University of Essex
Department of Mathematical Sciences

MA981 : DISSERTATION

Title of Your Project:Rainfall Prediction using machine learning

Your Name: Prabhat Dhar

Supervisor: **Your Supervisor:**

December 12, 2025

Colchester

Rainfall Prediction using Machine Learning technique

Abstract

In this project, we will use machine learning techniques in order to predict the rainfall happening and after that we will forecast the future rainfall for the next 30 days using ARIMA model(Auto Regressive Integrated Moving Average). One of the most difficult tasks in weather forecasting is predicting rainfall. Therefore, it is very important if we can estimate rainfall accurately and on time in advance since only then can we take safety measures in advance for ongoing building projects, transportation tasks, airline operations, agricultural tasks, and floods. Some of the external factors that can affect rainfall has been taken into accounts like temperature, humidity, wind direction , wind speed, season for predicting rainfall using machine learning techniques. Because of the harsh weather, it is growing harder to predict the amount of rain to fall. If we take an examples of countries like India which depend mostly on agriculture so here accurate prediction of rainfall is very important.

Using machine learning approaches, this study suggests a revolutionary real-time rainfall forecast method. For forecasting rainfall, our suggested study employs 4 distinct supervised machine learning algorithm strategies, including the random forest method, logistics regression, K Nearest Neighbors, and XG Boost and Pycaret which is an auto machine learning libraries. Before employing the classification method, all pre-processing procedures, such as data cleaning and normalisation, were completed on the dataset. The ARIMA (Auto Regressive Integrated Moving Average)model is then used to anticipate future rainfall.Both the geographic characteristic as well as the time horizon comprised has major influence on the weather forecasting.LSTM(Long Short Term Memory) techniques in also used for forecasting rainfall but in our case it does not perform good in terms of giving the output.
In terms of experimental result we see that XG Boast technique perform the best in term of accuracy with 86 percent the predicting the rainfall over other models like Logistics regression(83.86 percent), Random Forest Method(85.25 percent), Gaussian Naive Bayes Classifier (80.73 percent), K Nearesrt Neighbour(82.83 percent) and Latent Dirichlet Allocation(83.95 percent) .

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Predicting rainfall can assist prevent flooding, save human lives and property. Additionally, it aids in the management of water resources. Prior rainfall data is useful to farmers for improved crop management, which boosts the nation's economy. Rainfall prediction is a difficult assignment for meteorological experts because of variations in the timing and amount of the precipitation. Weather forecasting is the most popular service offered by the meteorological department for all nations worldwide. Due to the need for several specialists and the uncertainty of every calls, the process is quite hard.

Precipitation is the most essential component in nations where agriculture accounts for the bulk of economic activity, hence timely measurements of precipitation are important in terms of how they affect the country's overall economy. Our daily lives are greatly influenced by the climate. Since the beginning of human development, humans have been preoccupied with worrying about climate change. One of the most difficult problems in science and technology during the last couple of centuries has to do with forecasting the weather. Prediction is the phenomenon of foreseeing potential outcomes for a system. Instruments on the ground and satellites using remote sensing gather the most recent weather measurements. Many things, including controlling aviation operations, planning water resource management, issuing early flood warnings, and limiting transport and construction activities [**?**] [**?**], can benefit from accurate and timely rainfall forecast. Due to climate change, it is increasingly difficult to estimate rainfall accurately nowadays. Researchers have continually been attempting to forecast rainfall as accurately as possible by combining and enhancing data mining tools [**?**].

Supervised learning algorithm and unsupervised learning algorithms are two types of data mining tecniques.So supervised learning is a machine learning tasks where a function can be learnt on input-output pairs based on how the input is mapped with the output. [?] From labelled training data made up of a collection of training instances [?], it infers a function. Each example in supervised learning is a pair that includes an input item (usually a vector) and an intended output value (also called the supervisory signal). An inferred function is generated by a supervised learning algorithm from the training data, which may then be used to map fresh samples. The algorithm will be able to accurately determine the class labels for instances that are not yet visible in an ideal environment. This necessitates that the learning algorithm generalise in a "reasonable" manner from the training data to hypothetical circumstances. On the other side, unsupervised approaches don't need any training; instead of using pre-classified data, these methods employ algorithms to draw out hidden structure from unlabeled data. Recent studies have shown that for predicting rainfall, researchers favour integrated approaches because of their high accuracy.

To develop a prediction model for precise rainfall, forecasting is one of the greatest problems for academics from a number of domains, including meteorological data mining, environmental machine learning, functional hydrology, and numerical forecasting [?]. How to deduce previous forecasts and use future predictions is a frequent query in these challenges. The main process in rainfall is often made up of a number of smaller processes. Water resource managers and hydrologists [?] are now concerned about the altering rainfall pattern caused by climate change. According to Srivastava et al [?] and Islam et al [?], variations in rainfall amounts and frequency have a direct impact on stream flow patterns, their demand, the spatiotemporal distribution of runoff, ground water reserves, and soil moisture. As a result, these changes demonstrated the wide-ranging effects on the environment, ocean, biodiversity, agriculture, and food security.

Here in this project we will deal with the australia weather data and predict whether it will rain or not. There have been several previous attempts to find patterns for forecasting rainfall in Australia. It was discovered that ENSO significantly affects the variability of Australian rainfall [?]. There has been some investigation on how climatic indicators affect rainfall in Australia. The combined impact of climate indicators, however, has only been the subject of a relatively small number of research [?] [?] [?]. Predicting rainfall accurately is a major problem in many nations, particularly Australia where the climate can be quite unpredictable.

