These two methods have their own pros and cons.

In our project we have considered the over sampling method i.e. SMOTE. Over sampling works in opposite to under sampling method. It creates the duplicates of non dominant classes and adds to the data. The disadvantages of over sampling is it creates the duplicates without any threshold. That means it creates too many duplicates and we can not the duplicate values. Duplicated values are not the real values so we get some redundancy in the original data.

SMOTE analysis provides the solution to the drawbacks of the oversampling which provides the duplicates and unnecessary data. But in the SMOTE analysis avoids this creating of duplicates instead it creates the synthetic samples which are slightly different from original samples other than duplicates.

- Working of SMOTE analysis:
- In the step 1 we take a sample from the non dominant class
- In the step 2 from that random sample we will draw observations and identify the K nearest neighbors
- From the neighbors we will calculate the vector which is drawn from the current data point and the neighbor
- After finding the vector we will multiply the vector by the value between 0-1
- after multiplying the vector with random number our synthetic sample will be ready
- in the final step the synthetic sample will be added to the original samples

With this procedure we add the synthetic points from the neighboring points that is no too far from the original point and not the exact duplicate. The advantage is we can create the synthetic samples without any redundancy. The synthetic point is calculated by slightly moving from the original point towards the neighbor. This method affects the performance metrics precision, recall and F1 score as well. To create a robust model we should have the right performance metric. No matter how sophisticated methods you use choosing the wrong performance metric will not yield anything. So choosing the right performance metric plays an important role. Understanding the accuracy parameter gives the results of correct predictions in total samples i.e. in machine learning terms it is the sum of true positives and true negative divided by the total samples. But when we implement the models in real world scenarios focuses on false positives and false negatives.

An intuitive example of coronal virus case gives the better understanding of false positives and false negatives. If the patient is actually having corona virus but the predictive model predicted it as corona negative this comes under false negative and if the person has corona negative but the predictive model predicts it negative then that is false positive. In machine learning to know these factors we have two methods one is classification report which displays the precision, recall and F1-score for each test point on the scale of 0 to 100 for the each a. And the other method is confusion matrix this gives the over all breakdown of incorrect classifications and correct classifications with the model.These two methods can be implemented using scikit-learn library.

For any business case selecting and exploring false positives and false negatives plays a pivotal role to scale the project and weigh the contribution of the project towards the business

Github link : https://github.com/Jitheandra/Finalproject

goal.Precision tells the what portion of the positive predictions are correctly classified that is true positives divided by sum of true positives and false positives.Recall tells the story of true positives divided by sum of true positives and false negatives.Where as F1 score gives the advantages of both the methods precision and recall.That is average of precision and recall factors.

Researchers conclude that climate is a complex model in the analytical world this is called as hard to solve problem. However, with the development of computational methods makes a way to experiment with numerical algorithms to challenge these scenarios.Machine learning is algorithms are part them. Machine learning methods are easy to adopt unlike Deep learning methods which require specific computational power.

II. Motivation

Researchers conclude that climate is a complex model in the analytical world this is called as hard to solve problem. In general the accuracy of the predictions depend on the accuracy or variance of the data. In the traditional method predictions we can not preprocess or analyse the data but the advantage of machine learning methods gives us the facility of preprocessing or cleaning the that is ready to feed the data. However, with the development of computational methods makes a way to experiment with numerical algorithms to challenge these scenarios.Machine learning is algorithms are part them. Machine learning methods are easy to adopt unlike Deep learning methods which require specific computational power.The main goal of the machine learning is to find the patterns and draw observations. With this observations and patterns we train the model.While training machine learning models adjust the weights of the model when we apply new or unseen samples or test samples it will predict the results. In the training process we give input as samples and outputs. In the test phase we give input and ask the targets as predictions. These predicted outputs are compared using true labels to find the accuracy.

III. Objectives Main objectives

of the project are:

- Our project is a binary classification problem.
- The objective of our project is to classify or forecast the rainfall using meteorological parameters using machine learning techniques. It will classify whether raintomorrow is yes or no.
- The features of the project include ensemble models and SMOTE oversampling methods. SMOTE balances the imbalance classes for better prediction results
- Machine learning Decision tree, Gradient Boosting, AdaBoosting and K-Means clustering algorithms are used to predict next-day rain by training classification models on the target variable Rain Tomorrow.
- After the implementation of the algorithm we conduct a comparative analysis of the models.

