JAHANGIRNAGAR UNIVERSITY
Department of Statistics
SAVAR, BANGLADESH

RAINFALL PREDICTION: MACHINE LEARNING APPROACHES TO PREDICT RAINFALL LEVELS IN CHITTAGONG BASED ON METEOROLOGICAL FACTORS

Course Name: Machine Learning for Data Science
Course No: WM-ASDS22 (Section: B)
Masters in
Applied Statistics and Data Science
Spring 2023

by
Sk. Md. Rashid Abrar
ID: 20229048
Batch: 9th

Submitted to
Md. Habibur Rahman
Associate Professor, Department of Statistics
JAHANGIRNAGAR UNIVERSITY

ABSTRACT

This report focuses on the development and evaluation of various machine learning models to predict monthly rainfall levels in Chittagong, Bangladesh. The study utilizes atmospheric data collected over the period of 1964 to 2015, including variables such as temperature, dew point temperature, wind speed, humidity, and sea level pressure. The analysis of the relationship between these meteorological variables and rainfall amounts is the main goal of this study; doing so will help us better understand local weather patterns and improve the accuracy of our forecasts. We intend to generate models that can accurately forecast future rainfall levels based on the available meteorological data by utilizing various machine learning algorithms, including linear regression, decision tree classification, random forest classification and Artificial neural network (ANN). Through extensive experimentation and comparative analysis, the results of this study should clarify the extent to which each meteorological component influences precise rainfall forecasts or prediction, thereby assisting in the creation of more complex and accurate predictive models for weather forecasting in the Chittagong region. To conclude, this research aims to bridge the gap between meteorological variables and rainfall patterns by harnessing the power of machine learning.

Keywords: Rainfall prediction, machine learning, meteorological factors, temperature, dew point temperature, wind speed, humidity, sea level pressure, weather forecasting, predictive models, regression, decision tree, random forest, Artificial neural network (ANN)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This project explores the field of rainfall prediction using modern machine learning algorithms on a dataset gathered by the Chittagong data record station (22.35, 91.82) over the years 1964–2015. This dataset includes significant climatic factors like temperature, dew point temperature, wind speed, sea level pressure, humidity, and monthly rainfall totals.

1.1 Overview

A major objective in climate research is the precise prediction of rainfall patterns, which has significant consequences for a wide range of industries, including agriculture, as Bangladesh is highly dependent on agriculture, water resource management, disaster mitigation, and urban planning. This report embarks on a focused investigation into rainfall prediction, utilizing various machine learning algorithms. The data included important meteorological factors such as temperature, dew point temperature, wind speed, sea level pressure, humidity, and monthly rainfall totals, which help perform machine learning algorithms.

Machine learning has come to be a powerful tool for predicting rainfall levels due to its ability to identify deep relationships within dataset features. Notably, Parmar, Aakash, Mistree, Kinjal, and Sompura, Mithila (2017) showed in their review paper about machine learning approaches to predict or forecast rainfall in India by discussing the neural network approach. [1]. G. Geetha and R. S. Selvaraj (2011) showed how rainfall forecasting methods, such as statistical methods and Numerical Weather Prediction (NWP) models, are becoming saturated day by day, and neural network can interpret better forecasting or prediction in these cases. [2]. Olusola Samuel Ojo and Samuel Toluwalope Ogunjo (2022) worked on rainfall in different parts of Nigeria. They used artificial neural networks (ANN) to look at the effects of precipitation variability and predict rainfall amounts in Nigeria. [3].
These experiments demonstrate how machine learning could be used to understand the complex relationships that control rainfall variations and inspire us to work on this.

The Chittagong data record station has carefully collected a comprehensive collection of atmospheric data, such as temperature, dew point temperature, wind speed, humidity, sea level pressure, and monthly rainfall totals. This collection provides a one-of-a-kind platform for examining and performing several machine-learning algorithms to achieve our goal.

1.2 Stating the Motivation and objectives of this project

We now state our main motivation and object of this project.

Motivations:

1. **Climate Adaptation:** Predicting rainfall accurately helps in the development of solutions for coping with climate change, which is especially important in areas like Chittagong, which is the second-most predominant city in Bangladesh.
2. **Disaster Mitigation and Management:** Heavy precipitation can cause flooding. Forecasts that are accurate and on time help disaster management organisations respond to floods and other weather-related events.
3. **Agricultural Planning:** Bangladesh, which is a land of greenery, mostly depends on agriculture. So, well-prepared planning is one of the most important things to do to keep ourselves in a nifty position to deal with the agricultural revolution. Based on accurate rainfall predictions, farmers may optimize planting and irrigation schedules, reducing crop losses.
4. **Water Resource Management:** Accurate predictions for rainfall can be used optimally for water reservoirs and infrastructure development, improving water resource management.
5. **Environmental Sustainability:** By taking into account the need for water drainage and conservation, improved forecasts aid in sustainable urban design.

Objectives:

1. Investigate the relationships between meteorological variables (temperature, dew point temperature, wind speed, humidity, and sea level pressure) and rainfall levels in Chittagong using the data visualisation technique.
2. Develop predictive models using machine learning algorithms to accurately predict monthly rainfall levels.
3. Compare the performance of these machine learning approaches (linear regression, decision trees, random forests, and neural networks).
4. Evaluate the predictive accuracy of each model using appropriate evaluation metrics.
5. Provide insights into the most effective machine learning approach for rainfall prediction in the Chittagong region.

We hope to contribute to the advancement of rainfall prediction approaches by boosting prediction accuracy and enabling more informed decision-making in response to climatic changes.

Chapter 2

Methodology

In this methodology chapter, we will discuss how the machine learning approaches will work and how they will predict the precipitation level in Chittagong. In order to delve deep into the methods of these machine learning algorithms, we first need to know about the variables of the data. Also, data pre-processing is a vital step in the methodology in order to perform these algorithms. After that, machine learning estimation will be shown, which is a vital part of this chapter.

2.1 Data Description

The Dataset Contains 11 variables, although we only discuss and work on six of them: Temperature, Dew Point Temperature, Wind Speed, humidity, sea level pressure, and rainfall category. Each variable has a distinct meaning. Here, we have shown a table that represents the variable description, level of measurement, and suitable measures for all the variables.

Variable name	Variable Description	Level of measurement	Appropriate measures
TEM	Temperature	Interval Level	Mean
DPT	Dew Point Temperature	Interval Level	Mean
WIS	Wind Speed	Ratio level	Mean
HUM	Humidity	Nominal level	Mean
SLP	Sea Level Pressure	Interval Level	Mean
RAN	Rainfall CATEGORY	Nominal	Mode

Table 2.1: Variable's summary information of the Selected Dataset

2.2 Data Preprocessing

First of all, before going into visualizing and interpreting data, there are some pre-processing steps that need to be done.