Water is an essential component of practically every business in Australia, especially agriculture. Making educated policy, planning, and management choices and contributing to the more sustainable functioning of water resource systems requires an accurate [?] estimate of the availability of water. In many countries, notably Australia where the climate may be highly variable, precisely predicting rainfall is a significant challenge. In Australia, water is a necessary component of almost every industry, notably agriculture. An accurate assessment of the water availability [?] is necessary to make informed policy, planning, and management decisions and to contribute to the more sustainable operation of water resource systems. Since the start of the twenty-first century, climate change has been a serious problem in Australia [?] . Climate change is making Australia hotter and more vulnerable to heat waves, bushfires, droughts, and floods. Australia's yearly temperature has risen by 1.4 degrees centigrade since the turn of the 20th century, increasing at a rate that has doubled in the last 50 years. In southwestern Australia there is a decline of 10-20 percent of rainfall since 1970 while southeastern Australia has experienced a little less decline since the 1990s. In summer rainfall is more common in Australia rather than in winter.

The two types of rainfall forecasting models are physical models and data-driven models. All significant physical processes that affect rainfall are modelled using physical models based on physical principles. In order to forecast the future, data-driven models employ previous data. They are mostly based on statistical and computational intelligence methodologies. Comparative studies have demonstrated that data driven models outperform physical models in terms of precipitation prediction [?] [?]. The most popular techniques among the data driven models that were used: ARIMA(Auto Regressive Integrated Moving Average) , K Nearest Neighbors , Support Vector Machine and Multi Linear Regression model. The month determines the beginning, duration, and conclusion of the rainy season when predicting rainfall. Compared to seasonal rainfall data, monthly rainfall statistics more accurately depict the distribution of rainfall throughout the year. The amount of rainfall that falls each month has a significant impact on hydrological and agricultural processes. The quality of decision-making in these tasks can be enhanced by its precise prediction.

CHAPTER 2

Literature Review

P. Goswami and Srividya incorporated RNN and TDNN features, and their research showed that composite models outperform single models in terms of accuracy [?]. Multilayer Feed Forward Neural Networks (MLFNN) were utilised by C. Venkatesan et al [?] to forecast rainfall during the Indian summer monsoon. To forecast the weather, the Error Back Propagation (EBP) algorithm is learned and used. The analysis of three network models with two, three, and 10 input parameters. The output result was also contrasted with the statistical models of A. Sahai and co. India's Summer Monsoon Rainfall Prediction Using Monthly and Seasonal Time Series Used Error Back Propagation Algorithm [?]. For rainfall prediction, they used data from the past five years' worth of monthly and seasonal mean rainfall values. For the purpose of predicting annual rainfall in the Kerala region, N.Philip and K.Joseph deployed an ABF neural network. According to their research, ABFNN outperforms Fourier analysis [?].

A support vector machine-based model using data from the city of Sydney from 1957 to 2019 was created in [?]. The goal of this research is to conduct an exploratory analysis on the use of machine learning algorithms to model the phenomenon of rain, using a dataset of precipitation measurements, atmospheric conditions, and other characteristics from the major Australian cities over the previous ten years as an example. In order to assess its effectiveness and usefulness, several of the most significant machine learning algorithms were also applied to this data. [?] describes a collection of experiments where models based on Logistic Regression, Decision Tree, K Nearest Neighbor, Rule-based, and Ensembles are used to generate predictions of whether it will rain tomorrow or not based on the meteorological data for that specific day for key Australian cities. Although the ANN has many benefits for simulating physical processes, it has certain drawbacks when dealing with severely nonstationary time series since its precision is not very good [?]. In these situations, hybrid models can be applied to address the shortcomings of AI-based techniques and enhance their output. In signal processing analysis, the wavelet transform has been introduced, and many wave data analyzers have benefited from this advancement [?]. There have been several previous attempts to find patterns for forecasting rainfall in Australia. It was discovered that ENSO significantly affects the variability of Australian rainfall [?]. There has been some investigation on how climatic indicators affect rainfall in Australia. Using a variety of delayed climatic factors may help convey priceless predictive information in mid-term rainfall forecasting, claim [?]. Additionally, they discovered that general circulation model predictions for mid-term and monthly rainfall, particularly for POAMA (Predictive Ocean Atmosphere Model for Australia), are unsatisfactory. They accepted that the climatic factors were helpful for predicting rainfall in Queensland. Finding accurate and trustworthy models is therefore crucial for improved management of water resources and catastrophe prevention, especially for medium-term forecasts.

[?] used the K-NN, initially a non-parametric statistical method, to time series prediction. Forecasters like this strategy because it is straightforward and simple to utilise. By combining the future values of the past sequences that are most similar to the current sequence, the K-NN approach determines the sequence's subsequent value. The foundation of Support Vector Machines (SVM) is statistical learning theory [?]. It creates an ideal hyperplane by using a non-linear mapping function to translate the original data into a high-dimensional feature space. The findings in [?] these publications demonstrate that the majority of models do not successfully anticipate the intricate physical process of rainfall. Predictions of this nature are particularly difficult in Australia because of its frequently unpredictable climate. Therefore, the creation of reliable rainfall forecast techniques remains a crucial area of study. The fundamental [?] components of hydrology, climatology, and meteorology research across [?] the world have been to examine the variability, changes in pattern, and existence of trend in rainfall over various geographical horizons. The majority of [?] studies have employed both parametric and non-parametric techniques, such as the regression test [?] Mann-Kendall test, Kendall rank correlation test, Sen's slope estimation, and Spearman rank correlation test

[?]. Since the Mann-Kendall test is one of the most widely used global methods for trend identification in hydrology, climatology, and meteorology, it was used in the current study to identify the trend in rainfall. Non-parametric tests were utilised in this study because they may be used to independent time series data and are less susceptible to outliers [?].