IV. Related work

Rainfall prediction is the most important prediction category in machine learning because the damage caused by this one is not only limited to the agricultural industry it causes the loss of human lives. Sometimes the this is very high in few areas where the affect of the floods is high. This project is a research about Tamil Nadu region, India rainfall precipitation. Since this region is a coastal area we have considered this state to predict the rainfall. In Tamil Nadu, India we have analysed the rainfall precipitation in southeast and southwest direction. Rainfall precipitation for south east direction ranges from 431470 mm where as southwest direction it ranges from 171-828 mm. periodical rainfall is analysed from 2004 to 2014 using Holt Winter method and the performance metric Root Mean Squared Error(RMSE) is used and the result of 96.0 percent is achieved [9].

The second most state that is affected by the floods and rainfall is Kerala, India. In this paper we have observed the trends of summer monsoon, post-monsoon and pre-monsoon of this region. To explore the data we have implemented SFLN (Single Layer Feed Forward Neural Network) and Extreme Learning Machine algorithms are implemented. As Kerala is a coastal state from the data we have observed that summer monsoon hits the state heavily. After the implementation we have conducted the comparative analysis. From the comparative analysis we observed that ELM model out performed the SFLN model [5].

Predicting the climate changes is a difficult task in that rainfall prediction and the major challenge is finding the best features that are suitable for prediction. In this study we will consider two most important features that effect the rainfall most. In general rainfall depends on two important features those are humidity and temperature. With these two parameters we can predict the rainfall. The entire project is built in MATLAB environment and two different approaches are used one is subtractive clustering and Levenberg-Marquardt algorithm.In the end their performance is compared [16].

Rainfall prediction depends on several factors. The type of factors vary with the technology. According to few studies the factors are pressure, humidity and temperature. But these factors comes under modern methods. The advantages of modern methods are they can predict the results accurately. Other than modern methods we have traditional methods also. In traditional methods we have factors which affect the rainfall one of the factors is phases of moon.Though this is an unusual phenomena according to traditional methods this has

Github link : https://github.com/Jitheandra/Finalproject

huge impact on the rainfall prediction. To predict the phases of the moon we have deployed two machine learning model Logistic regression and Random Forest. Machine learning is used because of its efficiency [18].

Another impact factor for rainfall is landslides. This situation causes the property loss and loss of lives. In this paper we collect the rainfall data and perform average rainfall and cumulative rainfall and observe the historical landslide data. How the rainfall data affects the landslides. To perform this task AdaBoost classifier and AdaBoost BP neural network algorithms are used [17].

Machine learning performs best when the data is sufficient. To perform better we need sufficient data. Though it performs well on small datasets. Few prediction tasks need much data to process and understand the data. This much data can not be visualized using manual methods. With the advent of data visualization methods we can easily analyse the huge datasets [8].

Availability of the data and reliable data is important for machine learning tasks. Most of the data is collected from IMD(Indian Meteorology Data) which is a government resource. In this project SVM-RF support Vector Machine and Random Forest classifier is used to predict the results. This ensemble model outperforms the existing models. The performance metrics considered are MAE(Mean Absolute Error) and RMSE(Root Mean Squared Error) [19].

Rainfall prediction can be done in different types. In this project rainfall prediction of next year is calculated using the historical data. Data mining techniques like K-NN(K Nearest Neighbors) and Decision Tree perform well on the prediction datasets. These algorithms find the patterns from the data. These two approaches are applied to Austin data. [4]

Data collected from ITU(International Telecommunication Union) data is processed and from the data annual rainfall rate and inter-annual rainfall rate is analysed. In addition to annual rainfall rate monthly rainfall rate also analysed. The data contains 16 years of the rainfall data. After analysing experimental results and actual results are compared [6].

The stages of machine learning is finding the patterns, observing the data and training and learning the weights and adjusting them. With the invention of neural networks which are the adoption of human brain functionlity. Neural Networks are the sought after methods to solve the complex problems. In neural networks there are different types methods for example ANN(Artificial Neural Network) and FFNN(Feed Forward Neural Netwok) these methods used feed forward structure to reduce the error [13].