1. From the below figure, it can be seen that there are 6 variables and 624 observations in this dataset, and there are no missing values. All of the data types are of the float type except for one, which is nominal variable and comes with the object type. To find any missing values, the pandas info function is handy in this place.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 624 entries, 0 to 623
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   TEM     624 non-null    float64
 1   DPT     624 non-null    float64
 2   WIS     624 non-null    float64
 3   HUM     624 non-null    float64
 4   SLP     624 non-null    float64
 5   RAN     624 non-null    object
dtypes: float64(5), object(1)
memory usage: 29.4+ KB
```

FIGURE 2.1: Missing values checking

2. As there are no missing values, in order to find outliers, boxplot is a well-known graphical visualization technique that needs to be addressed and implemented.

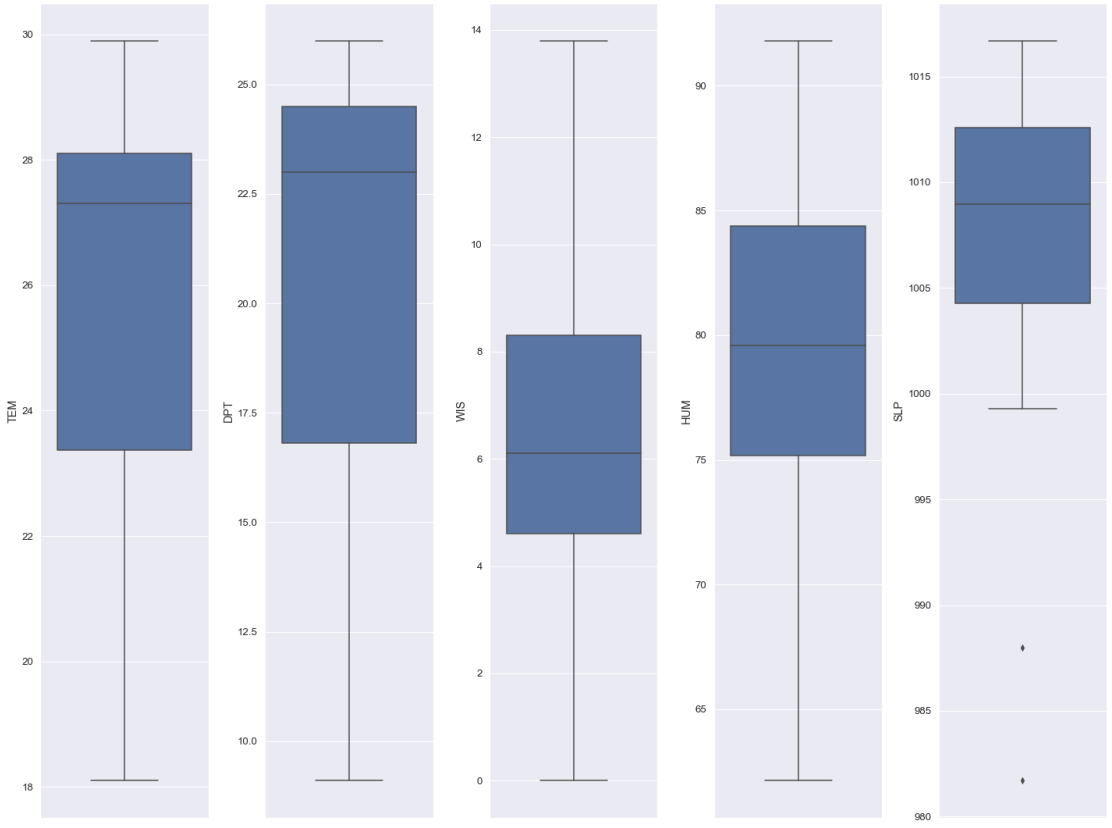

FIGURE 2.2: Boxplot

From the above figure, boxplots are shown for all five variables that represent quantity, and clearly RAN does not have a boxplot as it is a nominal value. Now it is time to try and interpret some of the boxplots:

- For the TEM variable, which stands for temperature, it can be seen that the overall distribution is not symmetric; instead, the boxplot can be inferred as having a bit of a negatively skewed distribution. The minimum value is close to 18.2, the maximum value is approximately 29.8, and the first quartile is close to 23.5. The middle observations that lie within the first and third quartiles have a negatively skewed distribution. From this boxplot, it is also apparent that high concentrations of observations lie in the range of 23.5 to approximately 26.5. Also, the mean lies approximately at 27.6. The most important finding from this boxplot is that there are no outliers here, which is good for our model estimation.
- Now for the DPT variable, which stands for Dew Point Temperature, it can be seen that the overall distribution is not symmetric; instead, this boxplot can also be inferred as having a bit of a negatively skewed distribution. The minimum value is close to 5, the maximum value is above 25, and the first quartile is close to 17. The middle observations that lie within the first and third quartiles have a negatively skewed distribution. From this boxplot, it is also apparent that high concentrations of observations lie in the range of 17 to approximately 24.5. Also, the mean lies approximately at 23. The most significant finding from this boxplot is that there are no outliers here, which is good for model estimation.
- For the WIS variable, which stands for Wind Speed, it can be seen that the overall distribution is mostly symmetric. The minimum value is 0, the maximum value is approximately 13.8, and the first quartile is close to 4.5. The middle observations that lie within the first and third quartiles have a bit of a negatively skewed distribution. From this boxplot, it is also apparent that high concentrations of observations lie in the range of 4 to approximately 8.2. Also, the mean lies approximately at 6.2. The most significant finding from this boxplot is that there are no outliers here, which is good for model estimation.