For the purpose of predicting rainfall, K. Htike and O. Khalifa [?] used data from yearly, biennial, quarterly, and monthly rainfall. For the purpose of forecasting rainfall, they trained four distinct Focused Time Delay Neural Networks (FTDNN). When annual rainfall data is used for training, the FTDNN model has the highest forecast accuracy. The K-mean clustering approach linked with the decision tree algorithm, known as CART, was developed by S. Kannan and S. Ghosh [?] and is used to generate rainfall states from large-scale meteorological data in a river basin. Using K-mean clustering, the daily rainfall state is determined from historical daily multi-site rainfall data. Short-term rainfall was forecasted by M. Kannan et al. [?] For prediction tasks, empirical method techniques are applied. A specific region's data from three distinct months throughout five years is evaluated. The items are grouped using clustering.

Supervised learning algorithms can be used in case of rainfall prediction because here the data or samples are labelled as whether it is going to rain or not. In dealing with such problems Artifical Neural Network(ANN) and Support Vector Machine(SVM) [?] are most used supervised learning algorith for this purpose. Here we will be using SVM as it is very efficient. Hyperplane or set of hyperplane gets generated in case of Support Vector Machine as it is parameter method due to which it gets genearted between the classifier to the closest data neighbour [?]. SVM constructs a maximum margin hyperplane based on the training samples that divides the group of points xi for which yi = 1 (i.e., rain) from the group of points xi for which yi = 1 (i.e., sun) (i.e., no-rain). This is done to maximise the distance between the hyperplane and the nearest point xi from either group. The limitations of statistical techniques such as auto regressive (AR), moving average (MA), auto regressive moving average (ARMA), and auto regressive integrated moving average (ARIMA) are as follows: AR model regresses past values, MA model uses past error as explanatory variables, and ARMA model can perform for stationary time series data [?]. However, the use of artificial intelligence (AI) models, such as machine learning approaches, has lately attracted attention. Unlike physical models, AI models perform very well since they do not require a large amount of information but can manage large and complex data sets if they are offered [?]. We can use this in non stationary data also.

There are just a few studies [?] in the literature that integrate multiple meteorological parameters and apply a machine-learning-based technique for rainfall prediction. The majority of these are concerned with the implementation and intercomparison of various machine learning algorithms. However, it is more interesting for the meteorological community to determine the relevant characteristics for rainfall forecast, their dependency, and their level of participation in the prediction. The BPNN model was used by the authors of [?] to construct a daily rainfall forecasting system for landslide early warning. The study suggested four back propagation neural networks with various topologies that might predict how much rain will fall the next day. By examining the correlation between the anticipated intensity and the specified rainfall thresholds, the likelihood of a landslide occurring is projected [?]. In order to compare the SVR model's performance with that of the BPNN and the autoregressive integrated moving average (ARIMA) models for predicting monthly rainfall intensity, a research is reported in [10]. The factors that produced the most accurate inferences were discovered by the authors using particle swarm optimization (PSO) techniques. Due to its capacity to extract long-term dependencies and recall data selectively, the long short term memory (LSTM) network is an artificial recurrent neural network that is of tremendous relevance when working with continuous time-series data. Deep Learning has been applied successfully in ANN during the last several years to solve complicated issues [?]. A collection of multilayer structures that are taught using unsupervised methods are collectively referred to as "deep learning." The major advancement is learning a non-linear, compact, and correct representation of the data using unsupervised approaches in the hopes that the new representation will aid in the current prediction job. A computational model called ANN was motivated by the human brain [?]. A large number of linked, well-structured, parallel-operating neurons make up an ANN. Neural networks can be divided into single-layer or multi-layer categories. The hidden layer is the one that exists between the input layer and the output layer. One input layer with weights assigned to each node and one output layer make up a single-layer feed forward (SLFF) neural network.

CHAPTER 3

Data

Bureau of Meteorology that comes under Australian goverment has provided with the weather forecast details in their website. Investigation of weather report in Australia has been done taken into several external factors like humidity, wind speed, wind direction, temperature etc from the year 2008 to 2020 and we will predict whether it is going to rain or not and then forecast amount of rainfall that will occur in next 30 days. All the necesaary pre-processing steps and cleaning of data has been done before doing the prediction. We has downloaded the dataset from www.kaggle.com and it contains 1,45,460 rows and 23 columns and the details of overview in data and no of missing values present are given below:

Table 3.1: Overview of data(table 1)

Column name	No. of missing value	missing value(percent)
Date	0	0.00
Location	0	0.00
MinTemp	1485	1.02
MaxTemp	1261	0.86
Rainfall	3261	2.24
Evaporation	62790	43.16
Sunshine	69835	48
Windgustdir	10236	7.09
Windgutspeed	10263	7.05
Winddir9am	10566	7.26
Winddir3pm	4228	2.90
WindSpeed9am	1767	1.21
WindSpeed3pm	3062	2.10
Humidity9am	2654	1.82

Table 3.2: Overview of data(table 1)

Column name	No. of missing value	missing value(percent)
Humidity3pm	4507	3.09
Pressure9am	15065	10.35
Pressure3pm	15028	10.33
Cloud9am	55888	38.42
Cloud3pm	59358	40.80
Temp9am	1767	1.21
Temp3pm	3609	2.48
RainToday	3261	2.24
RainTomorrow	3267	2.24

3.0.1 Data Pre-Processing

So from the data we can know that there are lots of missing value in the dataset and we have to deal with that.