Available data on rainfall contains location specific sometimes it is limited to one location. If the data is limited to only one location we can not make much out of the data. With the diverse data we can generalize the model. In this study we have China Meteorology Adminstration(CMA) open dataset is used. The data contains the information of different locations. In the dataset latitude , longitude, altitude , average temperature and prior prediction probability considered to make predictions [14].

In a hybrid model in this we have implemented rainfall prediction model along with crop yield prediction model. To predict the crop yield we have novel approaches like ARMA(Auto Regressive Moving Average), SARIMA(Seasonal Auto Regressive Integrated Moving Average), ARMAX and fuzzy logic models.In the end we have performed comparative analysis of the models [2].

India is majorly dependent on agriculture. New innovations like knowing the rainfall in advance gives the opportunity to plan the things early.By knowing the rainfall in advance helps the farmers in decision making in choosing the efficient crop yields. Machine learning models like Artificial Neural Networks(ANN), Back propagation Neural Network and Cascaded Neural Networks various types of Neural Networks are implemented and compared the results [3].

Rainfall or weather data is considered to be the volatile data. The data changes overtime and there are random changes in the data. To predict the fluctuated data we need standard machine learning models like Markov chains but the Markov chains alone can not give the alone can not predict the results accurately. So combining the Markov chains with Gray Markov model gives efficient results [11].

With the complexity of the data we have implemented the combination of embedded methods and various performance metrics. In the first stage we have implemented XGBoost algorithm and GRU on the time series data. To measure the data we have considered the various performance metrics MAE(mean absolute error), MSE(mean squared error) and MAPE(mean absolute percentage error). In the two models XGBoost has the advantages of processing data compared to GRU network. This can be applied to the other projects as well [1].

Like embedded methods multi-mode models also proven to be best approaches to solve the time series and complex data like rainfall prediction or atmospheric datasets.In this paper authors researched about the SVR - RBF multi model algorithm. The combination Support Vector Regressor and Radial Bias Function are used to find the rainfall [7].

Time series data is handled by different kinds of approaches in different ways. These time series data is different in different categories. For example we can consider time series data in hydraulic system also but the people find it difficult to explore this data due to its complexity. In this case Neural Networks come handy GRN(Generalized Regression Neural Network) neural networks handle the scenarios efficiently. Compared all other neural network structures GRN((Generalized Regression Neural Network)) has the simplest structure and error rate is also proven to be very low

Github link : https://github.com/Jitheandra/Finalproject

[20]. Rainfall prediction has various applications. Machine learning models like GAN(Generative Adversarial Networks) can predict the rainfall ground radar signals.GANs are the next generation machine learning algorithms in this project researchers used Conditional Generative Adversarial Networks for predicting the short term rainfall from 10min to 4hours [10].The primary source of India is agriculture. Though we have enough fertile soil in the country there is an increasing trend of farmers suicide hints the lack of prior knowledge of rainfall prediction. A proper and robust model predicts the rainfall in advance helps the farmers plan the crops and other things in an efficient manner. In the recent years Chennai, Tamil Nadu India hit by major floods which in turn damaged the huge crops [15]. In a new approach we have implemented the new neural network structures BPNN(Backward propagation neural network) and other sate of the art neural networks to check the working of neural networks on the rainfall data. After the implementation of these algorithm a comparative study is conducted to know the working of each algorithm [12].

V. Data Description

Dataset of the project are collected from open source repository kaggle.This dataset is weatherAUS. It contains various climate and weather parameters.Target variable has two types of categories whether it will rain or not the next daya i.e. yes or no. If the target column is yes the rainfall is more than 1mm. Dataset size is 14MB. The characterstics of the dataset are parameters contain mix of both numerical and categorical parameters. The characterstics of the data are:

- There are 145460 instances in the data
- There are 7 object columns
- Remaining 15 columns are numerical columns
- There are missing values in the data.Top 5 missing value columns are:sunshine, evaporation, Cloud 3pm, Cloud 9am, Pressure3pm, Pressure 9am.

