Report

Fidan Nabiyeva

December 15, 2023

1. **Abstract**

This report focuses on investigating data pertaining to rainfall in Australia. It aims to illustrate distinctions between traditional, ensemble and neural network-based classification algorithms in the context of predicting rainfall patterns.

1. **Introduction**

Understanding and predicting rainfall patterns is a critical aspect of meteorological research, particularly in regions such as Australia, where climatic conditions can exhibit significant variability.

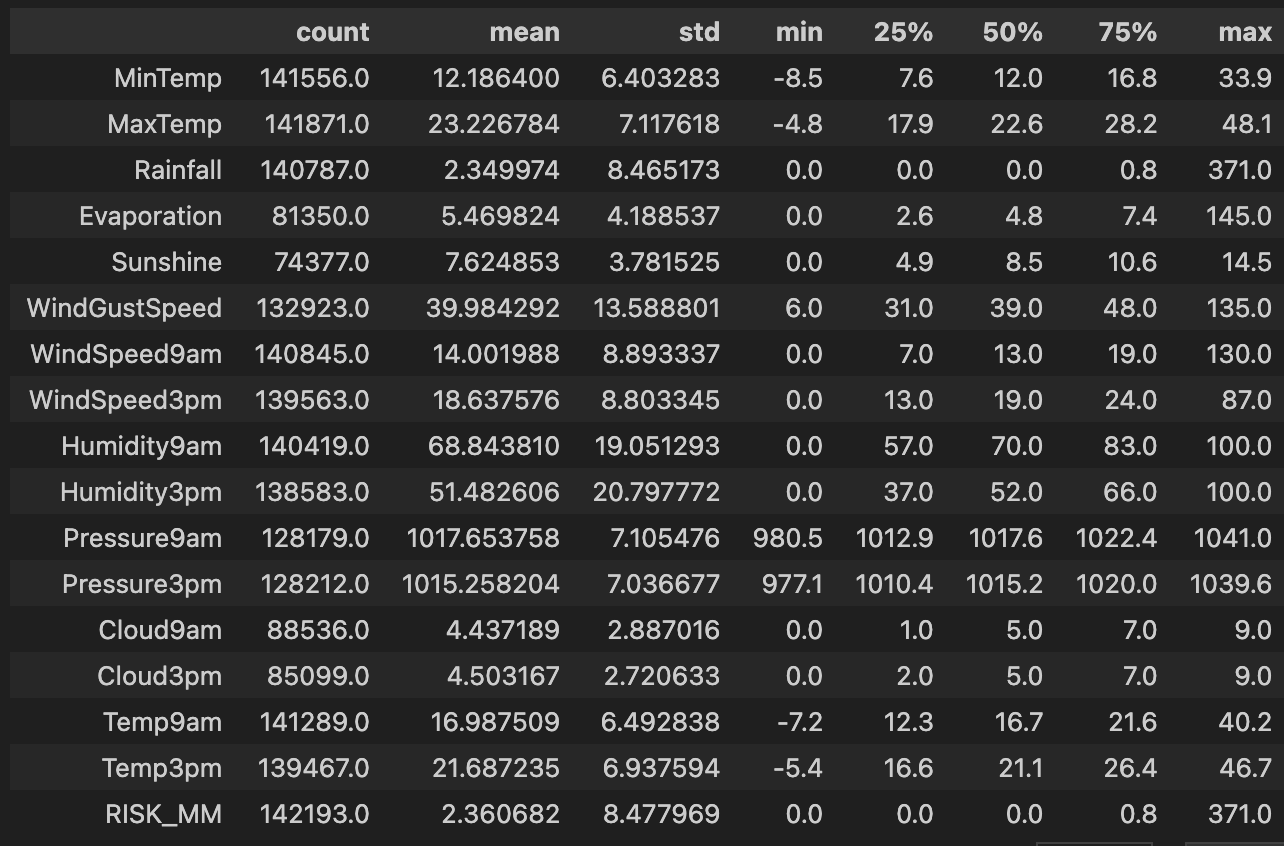
This report embarks on an in-depth exploration of a dataset specifically centered on rain in Australia. The primary objective is to unravel valuable insights into the intricate relationships within the data, and, notably, to showcase distinctions among traditional, ensemble, and neural network classification algorithms. As we delve into this analysis, we aim to enhance our understanding of meteorological phenomena and contribute to the development of effective predictive models for rainfall in the Australian context.