Random Imputation techniques is used to fill the missing value which has more than 30 percent of it in the data. A straightforward and well-liked method of data imputation includes estimating a value from a column based on the values that are already there, and then replacing any missing values in the column with the estimated statistics. The rest of the continuous columns which has low percentage of missing value we have selected that columns and used mean imputation techniques there and also the categorical columns are also filled with same technique in the label conding part. Here in this case in each variable of the observed value mean is found out and the missing values for that variable are imputation by this mean.

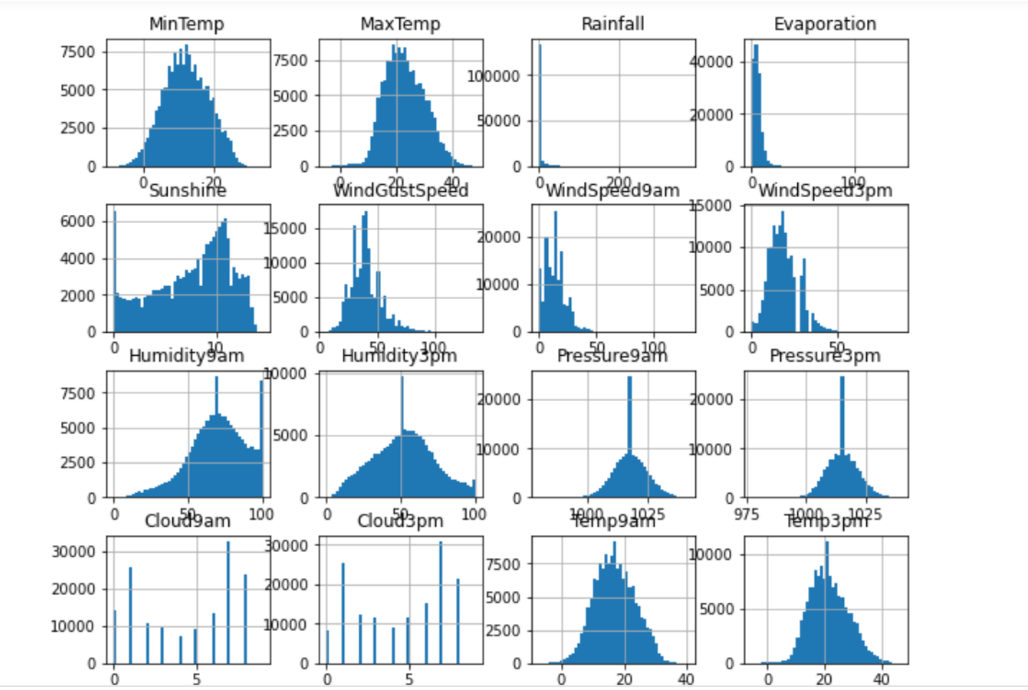


Figure 3.1: Visualization of data used(source code)

From the above figure using bar graph we have visualize the data. We can see that the rainfall and evaporation contains a lot of missing value.

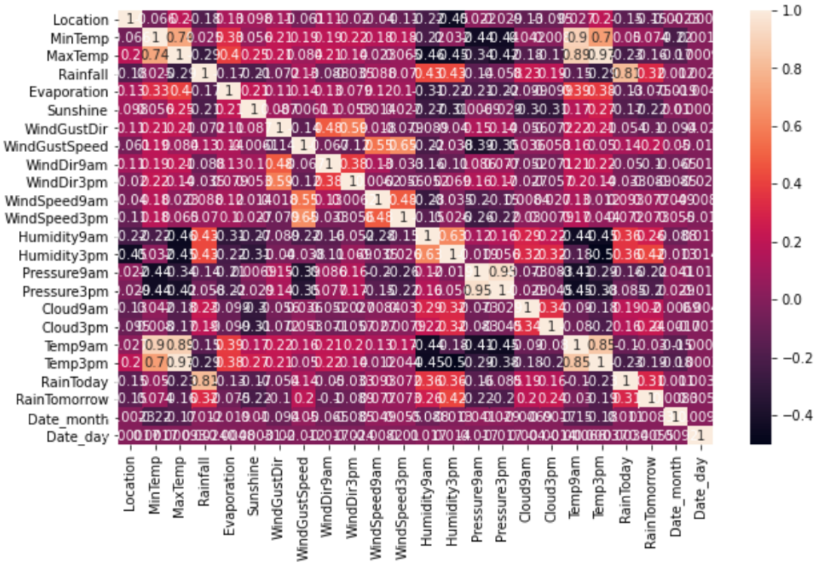


Figure 3.2: Heatmap(source code)

We can find out the presence of multi-collinearity using the heatmap. When the dataset has a large number of independent variable it is possible that few of these independent variable may be highly correlated [?]. The existence of highly correlation between the independent variable is called multi-collinearity and the presence of can destabilize the model and we have to remove this before building a model. So for avoiding multi-collinearity we have keep only one columns for each group of highly correlated variables and removed the others. Using boxplot we have checked whether outliers was present or not. Performing descriptive analytics before moving to predictive analytics is always a good practice. So with the help of descriptive analytics we can understand the variability in the model and visualization of data is done. Another important techniques is that using scatter plot we can reveal if there is any obvious relationship between the two variable is there or not. It helps us to find the functional relationship between the dependent or outcome variable and feature.