S.No.	Column name	Description
1.	Date	Date of the record
2.	Location	Location of recorded samples
3.	MinTemp	Minimum temperature of the day(°C)
4.	MaxTemp	Maximum temperature of the day(°C)
5.	Rainfall	Amount of rainfall recorded on the day(mm)
6.	Evaporation	Recorded evaporation using classA evaporation method
7.	Sunshine	Number of hours bright sunshine on the day
8.	WindGustDir	Wind gust direction(W/SE/SW/N/SSE)
9.	Windspeed	Wind speed in km/hr
10.	Pressure	Pressure throughout the day
11.	Humidity	Humidity of the day

We have constructed the correlation matrix for missing values using msno library. From the above graph we see that cloud 9 am is negatively correlated with wind gust direction and wind gust speed. Rainfall is strongly correlated with Rain Tomorrow. Wind gust speed is strongly correlated with wind gust direction.

Table I
Dataset feature description

VI. Proposed Framework

A. Preprocessing

In the machine learning always the first step is preprocessing the data. In the cleaning step defaults tasks are finding the missing values or null values, removing the duplicate rows/columns, renaming the columns as per the convenience. In our project we have identified missing values and sorted the missing value percentages in the descending order and dropped the missing values using pandas dropna method. As a new method we have constructed the correlation matrix for null values using msno library to check the correlation of missing value columns to one another.

B. Exploratory Data Analysis

In the exploratory data analysis step we have analysed the patterns in the data.For all the data analysis we have used plotly library for the dynamic visualizations.

- In the exploratory data analysis we have found the following observations:
- The target class is imbalanced. In the target class 'No' is the dominant class which has 78 percent samples. Yes category is the non dominant class which has 22 percent of samples
- Using plotly scatter plot we have explored the location wise data rainfall vs evoparation.
- Using plotly histogram we have explored the wind gust direction and wind gust speed.
- Using plotly histogram we have explore the pressure and rain today

C. SMOTE Analysis

Due to data imbalance we have applied SMOTE analysis to improve the non dominant class. SMOTE object is created using scikit-learn RandomOverSampler. This random sampler is trained using training data. After the training process we have visualized the data scatter plot to check the how the data improved. The target class improved and the distribution is almost equal for both the categories. After checking the data distribution, data is split into training and testing using sklearn train and test split function. The split ratio of train and test set is 80 to 20 ratio.

Github link : https://github.com/Jitheandra/Finalproject

D. Feature selection

After the data split ANOVA feature selection is applied to the data

ANOVA feature selection method is implemented using sklearn

There are various ways that we can implement ANOVA feature selection method

Using SelectK best function and the score function as fclassif and k is all

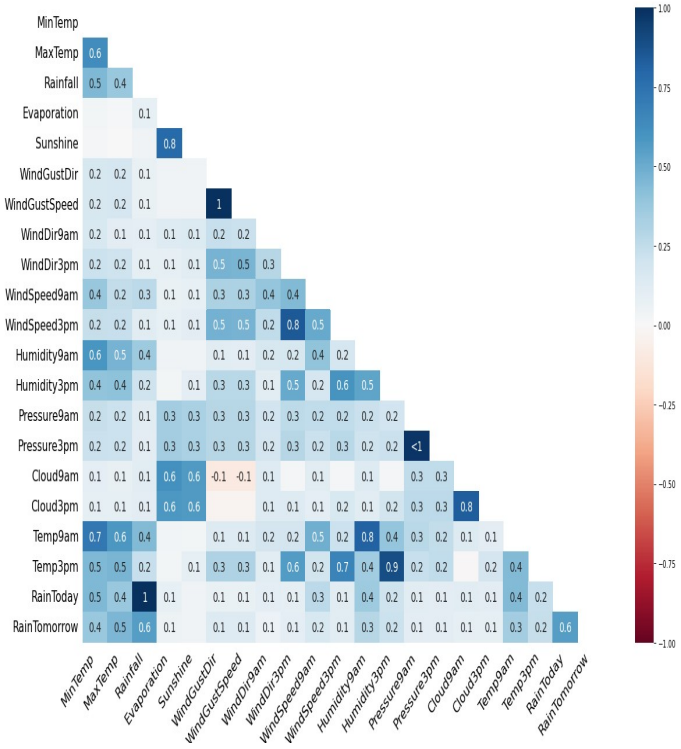

Figure 1. Missing value correlation matrix(msno)