- Now for the HUM variable, which stands for Humidity, it can be seen that the overall distribution is not symmetric; instead, this boxplot can also be inferred as having a bit of a negatively skewed distribution. The minimum value is close to 65, the maximum value is above 90, and the first quartile is close to 75.1. The middle observations that lie within the first and third quartiles have a negatively skewed distribution. From this boxplot, it is also apparent that high concentrations of observations lie in the range of 75.1 to approximately 84.7. Also, the mean lies approximately at 79.2, and the mean is somewhat symmetric. The most significant finding from this boxplot is that there are no outliers here, which is good for model estimation.
- For the SLP variable, which stands for Sea Level Pressure, we can see that the overall distribution is symmetric. The middle observations that lie within the first and third quartiles have a positively skewed distribution, and the minimum value is close to 999.8. From this boxplot, it is also noticeable that high concentrations of observations lie in the range of approximately 1004.275 to 1012.6. The most significant finding

from this box plot is that there are two outliers. Outliers can be found at either the upper or lower extreme, and for this plot, the outliers lie within the low extreme. So, to calculate outliers, an interquartile range is needed, which for this variable is approximately Q3-Q1 = 1004.275-1012.6 = 8.325. The outlier formula for this variable is Q1-1.5 * IQR = 991.78749. So, the values that are lower than 991.78749 are seen as the dotted ones in this boxplot as outliers. As there are only two values, we can simply remove them by taking all the values that are bigger than the lowest value and using these values for model estimation. We present SLP after removing the outliers in the below figure.

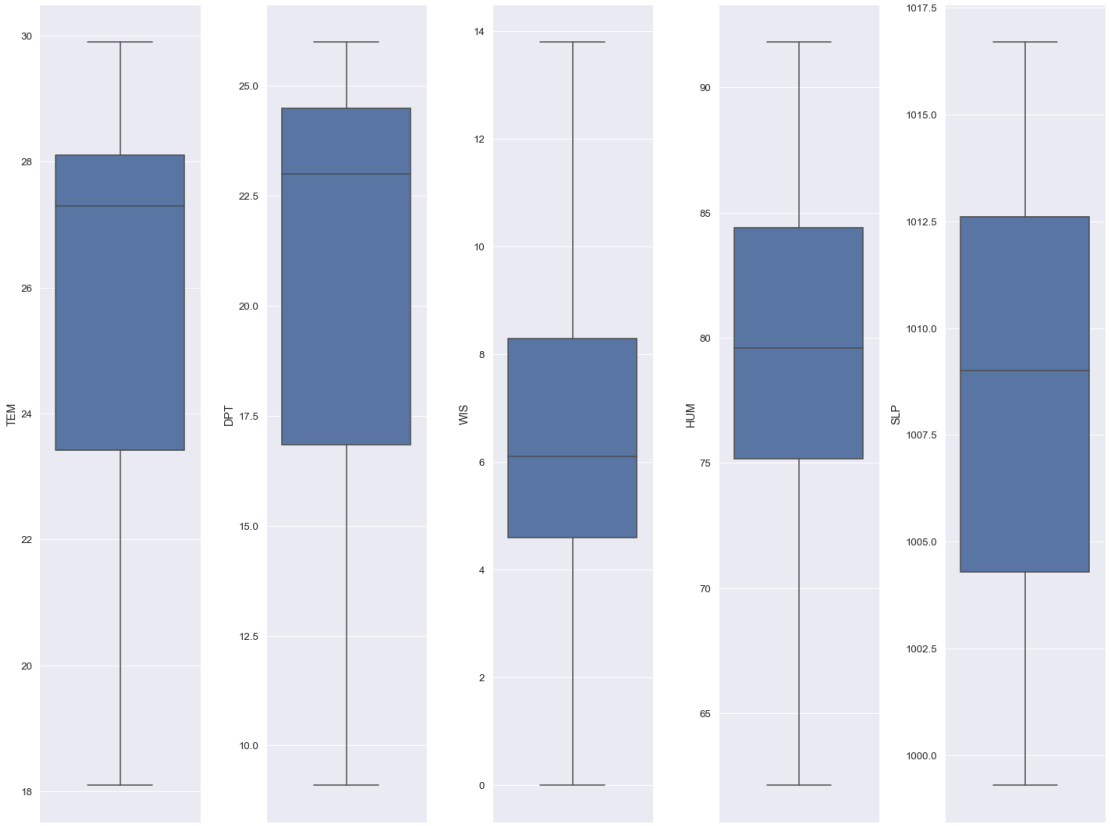

FIGURE 2.3: Boxplot after removing outliers

2.3 Model Estimation

After Removing outliers, we can now go on and apply several machine learning approaches. We work on four algorithms, including Logistic Regression, Decision trees, Random forests, and artificial neural networks (ANN).

First of all, the target variable, which is RAN in this case, and all other variables need to be separated. Then training and testing data need to be split in order to get a good model estimation. In this case, 30% of the data is saved for testing, while

70% is used for training. We will also use normalization for normal distribution after training and testing.

- **Logistic Regression:** Before estimating, it is important to set the best parameters for model estimation. So, in order to set these parameters, GridSearchCV is used. It is a method for determining the best parameter values from a given set of parameters in a grid. For each algorithm, this process is done, and the below code snippets show it for the logistic regression machine learning algorithm. Here the optimal parameters are C = 10.0, penalty = 'l1', and solver = 'saga'. After getting the optimal parameters, these are used in invoking the logistic regression function, which is provided by the Scikit-Learn library. Then, the model can be fitted using the fit method by providing a training dataset.