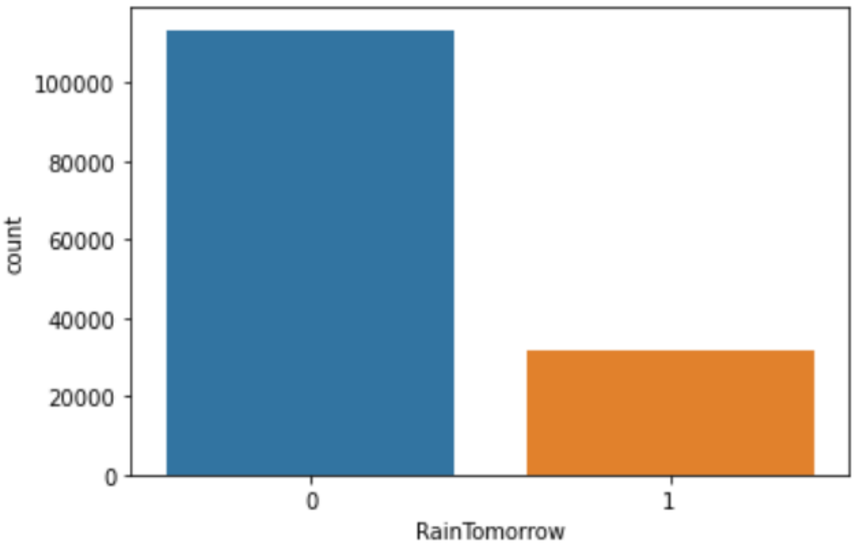


Figure 3.3: Imbalance target columns(source code)

From figure 3.3 we can say that the outcome variable that is the RainTommorrow columns are imbalance that it is in the ration of 78:22
Significant discrepancies in the distribution of the classes within a dataset are what constitute an unbalanced dataset. This indicates that a dataset has a bias toward a certain class. An algorithm trained on the same data will be biassed toward the same class if the dataset is biassed toward that class. We need to deal with the imbalance dataset, In the field of statistics of Â data analysis, oversampling and undersampling are approaches used to modify a data set's class distribution (i.e., the proportion of each class or category represented). These phrases are utilised in machine learning as well as statistical sampling and survey design methods.

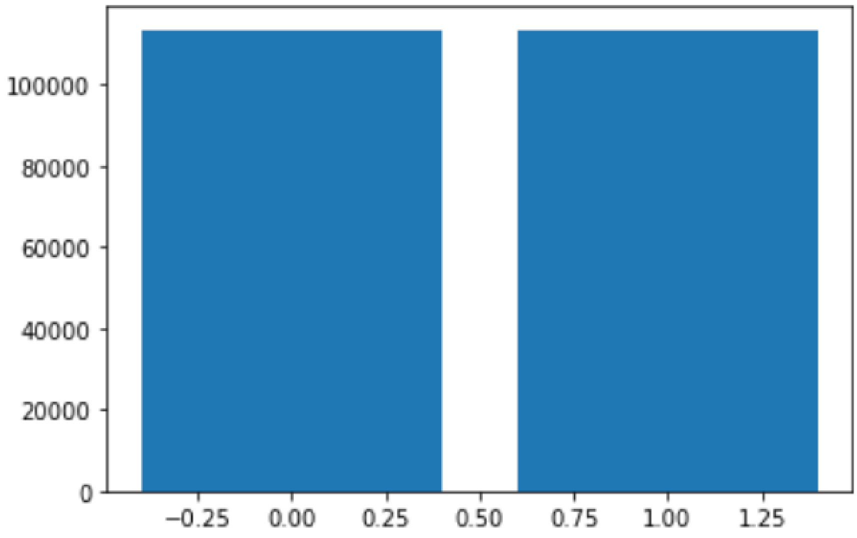


Figure 3.4: Balance target columns(source code)

In order to balance the data we have used the oversampling techniques.Random oversampling includes adding extra copies of some of the minority classes to the training data. Multiple oversampling attempts are possible. This is one of the first approaches that has been suggested, [?] and it has been shown to be reliable. Some of the minority class samples could be randomly selected with replacement rather than all of them being duplicated. A dataset used in a typical classification task can be oversampled using a variety of techniques (using a classification algorithm to classify a set of images, given a labelled training set of images). Synthetic Minority Over-sampling Technique (SMOTE) is the most used method. Â Consider some training data with s samples and f features in the feature space of the data to demonstrate how this approach functions. Keep in mind that these qualities are continuous for simplicity's sake. [?] Consider a dataset of birds for categorization as an illustration. Beak length, wingspan, and weight might represent the feature space for the minority class for which we wish to oversample (all continuous). Take a sample from the dataset and take into account its k nearest neighbours to oversample (in feature space). Take the vector between one of those k neighbours and the current data point to generate a synthetic data point. Divide this vector by a chance number x, which ranges from 0 to 1. [?] To produce the new, fabricated data point, add this to the existing data point.

We have used one hot encoding in the 2 columns namely RainToday and RainTomorrow. So A one-hot is a collection of bits in digital circuits and machine learning where the only valid value combinations are those that have a single high (1) bit and all other low bits (0) [**?**] . We have also performed label encoding by assigning values to some of the columns. Popular encoding methods for categorical data include label encoding. According to this method, an individual number is given to each label depending on its alphabetical order. Labels may be normalised using LabelEncoder. As long as they are hashable and comparable, it may also be used to convert non-numerical labels into numerical ones. There are some of the other pre-processing steps that is being done are:

- The dataset is divided into training and testing part that is 80 percent of the data is used in the training and rest 20 percent in testing part.
- We have groupby the two columns that is Location and RainTomorrow and used label encoder for our location column according to the target variable

We have also normalize the dataset that is used scaling to keep the value in range(0-1). A technique for normalising the variety of independent variables or features in data is called feature scaling. It is often carried out during the data preparation stage and is sometimes referred to as data normalisation in the context of data processing. Some machine learning algorithms' objective functions won't operate effectively without normalisation since raw data's value range varies greatly. For instance, many classifiers use the Euclidean distance to compute the distance between two locations. The distance will be determined by a specific feature if it has a wide range of values among the characteristics. To ensure that each feature contributes about proportionally to the final distance, the range of all characteristics should be standardised [**?**].