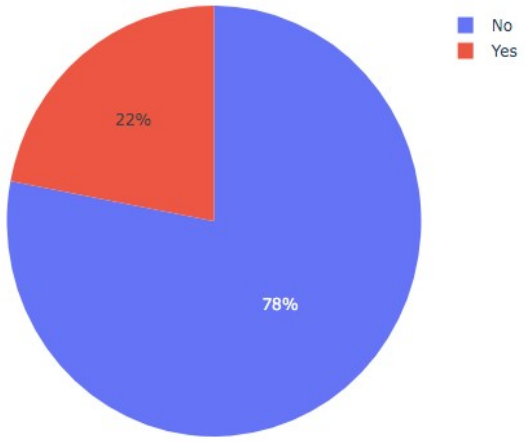

Figure 2. Distribution of targets

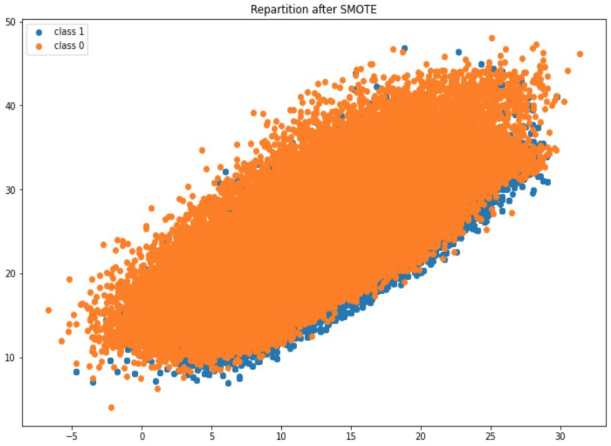

Figure 3. Target class SMOTE analysis

- feature selection object is trained using training data and transformed the training input and testing input that is training labels and testing labels.
- after transforming the data it gives the score of the feature importance of the each column. In the following graph we have visualized the feature importance of the each column i.e. feature

From the graph we can see only 5 features have highest importance in the data. With the these feature we will train the machine learning models. We will train 4 machine learning models decision tree, Adaboost, Gradient Boosting and Kmeans classifier.

E. Machine learning models he working of tree based algorithms work like human brain while taking the decisions. Generally these machine learning

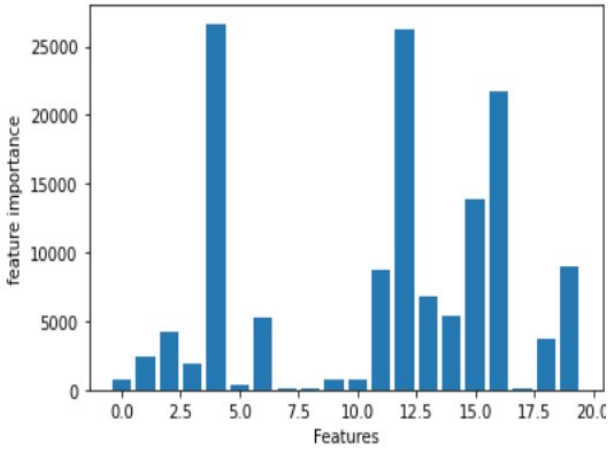

Figure 4. Feature importance

Github link : https://github.com/Jitheandra/Finalproject

algorithms are adopted from human brain functionality. Types of functionality of the algorithms are categorized into rule based and probabilistic approach. Humans also work on rule based approaches and probabilistic approaches. Rule based approaches are step by step process of solving the problem considering rules for example planning the vacation you will consider the price, distance these are the features and the rules are high price or low price and near or far away from the location.

Decision Tree classifier: Decision Tree is a widely used tree based supervised machine learning algorithm. This works on the rule based approach. If the target is yes or no i.e. rain tomorrow is yes or no we continuously split the data until we reach the root node.

Boosting methods: AdaBoost and Gradient boosting algorithms also used to train the model in the second step. These methods are ensemble models where they use the weak learners to train the data and based on the statistical approaches they choose the best predictions. These statistical methods can be mean and median etc.,

K-Means: For a change we have implemented the unsupervised learning algorithms. In that we have implemented K-Means clustering.

