Assignment-2

Rainfall Prediction (Data Science for Innovation, Autumn-2023)

Submitted To :

Dr. Ali Anaissi

Submitted By :

Hemang Sharma (24695785)
Jyoti Khurana (14075648)
Mahjabeen Mohiuddin (24610507)
Suyash Santosh Tapase (24678207)

INDEX

1. Literature review

1.1 In one of the studies , the author performed a comparative analysis of Artificial Neural Networks,Support Vector Machine and Adaptive Neuro Fuzzy Inference System on rainfall prediction.He did comparison of models in four ways:

- by using different lags as modelling inputs;
- by using training data of heavy rainfall events only;
- performance of forecasting for 1 hour to 6 hours and;
- performance analysis in peak values and all values.

According to results ANN performed better when trained with a dataset of heavy rainfall. For 1 to 4 hour ahead forecasting, the previous 2-hour input data was suggested for all three modelling techniques (ANN, SVM and ANFIS). ANFIS reflected better ability in avoiding information noise by using different lags of inputs. And finally during peak values, SVM proved to be more robust under extreme typhoon events.

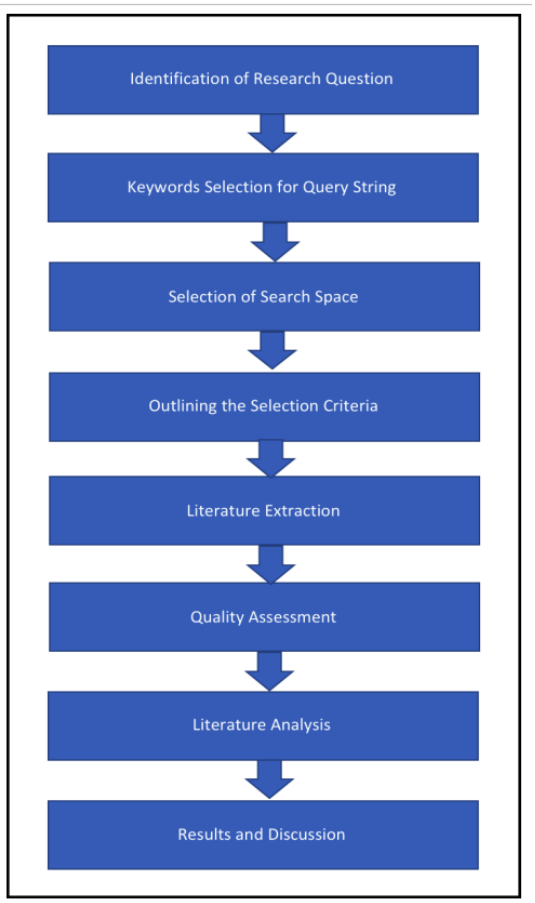

Figure: SLR Process

1.2 Another Researcher performed a comparative analysis of various data mining techniques for rainfall prediction in Malaysia such as: Random Forest,

Support Vector Machine, Naive Bayes, Neural Network, and Decision Tree. For this experiment, a dataset was obtained from various weather stations in Selangor, Malaysia. Before the classification process, Pre-processing tasks were applied to deal with the noise and missing values in the dataset. The results showed significant performance of Random Forest as it correctly classified large amounts of instances with small amounts of training data.

1.3 Thirumalai, Chandrasegar, et al. discusses the amount of rainfall in past years according to the crop seasons and predicts the rainfall for future years. The crop seasons are Rabi, Kharif and Zaid. Linear regression method is applied for early prediction. Standard deviation and Mean was also calculated for future prediction of crop seasons. This implementation will be used for farmers to have an idea of which crop to harvest according to crop seasons. Geetha, A., and G. M. Nasira. implements a model which predicts the weather conditions like rainfall, fog, thunderstorms and cyclones which will be helpful to the people to take preventive measures..They have used a hybrid approach that is combining two techniques, Random forest and Gradient boosting with many machine learning techniques like ada boost, K-Nearest Neighbour, Support vector machine, and Neural Network.These have been applied on the rainfall data of North Carolina from 2007– 2017 and also the performance is calculated by applying different metrics F-score,precision, accuracy, recall. Finally, eight hybrid models have been proposed and Gradient boosting-Ada boost has been the superior which exhibited good result