In the subsequent sections, we will elucidate our approach to data preparation for modeling, delve into the application of diverse classification algorithms, and discuss the discernible relationships that emerge from our comprehensive analysis.

1. **EDA**

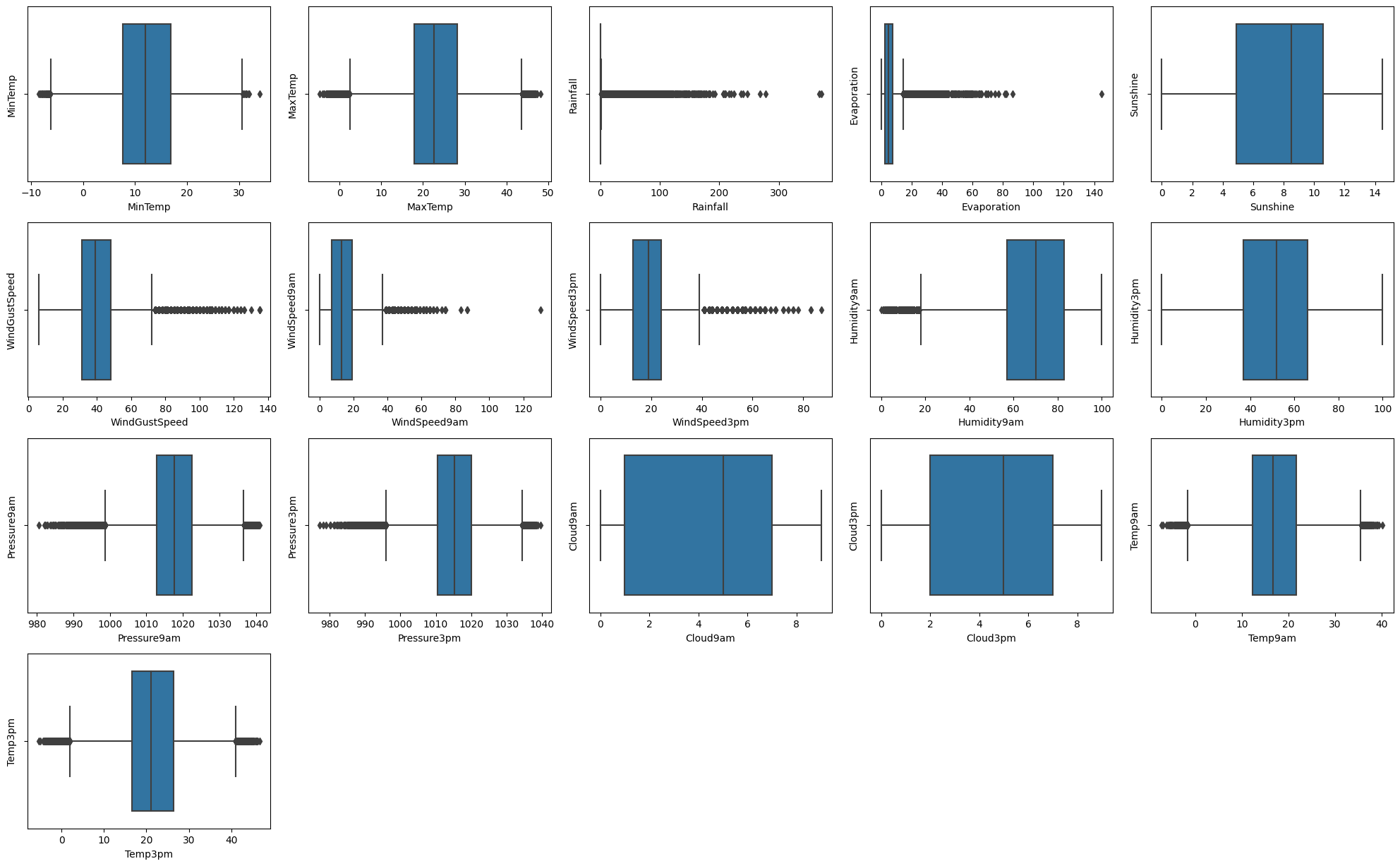
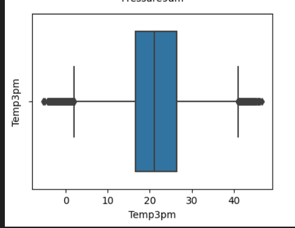
With **24** columns and **142,193** rows, our dataset provides a substantial foundation for analysis. To streamline our data, we eliminate the **'RISK\_MM'** column due to its redundancy with the **'RainTomorrow'** column. Of the remaining 24 columns, **seven** are **categorical**, and **17** are **numerical**. Notably, the **'Date'** column, initially categorized, is transformed into a **datetime** format to enhance its utility for analysis. This adjustment ensures a comprehensive dataset, featuring six categorical columns and 17 numerical columns, ready for further exploration and modeling.