These three different classification(binary classification) algorithms are implemented to check the model performances. In the result analysis we have compared the results of the these 4 algorithms. Various performance metrics are used to measure the performance other than accuracy. FOr any kind of model to understand the over all performance of the model accuracy parameter is not enough in addition to accuracy we have to explore false positives and false negatives which gives the whole picture of the performance of the models. Sklearn or scikit learn library provides the methods to calculate these parameters i.e. precision, recall and F1 score i.e. classification report and confusion matrix gives the overall picture of the false positives and false negatives.After constructing the confusion matrix we will implement that using seaborn heatmap. These reports are analysed in the result analysis.

VII. Results Analysis

The following figures show the classification report and confusion matrix of decision tree classifier. In the classification report we can observe category 1 highest recall and F1score that is 97 percent and 90 percent. Category 0 has the highest precision i.e. 97 percent. The accuracy of the model is 91 percent. In the confusion matrix we can observe total correct classifications are 15976 remaining are incorrect classifications. The incorrect classification rate is high for category 1 i.e. 1380 total 1380 samples are classified as 0 category but originally they are category 1.

	precision	recall	f1-score	support
0	0.97	0.84	0.90	8754
1	0.86	0.97	0.91	8844
accuracy			0.91	17598
macro avg	0.91	0.91	0.91	17598
weighted avg	0.91	0.91	0.91	17598

Figure 5. Decision tree classification report

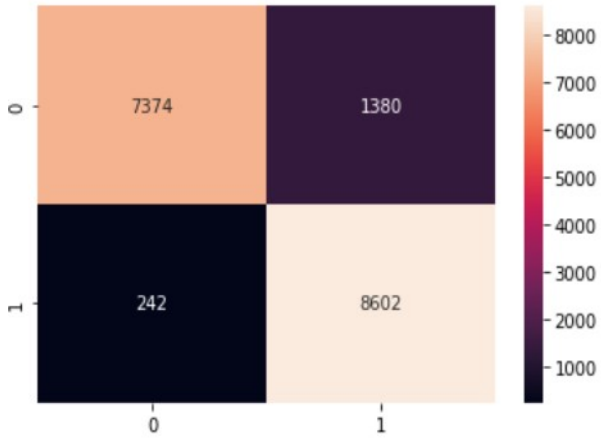

Figure 6. Decision tree confusion matrix

As a second step we have implemented boosting methods.One is Adaboost algorithm and the other one Gradient Boosting algorithm.

In the classification report of AdaBoost classifier we can observe the precision of category 0 has the highest score that is 97 percent and category 1 has the highest F1 score and recall that is 97 percent and 91 percent respectively.

From the confusion matrix of AdaBoost algorithm we can observe that the rate of correct classification is 14132. There is no much difference between the incorrect classifications of category 0 and category 1. The number of mis classification samples are 1743 for category 0 and 1723 for category 1.

From the Gradient Boosting algorithm we can summarize the rest performance like this: Classification report for other

	precision	recall	f1-score	support
0	0.80	0.80	0.80	8754
1	0.80	0.80	0.80	8844
accuracy			0.80	17598
macro avg	0.80	0.80	0.80	17598
weighted avg	0.80	0.80	0.80	17598

Github link : https://github.com/Jitheandra/Finalproject

Figure 7. Classification AdaBoost

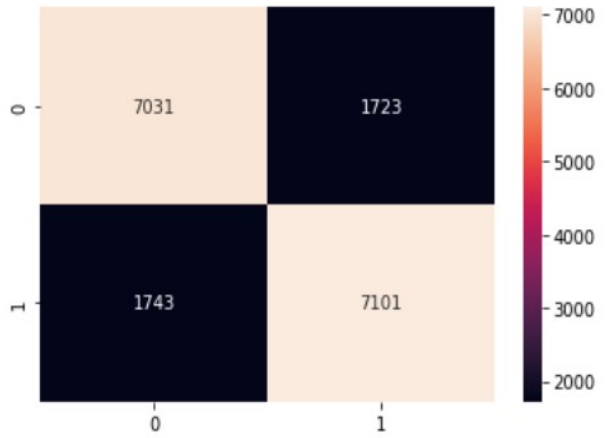

Figure 8. Confusion matrix AdaBoost

performance metrics are precision, recall and F1 score. From the classification report of Ada Boost classifier we can observe the test accuracy is 80 percent. For both the categories we have same precision, recall and F1 score that is 80 percent. From the confusion matrix we can observe incorrect classification samples are more or like same for category 0 and 1. For category 0 it is 1751 and for category 1 it is 1762.