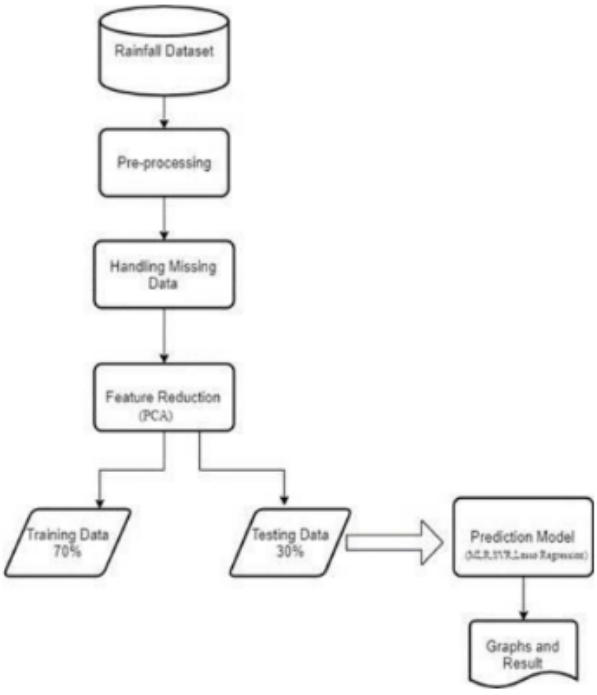

Figure : Rainfall prediction Model

1.4 Singh and Borah (2013), trained five architectures of a Feed-forward Back-propagation Neural Network algorithm containing only three layers (1 input, 1 hidden, and 1 output layer) to forecast the mean rainfall of the summer monsoon in India on a monthly and seasonal basis. The authors provide the prediction of rainfall amounts by combining the results given by the five trained Neural Networks. Only monthly values of rainfall and seasonal rainfall were used from two sources from the period 1871 to 2010, a dataset from a literature reviewed work and a dataset from the Indian Institute of Tropical Meteorology. Results showed that their ensemble approach performed better on the evaluation metrics of Means, Standard Deviations, Correlation Coefficient, Root Mean Square Error (RMSE), and Performance Parameter compared to a related work that uses a more complex Feed-forward Back-propagation Neural Network.