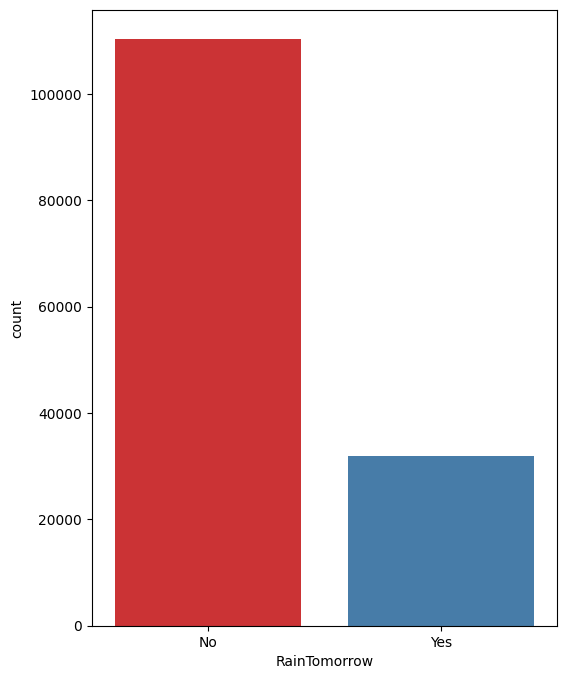
With the exception of 'Date,' 'Location,' and 'RainTomorrow,' all other columns exhibit null values. Additionally, upon inspecting the data description, evident outliers emerge in certain columns, identified by the substantial difference between the 75th percentile value and the maximum value within those respective columns.



After looking for boxplots of the full dataset we see that our obeservations are true.

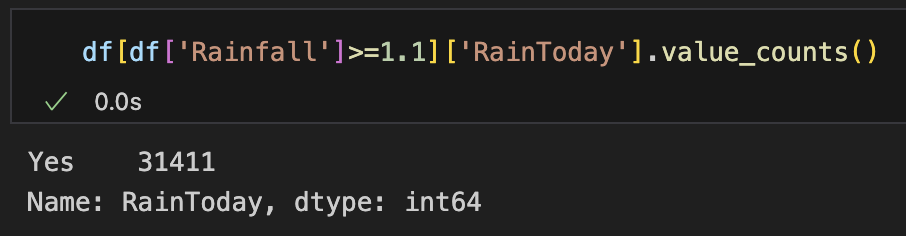


The primary outliers still exist in the 'Rainfall', 'Evaporation', 'WindGustSpeed', 'WindSpeed9am', and 'WindSpeed3pm' columns. However, outliers are also present in other columns. Nevertheless, we will disregard them since their impact is not significantly substantial. We will look at the outliers of 5 column more clearly and fix them.