	precision	recall	f1-score	support
0	0.80	0.80	0.80	8754
1	0.80	0.80	0.80	8844
accuracy			0.80	17598
macro avg	0.80	0.80	0.80	17598
weighted avg	0.80	0.80	0.80	17598

Figure 9. classification report Naive Bayes

In the last step we have implemented the unsupervised machine learning algorithm which is K-means clustering. This algorithm is implemented using sklearn library. For selecting the optimal number of clusters we have visualized the elbow plot. From the elbow plot we can see that the optimal number of clusters are 2. Since this is a unsupervised machine learning

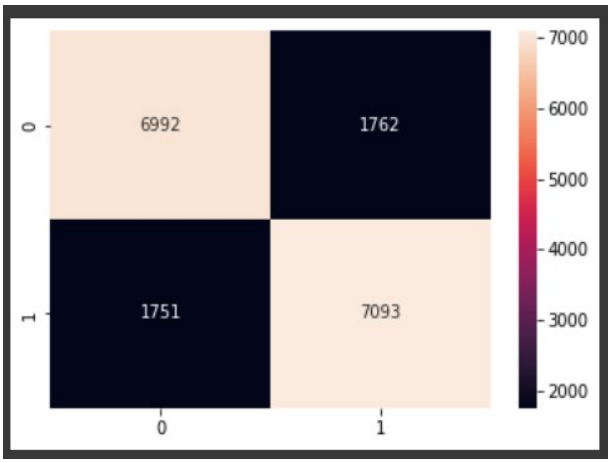

Figure 10. confusion matrix Naive Bayes

algorithm we can not publish the classification report and confusion matrix. The accuracy of the model is 88 percent.

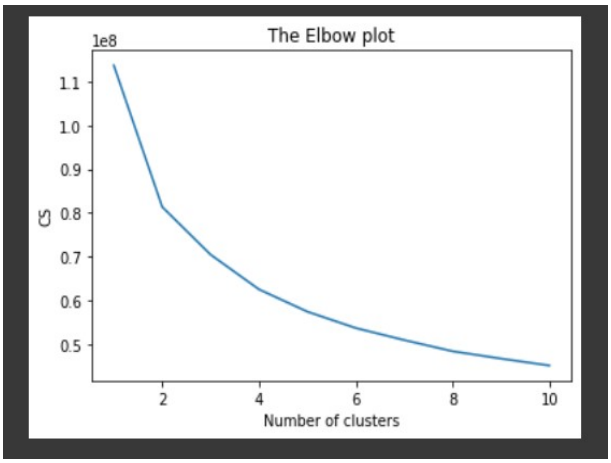

Figure 11. Elbowplot

VIII. Results Summary

To summarize the results in total we have implemented the 4 algorithms.3 supervised machine learning algorithms and one unsupervised learning machine learning algorithm. Four performance metrics are analysed for each algorithm. In the supervised learning category we have implemented Decision Tree, Ada Boost, Gradient Boosting algorithm. The accuracies of the algorithms Decision Tree, Ada Boost, Gradient Boosting and K-means clustering algorithm as follows 91, 80,80 and 88 percent. Amongst these algorithms Decision Tree performs well in terms of test accuracy. The next algorithm that performs well is K-Means clustering. The embedded methods AdaBoost and Gradient Boosting showed the same results holding 80 percent accuracy.

References

Github link : https://github.com/Jitheandra/Finalproject

[1] Jian Rong Ban, Qi Gou, and Ya Shi Li. Study on rainfall prediction of yibin city based on gru and xgboost. In *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pages 1–5. IEEE, 2022.