- **Decision Tree:** It is important to set the best parameters for model estimation when estimating a Decision Tree. So, in order to set these parameters and the below code snippets that show them for the Decision Tree Classifier machine learning algorithm, Here the optimal parameters are criterion = 'entropy', max_depth = 5, min_samples_leaf = 3, min_samples_split = 5. After getting the optimal parameters, these are used in invoking the DecisionTreeClassifier function, which is provided by the Scikit-Learn library. Then, the model can be fitted using the fit method by providing a training dataset.
- **Random Forest:** For the Random Forest Classifier machine learning algorithm, GridSerchCV is used to set the best parameters. Here the optimal parameters are criterion = 'entropy', n_estimators=29, max_depth = 8, min_samples_split= 5, min_samples_leaf= 1, and the below code snippets are given to show how these values are acquired.
- **Artificial Neural Networks (ANN):** We build an ANN model with eleven hidden layers to capture complex relationships in the data. Hidden layers are layers between the input and output layers of an ANN. They play a crucial role in capturing complex relationships and patterns in the data. In this case, the ANN model has eleven hidden layers, with varying numbers of neurons in each layer: 512, 256, 128, 128, 64, 64, 32, 32, 16, 16, 8, and 8 neurons. Early stopping is used to avoid overfitting, while accuracy and loss measurements are used to keep track of model performance.

Chapter 2 describes the approach that is required for this project, as well as some pre-processing and data descriptions for a better understanding of the methodology.

Chapter 3

Result and Discussion

This chapter showcases the results and discussion coming out of this project. As mentioned previously, we used four machine learning approaches, and for that, we needed to visualize and understand the data first. So, first we will discuss Exploratory Data Analysis (EDA), and then we will show the results and discuss each model's performance, which contain accuracy, precision, recall, and F1-score.

3.1 Exploratory Data Analysis (EDA)

3.1.1 Histogram

The histogram helps to depict the graphical representation of feature distribution. This distribution can be symmetric or asymmetric. The below figure shows the procedure to get the histogram of five variables from this Dataset. To understand them, first look at the plot that explains variable TEM. This plot indicates that this is a somewhat negatively skewed distribution. For another variable, DPT, it can be said that the distribution is also negatively skewed. WIS and HUM showcase a somewhat symmetric distribution. SLP can have a bit of an irregular distribution in terms of symmetry.

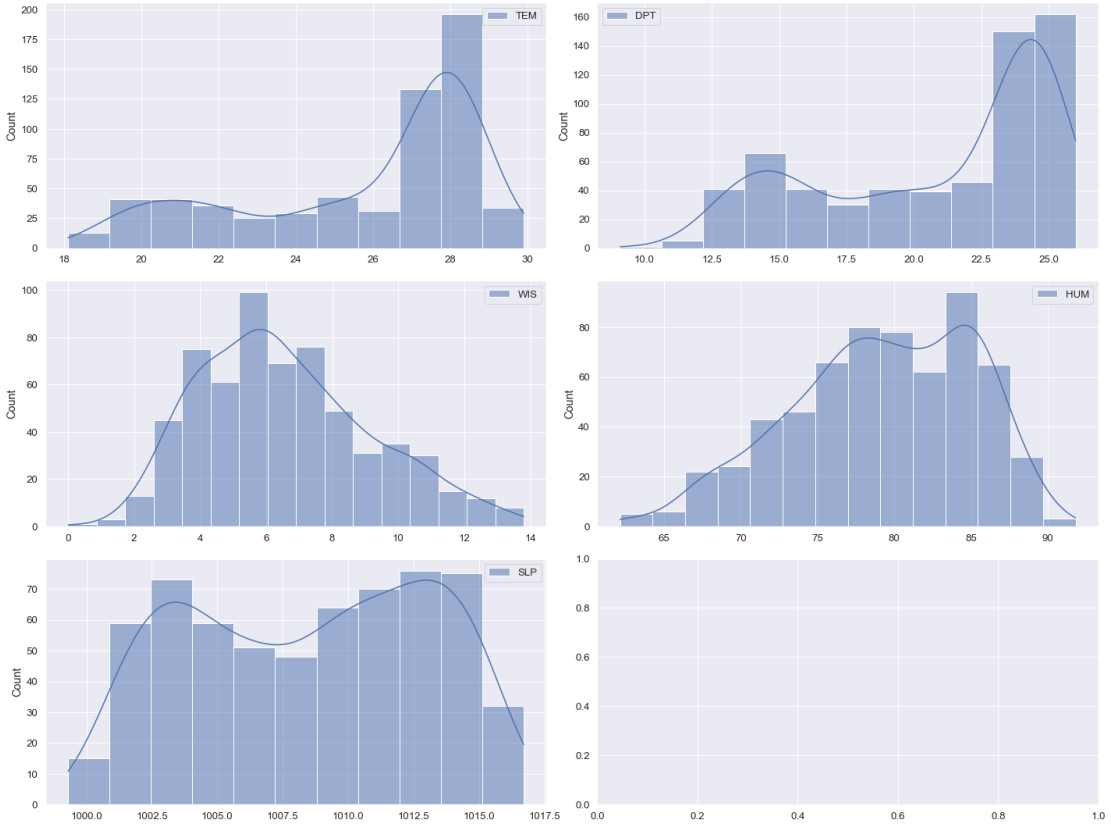

FIGURE 3.1: Histogram

3.1.2 Density plot

Density plots are particularly useful when dealing with the density of univariate data, where you want to understand the shape of the distribution and identify patterns or trends. It also shows how symmetric the variables are in a smooth manner.

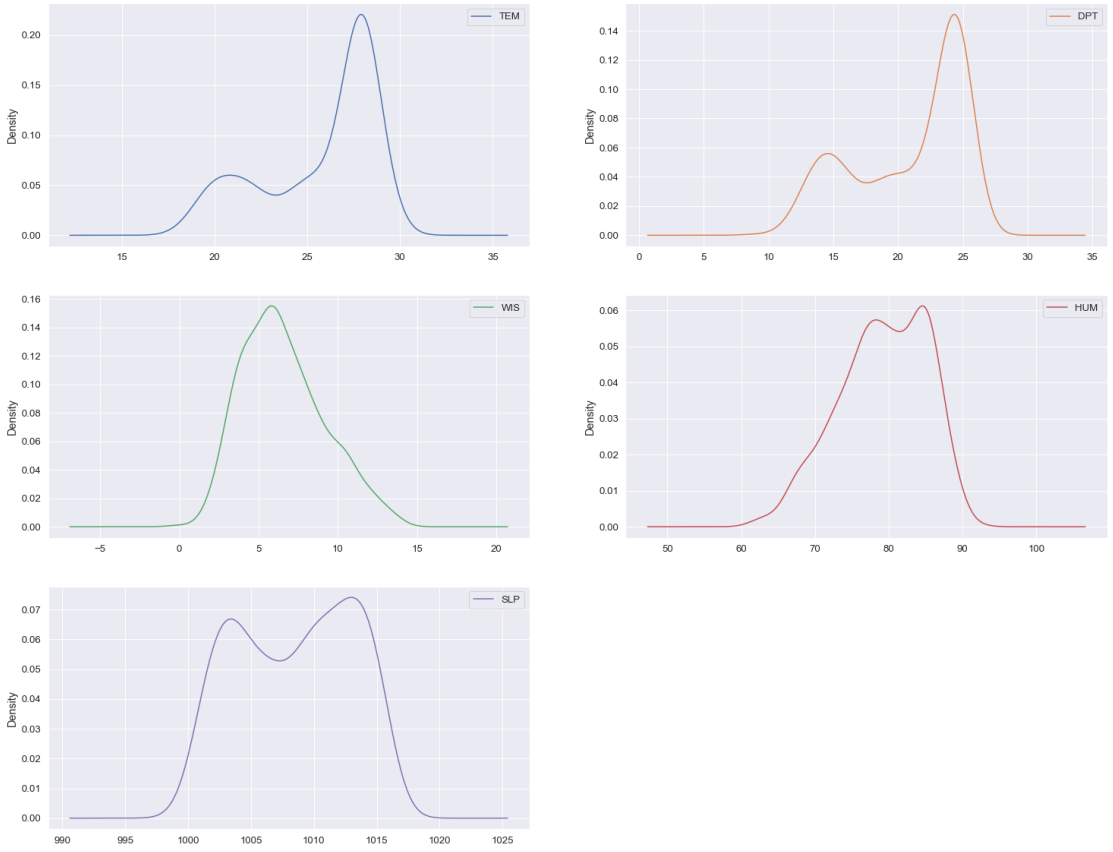

FIGURE 3.2: Density plot

3.1.3 Scatterplot

A scatterplot shows the relationship between a pair of values. The pair of values in a graph can be called a scatter diagram. When the dotted points are pointing upward, the relationship between the variables is positive. On the contrary, when the dotted points are downward, the relationship between the variables is negative. If neither happens, then it might be assumed that the variables are not correlated.