CHAPTER 4

Methodology

Here we will discuss about all the methods that we have used in rainfall prediction using machine learning techiques and finally we will say about the forecasting technique ARIMA(Auto Regressive Integrated Moving Average) which will say total rainfall that is going to happen in next 30 days. Here we have used stratified cross validation on the dataset in order to train all the model we used and we have'nt used any hyperparameter tuning for this.

4.0.1 Random Forest Method

One of the most used type of ensemble technique in machine learning is random forest method and it has good performance and scalibility that is why it is used in the industry. Random Forest Method is basically an ensemble of decision tree(classification and regression tree) where each samples is built from bootstrap samples and randomly selected subset of features without replacement. Usually decision tree does not have pruning and it is normally grown deep. The model accuracy in random forest method can be increased if we used more number of estimator. So some the hyperparameter used in random forest method is max depth, n estimator, max features, criterion and many others but here we have'nt used any hyperparameter tuning, we have normally used stratified cross validation and found out the accuracy in terms of classification report. The class that the majority of the trees choose is the output of the random forest for classification problems. The mean or average forecast of each individual tree is returned for regression tasks. The tendency of decision trees to overfit their training set is corrected by random decision forests. Although they frequently outperform decision trees, gradient enhanced trees are more accurate than random forests. But data features can impact how well they work. Bagging is the process of bootstrapping samples from the initial set to create several models and aggregating their outcomes for the outcome prediction. Bootstrapping and aggregating makes up the term bagging. Because the bootstrapping method reduces the model's variance without affecting its bias, the model performs better. As a result, whereas a single tree's predictions are very sensitive to noise in their training set, the average of several trees, provided the trees are uncorrelated, is not. Bootstrap sampling is a technique for de-correlating the trees by displaying them many training sets, as opposed to just training numerous trees on a single training set, which would result in heavily linked trees. Bagging repeatedly (B times) chooses a random sample with replacement of the training set and fits trees to these samples given a training set X = x1,..., xn with responses Y = y1,..., yn.

- Sample n training instances from X and Y with replacement; label these Xb and Yb.
- Train a regression or classification tree using the Xb, Yb data.

We have found out he confusion matrix where it gives us about the true positive, false positive , true negative and false negative and then we find out the accuracy of random forest method using the classfication report. We have also found out the ROC curve value(Receiver Operating Characteristic) and the AUC(Area under the curve) to show how the model performs. Receiver Operating Characteristic(ROC) curve tells us how the overall performance of the model is behaving and help us in selecting model. The plot between senstivity on vertical axes(true positive) and 1-speficity(false positive on horizontal axes is known as Receiver operating characteristic curve. ROC curve returns different cut off values and their corresponding false positive and true positive rates. With the help of these values we can create the Receiver Operating Characterictic curve and then it return the Area under the curve value.

4.0.2 Naive Bayes

In the field of statistics Naive Bayes classifier can be defined as the group of probabilistic classifier which we get after using Bayes Theorem with high independence proof between the features. Despite being among the simplest Bayesian network models [**?**], they may reach great levels of accuracy when used in conjunction with kernel density estimation [**?**]. In case

of learning problem if we are using Naive Bayes classifier we need lot of linear parameter in the number of variable as the scaling factor is very high. If we used closed form expression then the training can be done using maximum likelihood and we get linear time rather than iterative approximation that takes time in case of classifier. Simple Bayes and independent Bayes are two names for naÃ¯ve Bayes models that may be found in statistics literature. All of these titles refer to the classifier′s decision rule using the Bayes theorem, however naive Bayes is not (by definition) a Bayesian approach. Naive Bayes is a straightforward method for building classifiers. These models provide class labels to problem cases, which are represented as vectors of feature values, and the class labels are chosen from a limited set. For training such classifiers, there isn't just one technique, but rather a family of algorithms built on the premise that, given the class variable, the value of one feature is independent of the value of every other feature.
Using Bayes theorem conditional probability can be written as: p(Ck | x) = p(Ck)*p(x | (Ck)/p(x)
Thus we can sau Bayesian probability as:
posterior = (prior * likelihood) / evidence
Naive Bayes classifier works very well in case of tough real world scenario and that is why we are using it despite having some easy assumption. However, a thorough comparison of Bayes classification with other algorithms in 2006 revealed that alternative strategies, such as boosted trees or random forests, perform better [**?**].

Gaussian Naive Bayes

Gaussian Naive Bayes classifier has this advantage over other machine learning algorithm that we need less number of training data in order to estimate parameter for classifier. A assumption that is required to make if we are using continuous data is that each class should be equally divided according to a normal in case of continuous values

CHAPTER 5

Result and Discussion

Table 5.1: classification report of Random Forest Method

Classification Report	Precision	Recall	F1-Score	Support
0	0.87	0.96	0.91	22726
1	0.76	0.48	0.59	6366
Accuracy			0.85	29092
macro avg	0.81	0.72	0.75	29092
weighted avg	0.84	0.85	0.84	29092

Table 5.2: Classification report Gaussian Naive Bayes

Classification Report	Precision	Recall	F1-Score	Support
0	0.88	0.87	0.88	22726
1	0.56	0.57	0.56	6366
Accuracy			0.81	29092
macro avg	0.72	0.72	0.72	29092
weighted avg	0.81	0.81	0.81	29092

Table 5.3: Classification report K Nearest Neighbors

Classification Report	Precision	Recall	F1-Score	Support
0	0.87	0.92	0.89	22726
1	0.64	0.50	0.56	6366
Accuracy			0.83	29092
macro avg	0.75	0.71	0.73	29092
weighted avg	0.82	0.83	0.82	29092