[2] Shivam Bang, Rajat Bishnoi, Ankit Singh Chauhan, Akshay Kumar Dixit, and Indu Chawla. Fuzzy logic based crop yield prediction using temperature and rainfall parameters predicted through arma, sarima, and armax models. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE, 2019.

[3] Cmak Zeelan Basha, Nagulla Bhavana, Ponduru Bhavya, and V Sowmya. Rainfall prediction using machine learning & deep learning techniques. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 92–97. IEEE, 2020.

[4] S Biruntha, BS Sowmiya, R Subashri, and M Vasanth. Rainfall prediction using knn and decision tree. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1757–1763. IEEE, 2022.

[5] Yajnaseni Dash, SK Mishra, and BK Panigrahi. Rainfall prediction of a maritime state (kerala), india using slfn and elm techniques. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pages 1714–1718. IEEE, 2017.

[6] Pedro Garcia-del Pino, Domingo Pimienta-del Valle, Ana Benarroch, and Jose M Riera. Evaluation of recent prediction models using a long-´ term database of rainfall rate data. In *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pages 20–24. IEEE, 2017.

[7] Liu Huang, Xuejun Liu, and Heyi Wei. Urban rainfall forecasting method based on multi-model prediction information fusion. In *2020 6th International Conference on Information Management (ICIM)*, pages 210–214. IEEE, 2020.

[8] Yogesh Kumar Joshi, Udit Chawla, and Shipra Shukla. Rainfall prediction using data visualisation techniques. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 327–331. IEEE, 2020.

[9] D Karthika and K Karthikeyan. Analysis of mathematical models for rainfall prediction using seasonal rainfall data: A case study for tamil nadu, india. In *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pages 01–04. IEEE, 2022.

[10] Yerin Kim and Sungwook Hong. Very short-term rainfall prediction using ground radar observations and conditional generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–8, 2021.

[11] Cheng Liu, Yi-Mei Tian, and Xiao-Hua Wang. Study of rainfall prediction model based on gm (1, 1)-markov chain. In *2011 International Symposium on Water Resource and Environmental Protection*, volume 1, pages 744–747. IEEE, 2011.

[12] Aishwarya Himanshu Manek and Parikshit Kishor Singh. Comparative study of neural network architectures for rainfall prediction. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pages 171–174. IEEE, 2016.

[13] Santosh Kumar Nanda, Debi Prasad Tripathy, Simanta Kumar Nayak, and Subhasis Mohapatra. Prediction of rainfall in india using artificial neural network (ann) models. *International Journal of Intelligent Systems and Applications*, 5(12):1, 2013.

[14] Jinghao Niu and Wei Zhang. Comparative analysis of statistical models in rainfall prediction. In *2015 IEEE International Conference on Information and Automation*, pages 2187–2190. IEEE, 2015.

[15] J Refonaa and M Lakshmi. Cognitive computing techniques based rainfall prediction—a study. In *2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)*, pages 142–144. IEEE, 2017.

[16] Sandeep Kumar Sunori, Amit Mittal, Sudhanshu Maurya, Pushpa Bhakuni Negi, Shweta Arora, Kavita Ajay Joshi, and Pradeep Juneja. Rainfall prediction using subtractive clustering and levenberg-marquardt algorithms. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1458–1463. IEEE, 2021.

[17] Feifei Tang, Xifei Wang, Zhangpeng Wang, Min Xiao, Ying Ma, and Yun Wan. Displacement prediction of rainfall induced landslide based on adaboost bp neural network. In *2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS)*, pages 794–799. IEEE, 2022.

[18] Dinesh K Vishwakarma, Amandeep Singh, Amit Kushwaha, and Ayush Sharma. Comparative study on influence of moon's phases in rainfall prediction. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–8. IEEE, 2021.

[19] Veera Ankalu Vuyyuru, Giduturi Apparao, and S Anuradha. Rainfall prediction using machine learning based ensemble model. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–4. IEEE, 2021.

[20] Zhi-liang Wang and Hui-hua Sheng. Rainfall prediction using generalized regression neural network: case study zhengzhou. In *2010 International conference on computational and information sciences*, pages 1265–1268. IEEE, 2010.

Github link : https://github.com/Jitheandra/Finalproject