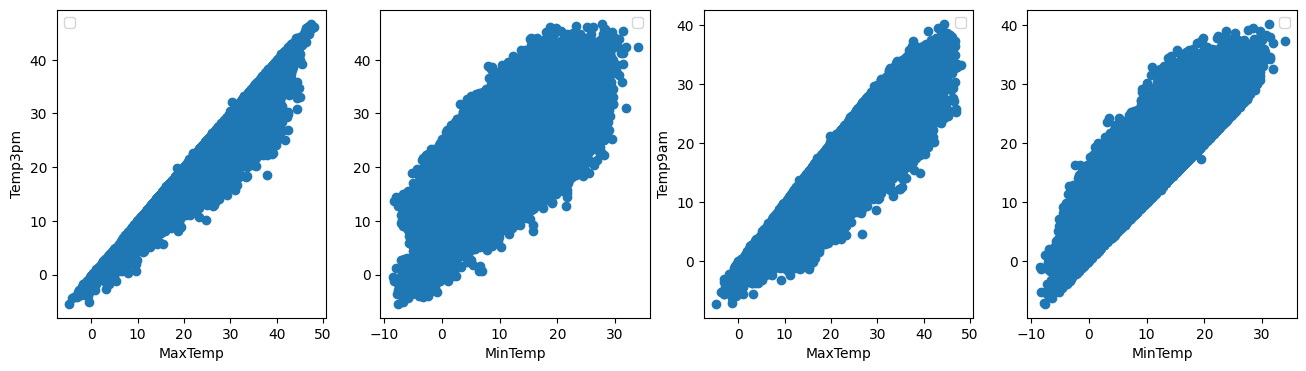
The visualization of the frequency distribution of the 'RainTomorrow' variable highlights a noticeable imbalance in the dataset. This imbalance, where one class significantly outnumbers the other, can impact the performance of classification models.

Upon closer examination of the 'Rainfall' and 'RainToday' columns, a pattern emerges where 'RainToday' is marked as 'yes' when 'Rainfall' values exceed 1.1. This observation indicates a potential relationship between rainfall values and the 'RainToday' variable. Consequently, when addressing outliers in 'Rainfall' values, setting an upper limit of 2 could be considered, given the observed pattern. Additionally, when handling null values, this relationship should be taken into account for a more accurate representation of the data.





A notable relationship is evident between 'Temp3pm' and 'MaxTemp', as well as between 'Temp9am' and 'MinTemp'. This suggests a correlation between afternoon temperatures and maximum temperatures, as well as between morning temperatures and minimum temperatures.



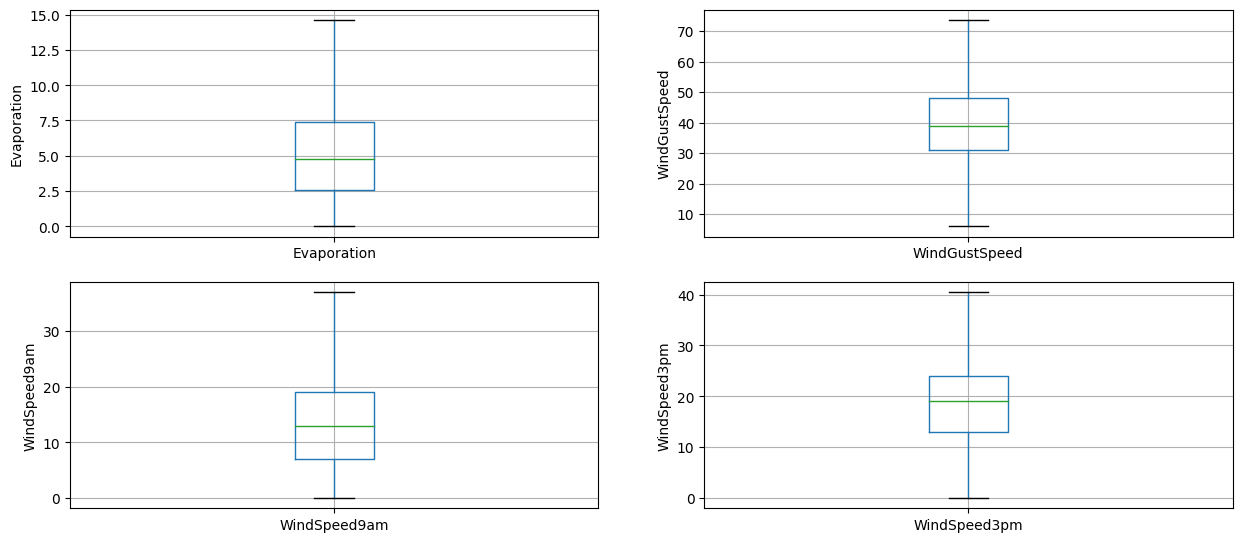
1. **Preprocessing**
   1. ***Fixing Invalid values and null values***