Table 5.4: Classification report XG Booast Classifier

Classification Report	Precision	Recall	F1-Score	Support
0	0.88	0.95	0.91	22726
1	0.75	0.55	0.63	6366
Accuracy			0.86	29092
macro avg	0.81	0.75	0.77	29092
weighted avg	0.85	0.86	0.85	29092

Table 5.5: Classification report Logistic Regression

Classification Report	Precision	Recall	F1-Score	Support
0	0.86	0.95	0.90	22726
1	0.71	0.45	0.55	6366
Accuracy			0.84	29092
macro avg	0.78	0.70	0.72	29092
weighted avg	0.83	0.84	0.82	29092

,

Table 5.6: Classification report LDA

Classification Report	Precision	Recall	F1-Score	Support
0	0.85	0.92	0.91	22726
1	0.73	0.47	0.55	6366
Accuracy			0.84	29092
macro avg	0.79	0.70	0.75	29092
weighted avg	0.83	0.83	0.82	29092

So from the above table 5.1 to 5.6 we use used classification report in order to find the accuracy of the model that we have used like Random Forest Method, Gaussian Naive Bayes, Logistic Regression, K Nearest Neighbors , XG Boost Classifier and LDA . By looking at the accuracy we see that almost all the model perform very well in this data with accuracy around 82-86 percent range. XB Boost performs the best in our case with the accuracy of 86 percent but here we have not used any hyperparameter tuning like in Logistic Regression. The baseline accuracy in case of logistic regression was 83.94 percent and after using the hyparameter tuning with the best parameter value(C=1 and penalty = l2) the accuracy increased to 84.21 percent. The reason why XG Boost has the highest accuracy may be due to A regularised (L1 and L2) objective function with a function of convex loss and a model complexity penalty term is minimised using XGBoost. The training process iteratively continues in order to make the final prediction, and new trees are added that predict the residuals or errors of older trees that are then combined with older trees. This process is known as gradient boosting because, when adding new models, it uses a gradient descent algorithm to minimise the loss. LDA(Latent Dirichlet allocation) also has the accuracy of 84 percent and here also we have used hyperparameter tuning like(solver = 'svd','lsqr','eigen') and shrinkage in the range(0,1). Gaussian Naive Bayes has the lowest accuracy because here data is not normally distributed and also the variable are not continuous and we have used lots of data for training and this algorithm don't work well for all these case. The accuracy of random forest method and K Nearest Neighbor are 85 and 83 percent respectively.

Using the concept of sensitivity, specificity, precision and F score we can measure the model performance. Sensitivity(also known as recall or true positive rate) is the ability of model to correctly specify the positive and negative(specificity also known as true negative rate). Classification_report() method in sklearn.metrics gives the full report of precision, recall and F score for each class. The accuracy of predictions made by a classification algorithm is evaluated using a classification report. how many of the forecasts came true and how many didn't. More specifically, the metrics of a classification report are predicted using True Positives, False Positives, True Negatives, and False Negatives.

Now we will discuss little bit about measuring accuracy like sensitivity, specificity, precision and F score.

- Sensitivity: TP/(TP+FN). It is the conditional probability that predicted class is positive given that actual class is positive.

- Precision: TP/(TP+FP). Here in case of precision we get the actual as well as the predicted class positive.

- F score: 2* Recall* precision/(recall + precision). It combines both the precision and recall.

Accuracy = TP+TN/TP+TN+FP+FN

Table 5.7: Algorithm performance

Model	ROC(Receiver operating characteristic)	AUC(Area Under Curve)
Random Forest Method	0.718	0.88
Gaussian naive bayes	0.722	0.82
K Nearest neighbors	0.708	0.79
XG Boost Classifier	0.746	0.89
Logistics Regression	0.72	0.84
LDA	0.71	0.83

From table 5.7 we see the algorithm performance using the curve. Using ROC(Receiver Operating Characteristic) curve the overall performance of the model can be understood and it will help us in selecting which model is better. If we are given a random pair of positive and negative class record, Receiver Operating Characteristic gives proportion of pairs that are correctly classified. The plot between sensitivity(true positive rate) on the vertical axes and 1-specificity on the horizontal axes is known as Receiver Operating Characteristic curve.A method draw _roc() is written which takes the predicted value and classes and draws the curve.

Different threshold values and correspondng false positive and true positive rates is found out using metrics.roc_curve() ROC(Receiver Operating Characteristic) curve is created using the value and roc_auc_score() give us the area under the curve. If the model area under curve is higher that model will be preferred and it is also used for model selection like the Receiver Operating Characteristic.

For practical application of the model we need to get the area under value of more than 0.7 and in our case all the model that we have used has the value of more than 0.7. Any Area under curve value greater than 0.9 suggests that the model has performed outstandingly well. If the dataset is imbalanced then there is a chance that the Area Under Curve value might be greater than 0.9 however sensitivity and specificity may be poor. In our case the dataset was imbalanced but we have balanced the data in the data pre- processing part by using the oversampling techniques. In our case XG Boost classifier that the best area under curve and receiver operating characteristic value og 0.89 and 0.746 respectively. This model has the best performance and we have also seen for table 5.4 it has the best accuracy of 86 percent. Other model also has good performance if we see random forest method, Logistic regression , LDA in terms of area under curve(AUC) and Receiver Operating Characteric(ROC) of around 0.83 to 0.84 range. K Nearest Neighbors has the area under curve value of 0.79 which says model performance is not as good as the other model.