From the below scatter plots, it can be attested that variables TEM and DPT have a strong positive correlation, TEM and WIS have a weak positive correlation, and TEM and HUM have a moderate positive correlation. Conversely, TEM and SLP have a strong negative correlation. Applying this same idea, all other variables correlations or relationships can be interpreted using a scatter plot.

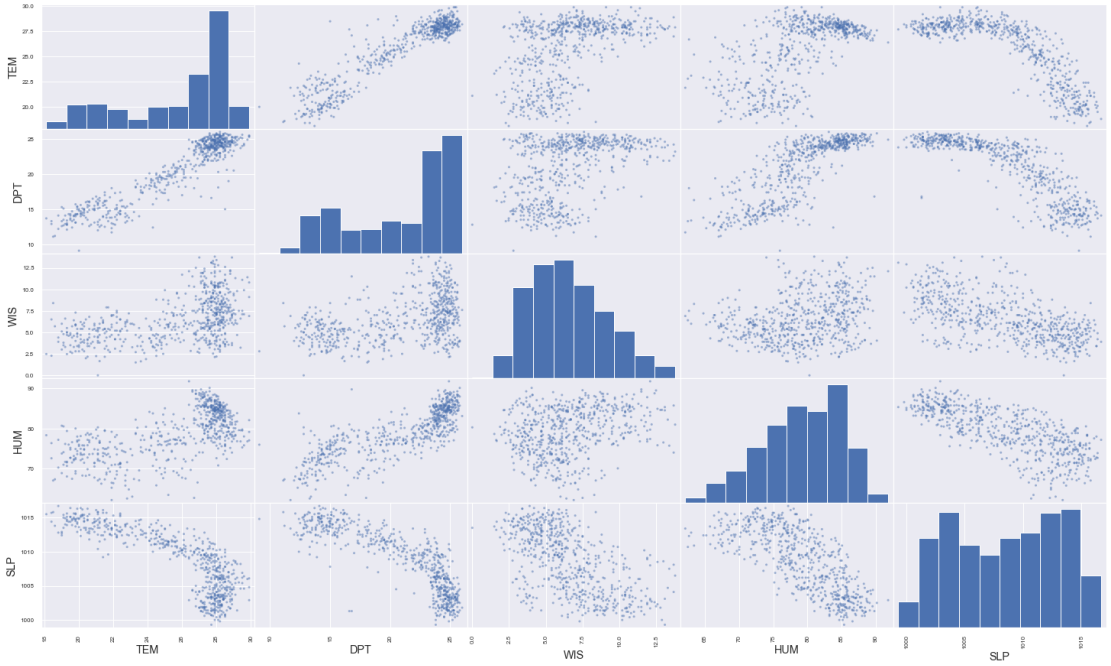

FIGURE 3.3: Scatter plot

3.1.4 Bar Diagram

A bar diagram, is mostly used to describe the frequency of the categorized data. The below plot shows that the Rainfall category observations for Light Rain are 223 units, Moderate And High Rain 195 units, and No Rain And Trace are 204 units.

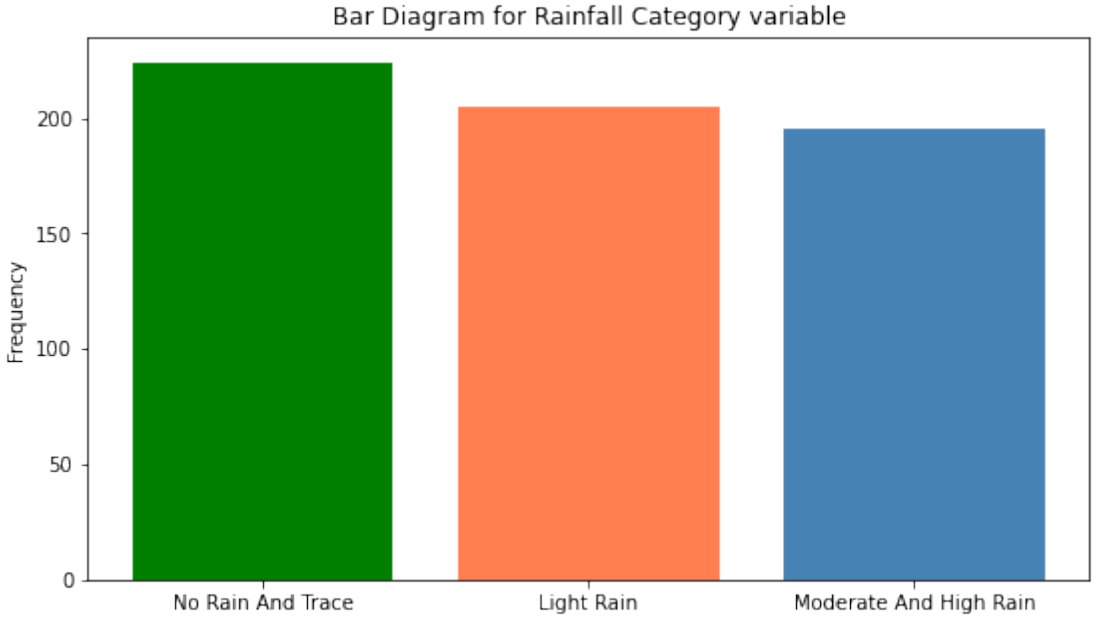

FIGURE 3.4: Bar diagram

3.1.5 Pie Chart

A pie chart also works on categorized data, although it tends to show the percentage of the categories. The pie chart below shows that 35.9 percent shows No rain and Traces, 31.25 percent shows Moderate and High Rain, and 32.85 percent shows Light Rain in the category given.

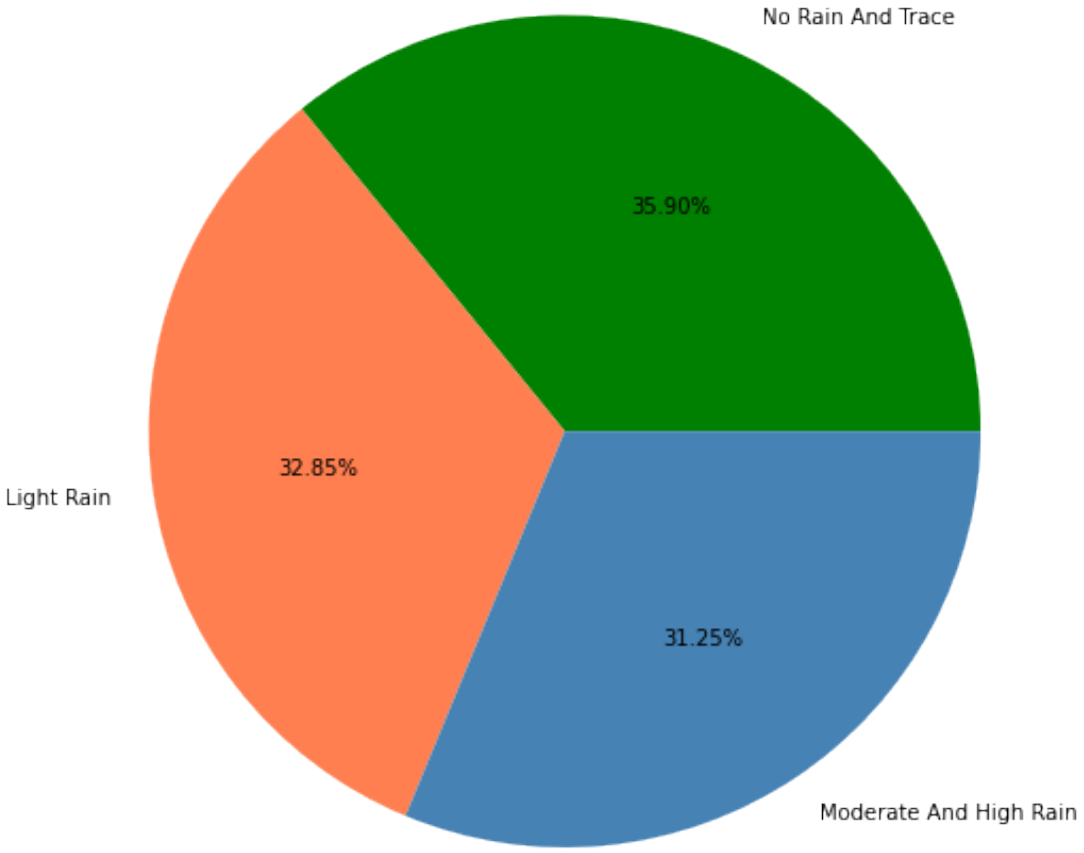

FIGURE 3.5: Pie Chart

3.1.6 Pair Plot

This pairplot is based on two fundamental figures: the density plot and the scatter plot. The diagonal density plot shows the distribution of a single variable, but the upper and lower triangular scatter plots indicate the relationship between two variables. The above pairplot diagram can be helpful in understanding how pairplot works. Here, the blue dots refer to No Rain Traces, The pink dot represents Light Rain, and the orange dots refer to Moderate and High Rain. If we look at the TEM variable, then we can interpret that if the temperature is close to 30 degree

Celsius, then Moderate and High Rainfall is likely to happen. If the temperature is approximately 23 degree Celsius, then Light Rain is likely to result, and if the temperature is approximately 21, then No Rain or Trace is likely to happen. For DPT, we can interpret that if the temperature is higher than 25 degrees Celsius, then Moderate and High Rain is likely to happen. If the temperature is approximately 17 degree Celsius, then Light Rain is the likely result that occurs, and if the temperature is approximately less than 14, then No Rain or Trace is likely to happen.

Likewise, other density plots can also be interpreted from the pairplot. Also, the scatter plot portion is already interpreted and described in the scatter plot section.

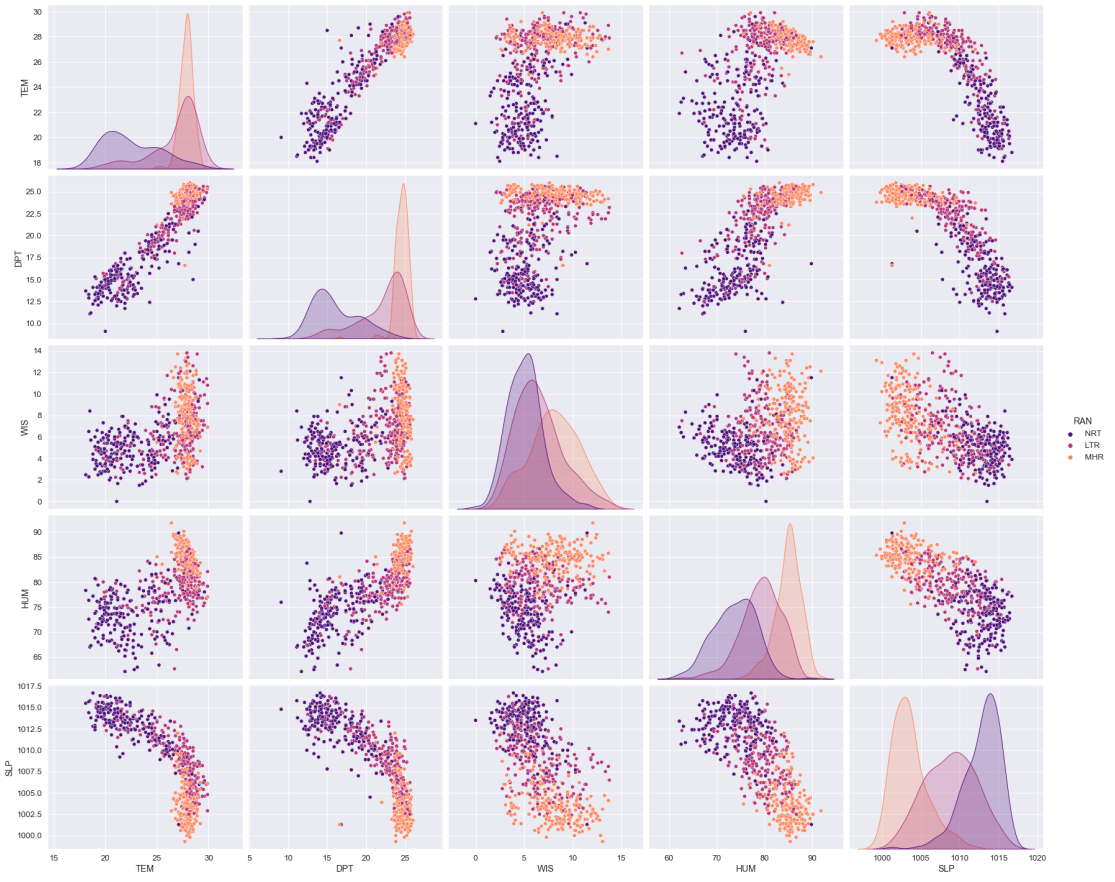

FIGURE 3.6: Pair Plot

3.1.7 Correlation Matrix

The correlation matrix, which calculates the correlation coefficients between variables, is also being calculated. Also, spearman rank correlation is used as a method in the hope of handling outliers. The figure shows the correlation matrix between the variables. The Seaborn Library is used to showcase the matrix in an organized manner.

The values above or below the diagonal 1's are considered in this correlation matrix. These values indicate how strong the correlation is between variables. To