Null values for 'MaxTemp' and 'MinTemp' have been filled with mean values, while 'Temp3pm' and 'Temp9am' have been imputed based on 'MaxTemp' and 'MinTemp.' Other numerical columns like pressure and humidity have been filled with mean values.

Categorical columns, such as 'RainToday' and 'WindGustDir,' are handled by filling 'RainToday' based on rainfall, using mode values for other categorical columns, and filling 'WindGustDir' with the mode specific to each location.

* 1. ***Outlier cleaning***

For the selected columns 'Rainfall,' 'Evaporation,' 'WindGustSpeed,' 'WindSpeed9am,' and 'WindSpeed3pm,' the handling of outliers has been carried out using the Interquartile Range (IQR) method. Outliers have been replaced with upper and lower limit values determined by the IQR, ensuring a more robust and representative dataset for these specific features.



* 1. ***Normalization and Encoding***

In preparation for modeling, the dataset's categorical columns are appropriately encoded. The binary categorical columns, 'RainToday' and 'RainTomorrow,' are encoded using label encoding, while the remaining four categorical columns are subjected to one-hot encoding. This ensures that the data is in a suitable format for input into classification models, considering the distinct requirements of binary and multi-category variables.

1. **Classification**

Classification is a supervised machine learning method employed to predict the class of new data based on a labeled dataset. In this specific dataset, the task involves predicting whether there will be rain tomorrow, with two distinct classes: 'Rain Tomorrow' and 'No Rain Tomorrow.' The primary objective is to achieve the highest possible accuracy score. For the classification process, 70% of the dataset is allocated for training, and the remaining 30% is reserved for testing to assess model generalization. Three categories of classification algorithms are considered: traditional, ensemble, neural network.

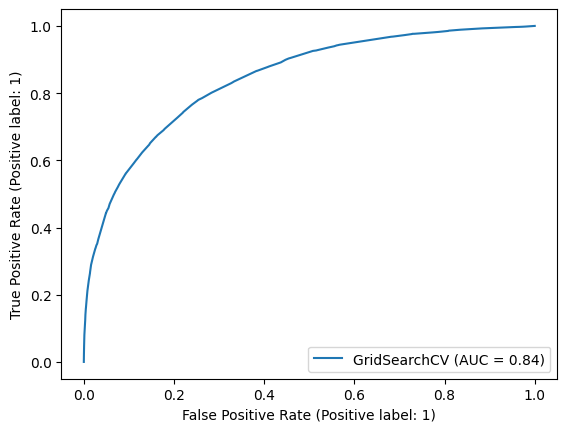
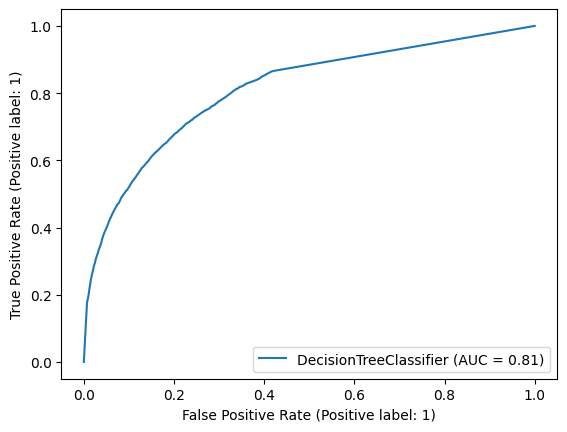
* 1. ***Traditional***