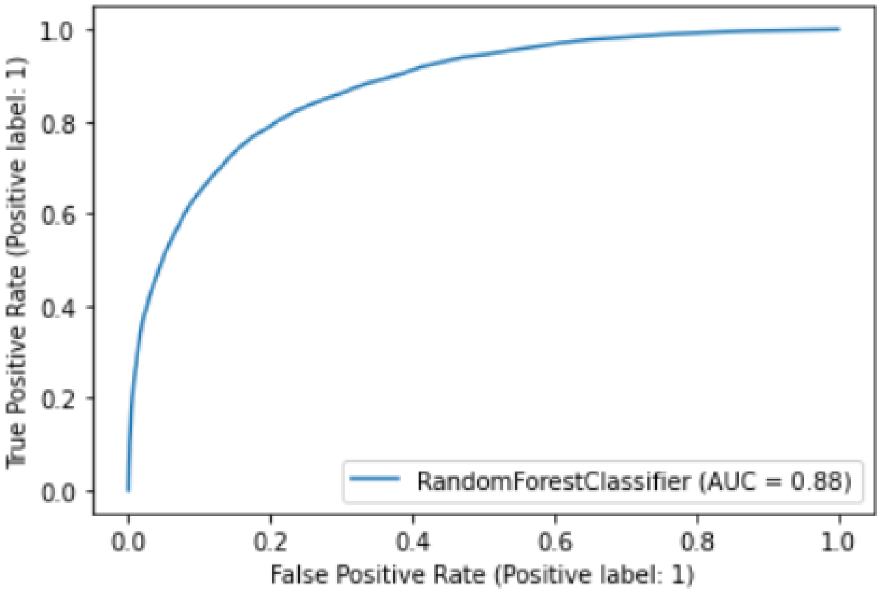
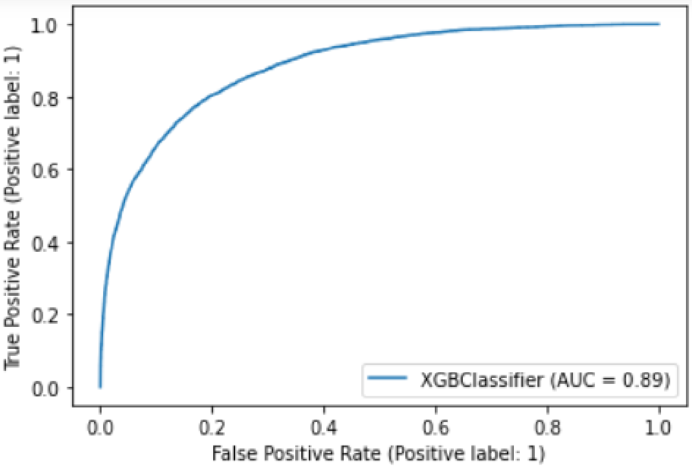


Figure 5.1: ROC CURVE(source code)

The above figure 5.1 is the receiver operating characteristic curve of random forest method.



The above figure 5.2 is the receiver operating characteristic curve of XG Boost Classifier.

Figure 5.2: ROC CURVE(source code)

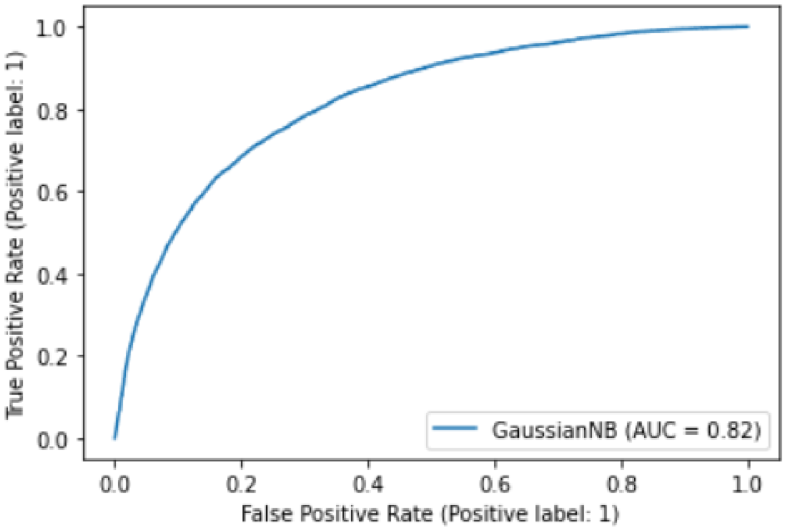


Figure 5.3: ROC CURVE(source code)

The above figure 5.3 is the receiver operating characteristic curve of Gaussian Naive Bayes.

Table 5.8: Confusion Matrix

Random Forest method	True Positive	False Negative
False positive	21754	972
True Negative	3318	3048

The above table 5.8 is about the confusion matrix of random forest method where on the top left is the true positive whose value is 21754 and the top right is false negative whose value 972. The bottom left and right is false positive and true negative respectively with value 3318 and 3048.

Table 5.9: Confusion Matrix

Gaussian Naive Bayes	True Positive	False Negative
False positive	19841	2885
True Negative	2720	3646

The above table 5.9 is about the confusion matrix of Gaussian naive bayes.

Table 5.10: Confusion Matrix

K Nearest Neighbors	True Positive	False Negative
False positive	20945	1781
True Negative	3214	3152

The above table 5.10 is the confusion matrix of K Nearest Neighbors

Table 5.11: Confusion Matrix

XG Boost	True Positive	False Negative
False positive	21543	1183
True Negative	2896	3470

The above table 5.11 is the confusion matrix of XG Boost classifier

CHAPTER 6

Conclusions

And here is the final chapter showing how clever you are ....