understand correlation, consider the correlation between DPT and TEM, which is 0.84, implying that these two variables have a strong positive correlation. The correlation between DPT and WIS is 0.42, indicating that these two variables have a moderately positive correlation. Furthermore, the correlation between DPT and SLP is -0.87, indicating that these two variables have a strong negative correlation. It can also be found that the correlation between DPT and HUM is 0.83, indicating that these two variables have a strong positive correlation. These are some variables that are discussed, and other variables that are not discussed also work the same way.

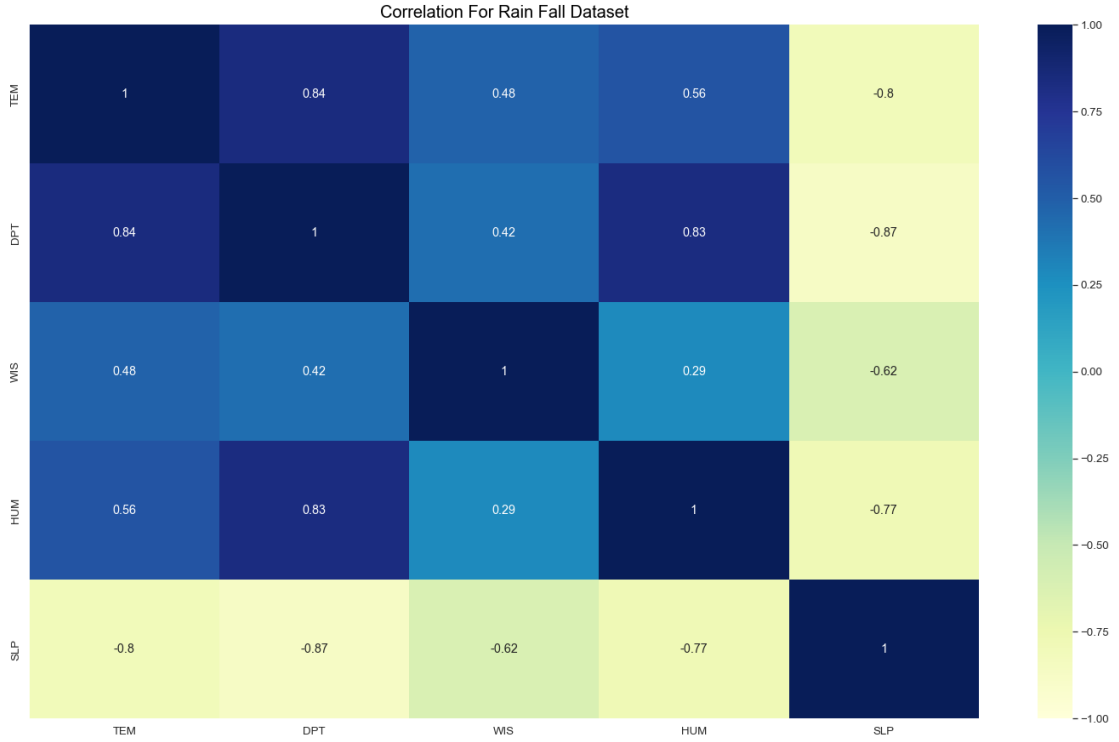

FIGURE 3.7: Correlation matrix

3.2 Model performances

Now that we have discussed Exploratory Data Analysis (EDA), we will move on and discuss the results of the models that are implemented in this project, and further discussion will happen on this model's ROC curve, Tree plot, and accuracy and loss plot for the ANN, which also determine the model's performances.

1. **Logistic Regression:**

   - The logistic regression model achieved an accuracy of 0.777 on the train set and an accuracy of 0.7647 on the test set, which means 76% predictions are correct. The precision for LTR is 0.73, that is, 73% positive predictions are correct for LTR. Similarly, MHR's precision is 0.77, and NRT's precision is 0.80. For recall, LTR, MHR, and NRT are respectively 0.63, 0.89, and 0.81, which means 63% positive cases are predicted as positive for LTR, 89% positive cases are predicted as positive for MHR, and 80% positive cases are predicted as positive for NRT, and the F1-scores for each class are 0.68, 0.82, and 0.80, respectively.
   - Now we will present the ROC curve for this model. In order to understand if a ROC is good or bad, we need to know if the true positive rate, or sensitivity, will increase, and if the area under the curve (AUC) is close to 1, the model is performing well. So, from the below curve, it can be seen that the model is performing well given the earlier statement stated above, and the ROC curve score is 0.83, 0.96, and 0.93 for LTR, MHR and NRT respectively.

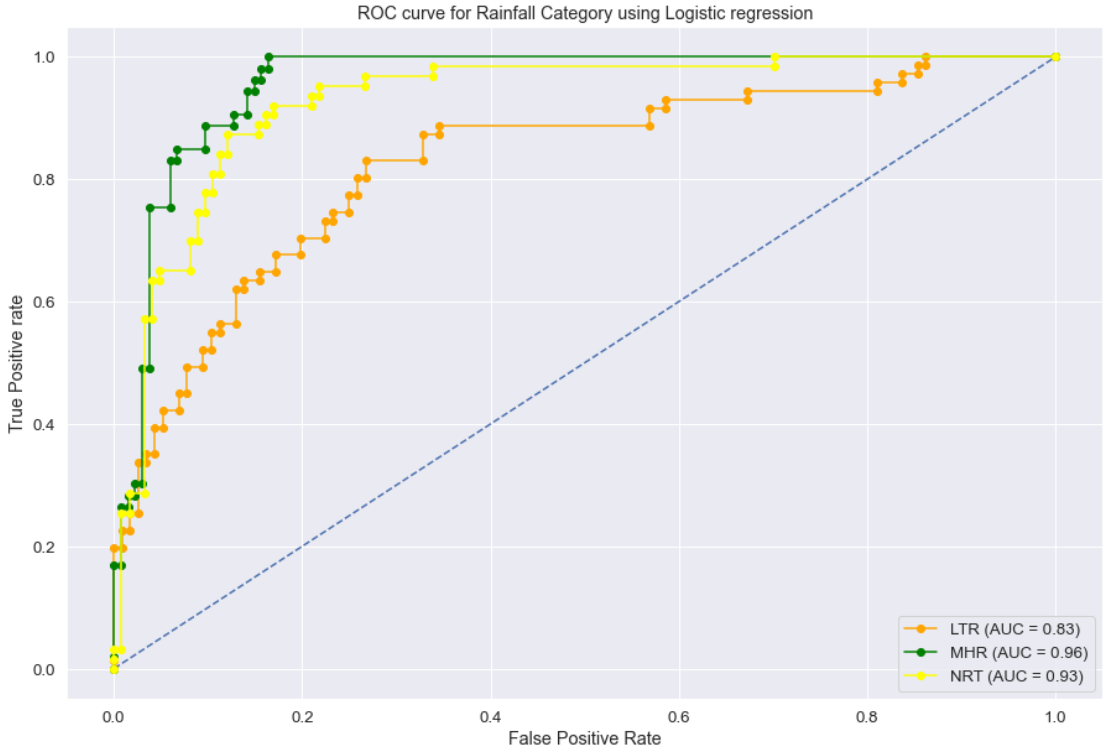

FIGURE 3.8: ROC Curve for Logistic Regression

2. **Decision Tree:**

   - The Decision Tree model achieved an accuracy of 0.8390 on the train set and an accuracy of 0.7166 on the test set, which means 71% predictions are correct. The precision for LTR is 0.67, that is, 67% positive predictions are correct for LTR. Similarly, MHR's precision is 0.71, and NRT's precision is 0.77. For recall, LTR, MHR, and NRT are respectively 0.51,

0.91, and 0.79, which means 51% positive cases are predicted as positive for LTR, 91% positive cases are predicted as positive for MHR, and 79% positive cases are predicted as positive for NRT, and the F1-scores for each class are 0.58, 0.79, and 0.78, respectively.

- Now we will present the ROC curve for this model. In order to understand if a ROC is good or bad, we need to know if the true positive rate, or sensitivity, will increase, and if the area under the curve (AUC) is close to 1, the model is performing well. So, from the below curve, it can be seen that the model is performing well given the earlier statement stated above, and the ROC curve score is 0.78, 0.95, and 0.90 for LTR, MHR and NRT respectively.

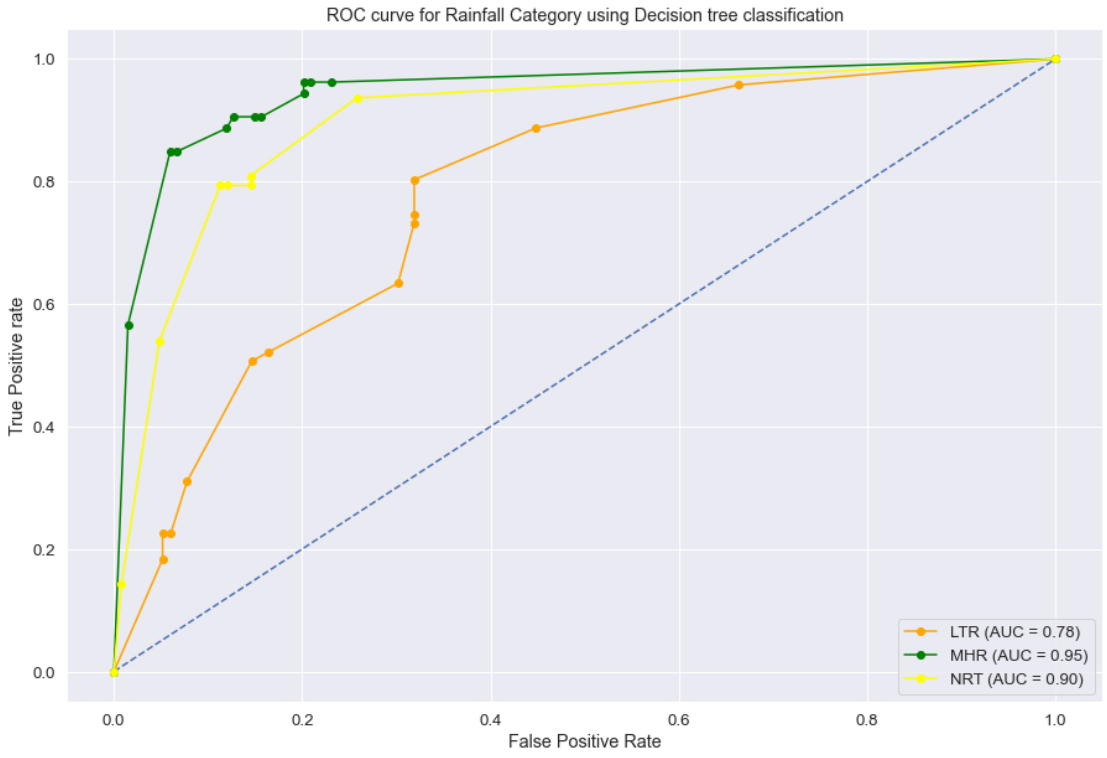

FIGURE 3.9: ROC Curve for Decision tree