***Decision Tree***

A tree-based model that recursively splits the dataset based on features to make predictions, commonly used in classification tasks. It's effective in capturing non-linear relationships and providing interpretable insights into decision-making processes.

Without hyperparameter tuning, the initial accuracy score stands at **0.8160**. However, post hyperparameter tuning, a notable improvement is observed, elevating the accuracy **to 0.8369**. This enhancement underscores the impact of optimizing model parameters on predictive performance.

Before: After:



The best hyperparameters for the model is :

{'criterion': 'entropy',

'max\_depth': 7,

'min\_samples\_leaf': 5,

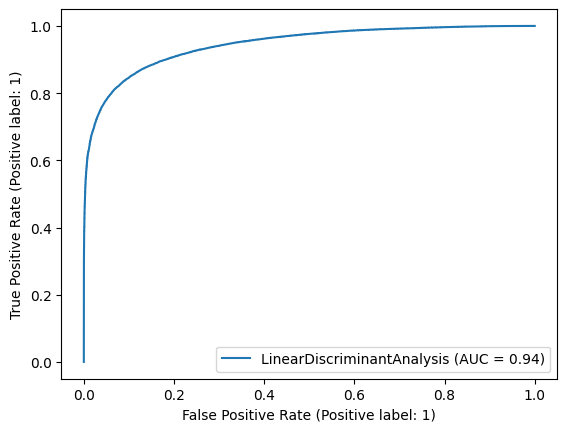
'min\_samples\_split': 5}

Due to the initially low accuracy, oversampling techniques have been applied to rebalance the dataset, and alternative algorithms are being explored. This strategic adjustment aims to address class imbalances and improve overall model performance.

***Linear Discriminant Analysis (LDA)***

A dimensionality reduction technique and classifier used for classification tasks. After making predictions our accuracy score become **0.8733**.

Also roc curve is more better than decision tree is like that



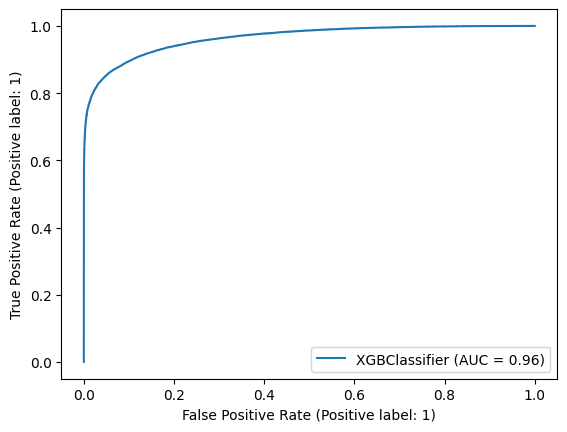
***K-Nearest Neighbors (KNN)***

K-Nearest Neighbors (KNN) is a method that classifies data points based on the majority class of their k-nearest neighbors. After experimenting with hyperparameters like n\_neighbors=5, n\_jobs=5, and algorithm='brute', the initial accuracy was **0.7700**. However, upon optimizing the number of neighbors using an accuracy graph, it was determined that 2 neighbors yielded the best performance, resulting in an improved accuracy of **0.7964**.

* 1. ***Ensemble***

***XGBoost***

XGBoost, an efficient and scalable gradient boosting framework, is recognized for its high performance. During model training, the accuracy reaches an impressive **0.9013**, indicating both speed and accuracy are notably high. Given these results, XGBoost emerges as a rational choice for this task. Additionally, the ROC curve underscores the algorithm's robust performance.