- To perform a decision tree, we need to understand what hyperparameters and splitting values are needed for that. In Chapter 2, we mentioned the parameter values. Here we will show the results after splitting and how the tree looks after building the model. The below plot shows that variable DPT is the decision node for this mode, which satisfies entropy criteria, and it is measured that information gain is higher in this decision node.

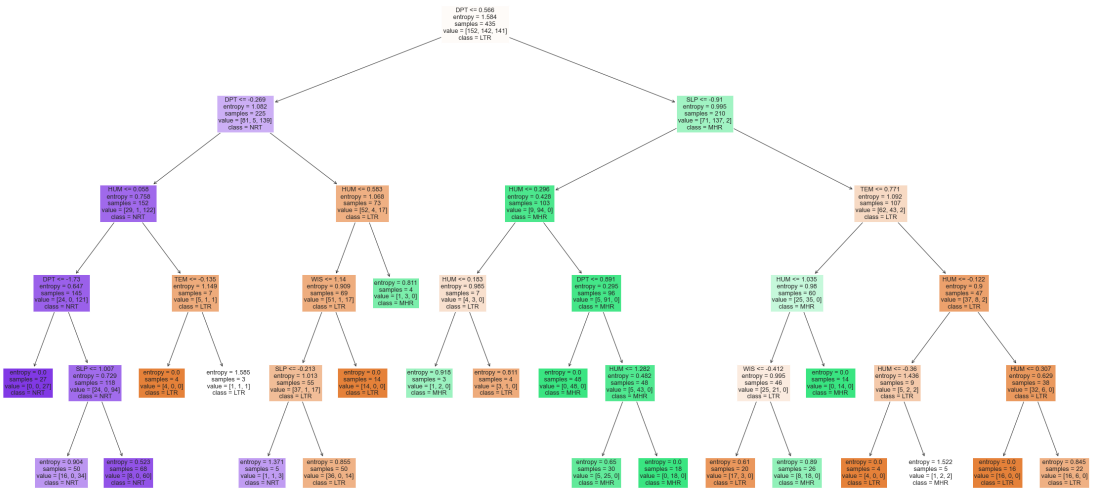

FIGURE 3.10: Decision tree

3. **Random Forest:**

   - The Random Forest model achieved an accuracy of 0.9494 on the train set and an accuracy of 0.77005 on the test set, which means 77% predictions are correct. The precision for LTR is 0.73, that is, 73% positive predictions are correct for LTR. Similarly, MHR's precision is 0.78, and NRT's precision is 0.80. For recall, LTR, MHR, and NRT are respectively 0.65, 0.89, and 0.81, which means 65% positive cases are predicted as positive for LTR, 89% positive cases are predicted as positive for MHR, and 81% positive cases are predicted as positive for NRT, and the F1-scores for each class are 0.69, 0.83, and 0.80, respectively.
   - Now we will present the ROC curve for this model. In order to understand if a ROC is good or bad, we need to know if the true positive rate, or sensitivity, will increase, and if the area under the curve (AUC) is close to 1, the model is performing well. So, from the below curve, it can be seen that the model is performing well given the earlier statement stated above, and the ROC curve score is 0.81, 0.96, and 0.92 for LTR, MHR and NRT respectively.

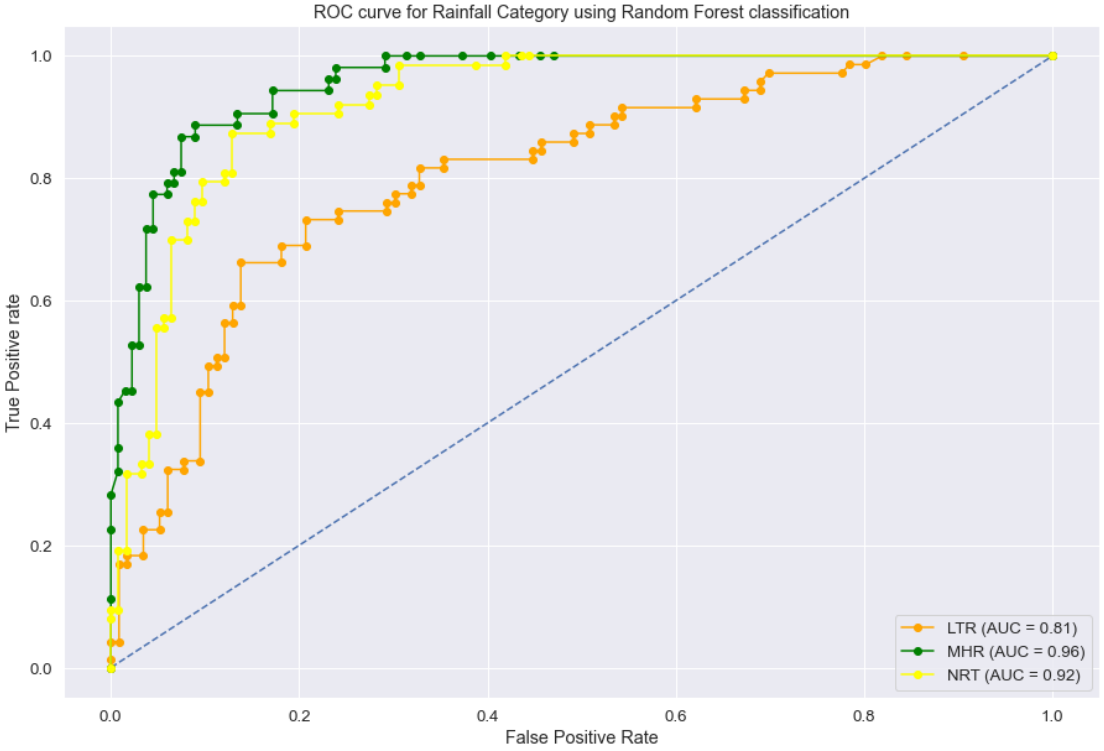

FIGURE 3.11: ROC Curve for Random forest

- Random forest is an ensemble technique, and it relies on combining different classifiers using some aggregation technique. As it is not possible to show all the trees using plot. Here, only the first tree is shown. From the tree, it can be interpreted that HUM works as the decision node, which satisfies entropy criteria, and it is measured that information gain is higher in this decision node. In random forest classification, the contribution of variables matters most. This plot helps to interpret just the one tree of a random sample to get some idea of estimating the model on the entire dataset.

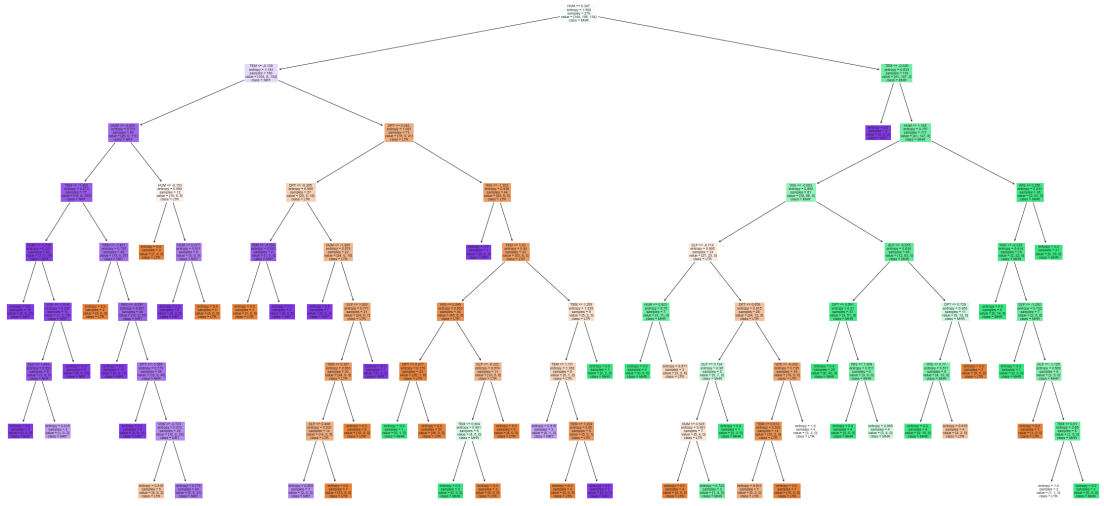

FIGURE 3.12: Random Forest Tree

4. **Artificial Neural Networks (ANN):**

   - Artificial Neural Networks achieved an accuracy of 0.8023 on the train set and an accuracy of 0.7807 on the test set, which means 78% predictions are correct. The precision for LTR is 0.78, that is, 78% positive predictions are correct for LTR. Similarly, MHR's precision is 0.80, and NRT's precision is 0.77. For recall, LTR, MHR, and NRT are respectively 0.59, 0.91, and 0.86, which means 59% positive cases are predicted as positive for LTR, 91% positive cases are predicted as positive for MHR, and 86% positive cases are predicted as positive for NRT, and the F1-scores for each class are 0.67, 0.85, and 0.81, respectively.
   - Now we will present the ROC curve for this model. In order to understand if a ROC is good or bad, we need to know if the true positive rate, or sensitivity, will increase, and if the area under the curve (AUC) is close to 1, the model is performing well. So, from the below curve, it can be seen that the model is performing well given the earlier statement stated above, and the ROC curve score is 0.84, 0.96, and 0.92 for LTR, MHR and NRT respectively.

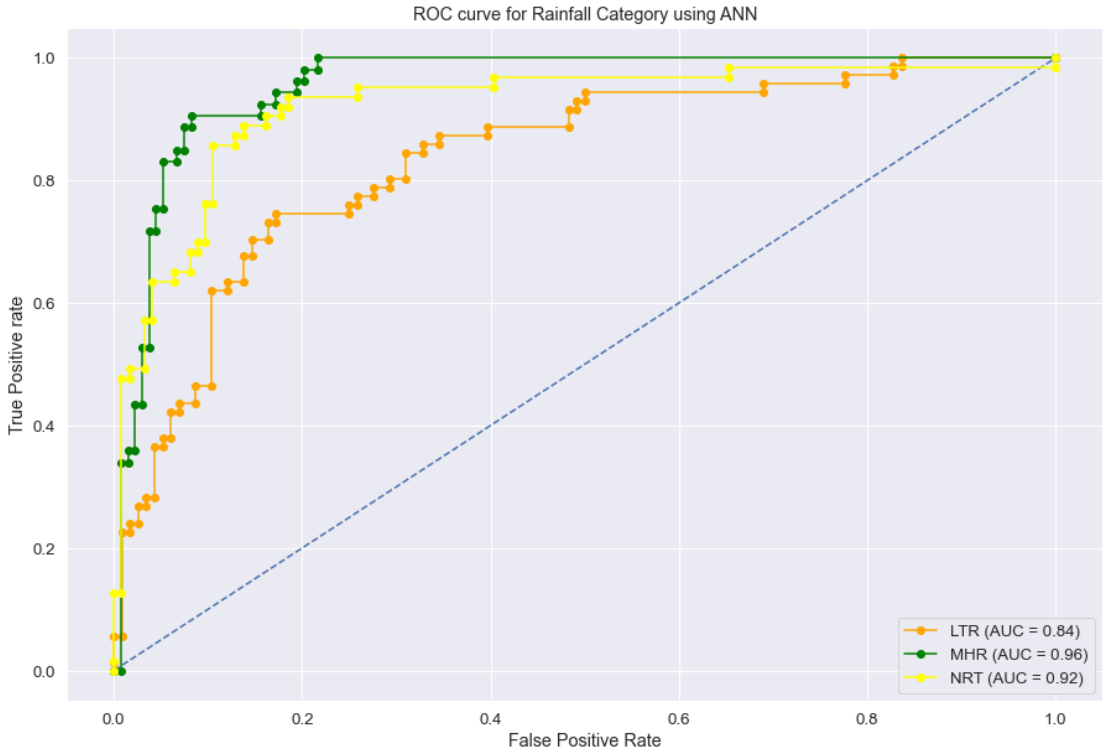

FIGURE 3.13: ROC Curve for ANN

   - The accuracy graph shows a visual representation of how the performance of an Artificial Neural Network (ANN) model evolves during the training process. It plots the model's accuracy on the training data and validation data over each epoch (training iteration). The below graph presents us with a blue and an orange line that start their accuracy measurements at 0.3 for the blue line and 0.45 for the orange line. After iterating the model for approximately 50 epochs, the blue line stops at

0.9 and the orange line stops somewhat at 0.75. This blue line actually shows the training accuracy measurements, and the orange line shows the testing accuracy measurements.

FIGURE 3.14: Accuracy Graph for ANN

- The loss graph plots the model's loss on the training data and validation data over each epoch (training iteration). The below graph presents us with a blue and an orange line that start their loss measurements at 1.5 for the blue line and approximately 0.88 for the orange line. After iterating the model for approximately 50 epochs, the blue line stops at 0.45 and the orange line stops somewhat at 0.82. This blue line actually shows the training accuracy measurements, and the orange line shows the testing accuracy measurements.

FIGURE 3.15: Loss Graph for ANN

Chapter 3 shows model results and discusses the performances through evaluation matrices and ROC curves.

Chapter 4

Summary and Conclusion

In this study, we explored and compared the performance of several machine learning algorithms for predicting rainfall levels based on meteorological factors. The model's results and discussions indicate that, among the four involved models, Artificial Neural Networks (ANN) perform the best. We can attest to that from the accuracy score and, more importantly, from the evaluation matrices. The evaluation matrix consists of precision, recall, and the F-1 score. The F-1 score combines two important aspects of classification, precision and recall, into a single value that represents a balance between them. So, Understanding the scores can help us determine which model performs the best. As the dataset is balanced, we will compare the macro-averages of each model. From the classification report, we know that the macro-average of Logistic regression is 0.77, Decision tree is 0.72, Random forest is 0.77, and Artificial Neural Networks (ANN) are 0.78. So, The model with the highest macro-average F1-score (ANN) might be preferred as the best-fit model.

These models can help predict accurate rainfall levels, contributing to better decision-making in various sectors.

Finally, this project demonstrates the significance of machine learning algorithms in predicting rainfall levels and gives useful insights for future weather forecasting and climate research.

References

[1] Aakash Parmar, Kinjal Mistree, and Mithila Sompura. Machine learning techniques for rainfall prediction: A review. *International Conference on Innovations in information Embedded and Communication Systems (ICIIECS)*, 03 2017.

[2] G. Geetha and R. S. Selvaraj. Prediction of monthly rainfall in chennai using back propagation neural network model. *International Journal of Engineering Science and Technology (IJEST*, 01 2011.

[3] Olusola Samuel Ojo and Samuel Toluwalope Ogunjo. Machine learning models for prediction of rainfall over nigeria. *Scientific African*, 06 2022.