***Random Forest Classifier***

The Random Forest Classifier, an ensemble method constructing multiple decision trees for enhanced accuracy, showcases notable performance with an accuracy score of **0.8648**. Its efficiency is further emphasized by the low computation time. Not as good as the XGBoost algorithm but better than traditional algorithms.

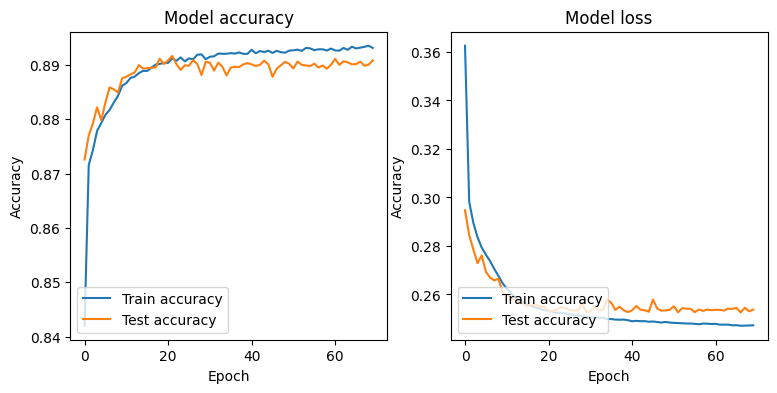
***Voting Classifier***

The Voting Classifier, configured with a soft voting strategy and comprised of individual classifiers including XGBoost, Random Forest, and Linear Discriminant Analysis (LDA), achieves an impressive accuracy of **0.8983**. This ensemble model harnesses the diverse strengths of its constituent classifiers, synergistically combining them to enhance overall predictive performance. The incorporation of complementary algorithms contributes to the robustness and effectiveness of the Voting Classifier in accurately predicting rainfall outcomes.

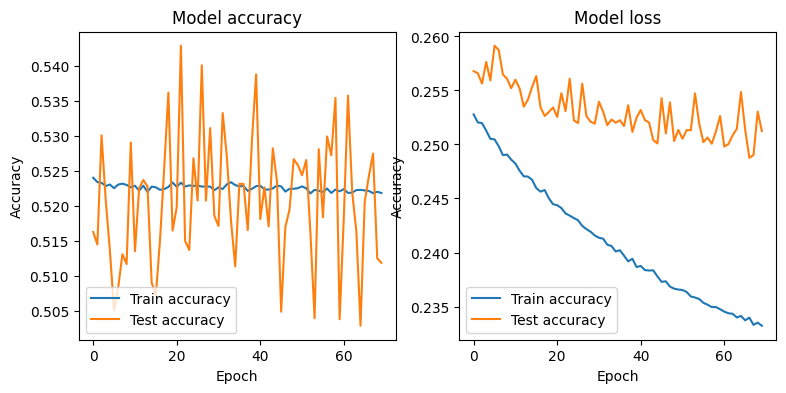
* 1. ***Neural Network Algorithms***

The Sequential model, built with the Keras library from TensorFlow, utilizes a sequential architecture for deep learning. It incorporates layers with diverse activation functions. The model is trained using two different optimizers: Adam and Stochastic Gradient Descent (SGD). Despite both optimizers yielding similar accuracy scores of **0.8908**, a closer examination of the training process reveals nuanced differences. To handle the large dataset efficiently, a batch size of 100 and 70 epochs are selected.

With Adam optimizer:



With Stochastic Gradient Descent optimizer:



1. **Conclusion**

In summary, our analysis underscores the superior performance of ensemble and neural network algorithms compared to traditional methods in predicting rainfall outcomes. While neural network algorithms exhibit a notable slowdown in processing speed as the dataset size increases, ensemble algorithms strike a favorable balance between time efficiency and accuracy. Among these ensemble models, XGBoost stands out as the top-performing algorithm, achieving the highest accuracy. Therefore, for this specific task of rainfall prediction, the ensemble approach, particularly with the XGBoost algorithm, proves to be the most effective and efficient choice.

Intuitions