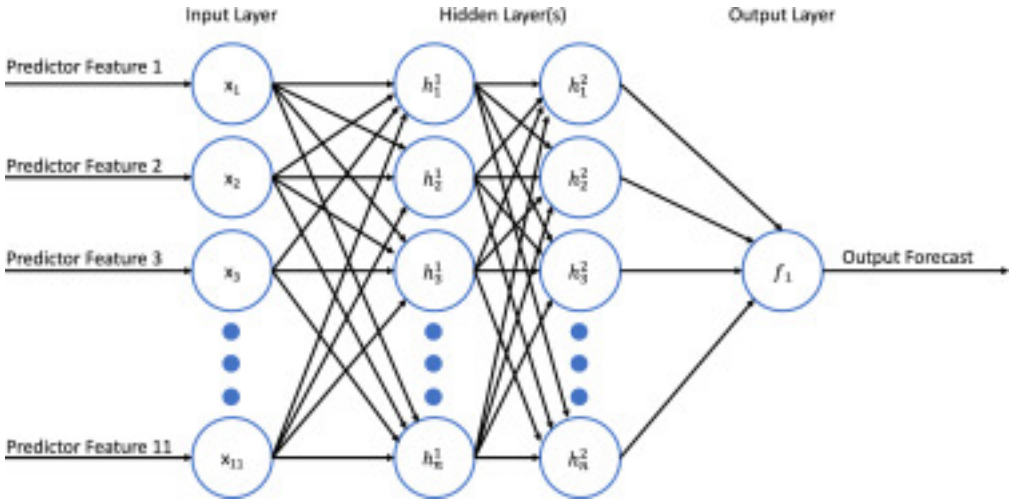

Figure : Neural Network Architecture

1.5 In this reference journal, researchers predicted the rainfall by using the proposed Wavelet Neural Network Model (WNN), an integration of Wavelet Technique and Artificial Neural Network (ANN). To analyse the performance, monthly rainfall prediction was performed with both the techniques (WNN and ANN) by using a dataset of Darjeeling rain gauge station in India. Statistical techniques were used for performance evaluation and according to results WNN performed better than ANN.

2. Setup

Research Questions:
1. What are the chances of rainfall in Australia tomorrow?
Null hypothesis: Significantly there is no difference between the predicted chance of rainfall in Australia tomorrow and a random guess.
Alternative hypothesis: The predicted forecast of rainfall in Australia tomorrow is significantly different from a random guess.

2. What is the accuracy of machine learning models in predicting rainfall?

Null hypothesis: There is no significant difference between the accuracy of machine learning models and random guessing in predicting rainfall.
Alternative hypothesis: The predicted accuracy is significantly different from random guessing in predicting rainfall

Various approaches are used to do the analysis of rain forecasting and they are described in the next section.And to evaluate the approach below measures are used:

- Accuracy - This metrics will measures the overall correctness of the predictions made by different models and it can be calculated by dividing the number of correct observations by the total number of observations
- Reliability - It assesses how well the models' predicted probabilities of rainfall align with the observed rainfall. This can be measured using calibration plots or reliability diagrams, where the predicted probabilities are compared to the actual frequencies of rainfall occurrence.
- Receiver Operating Characteristic (ROC) Curve - Area under the receiver operating characteristic can also be used to assess the performance of the models.

3. Approach

Our approach is simple, first we have performed Hypothesis testing followed by Kruskal Wallis test, and we created our model. We have used RandomForestClassifier, GradientBoostingClassifier and DummyClassifier for our model.

Hypothesis testing:

Result of mean computation on features:

MinTemp Mean is: 13.067384
MaxTemp Mean is: 23.757944

Rainfall Mean is: 2.607705
WindGustSpeed Mean is: 40.702467
WindSpeed9am Mean is: 15.289355
WindSpeed3pm Mean is: 19.530689
Humidity9am Mean is:67.574395
Humidity3pm Mean is: 50.937849
Pressure9am Mean is: 1017.232752
Pressure3pm Mean is:1014.843160
Cloud9am Mean is: 4.389851
Cloud3pm Mean is:4.464132
Temp9am Mean is:17.721812
Temp3pm Mean is: 22.254868

Result of standard deviation computation on features:

Standard Deviation of MinTemp is 6.555005
Standard Deviation of MaxTemp is7.168013
Standard Deviation of Rainfall is 9.561675
Standard Deviation of WindGustSpeed is 13.460367
Standard Deviation of WindSpeed9am is 8.575474
Standard Deviation of WindSpeed3pm is 8.577614
Standard Deviation of Humidity9am is 19.042966
Standard Deviation of Humidity3pm is 20.979821
Standard Deviation of Pressure9am is 6.974059
Standard Deviation of Pressure3pm is 6.918889
Standard Deviation of Cloud9am is 2.869173
Standard Deviation of Cloud3pm is 2.720248
Standard Deviation of Temp9am is 6.682983
Standard Deviation of Temp3pm is 7.037965

It is evident from the obtained mean and standard deviation results that Anova test does not suit here for the hypothesis testing as all the features have different standard deviation values, so we have performed the Kruskal Wallis test.

Kruskal Wallis test:

The computation result of Kruskal Wallis test is 264106.9374832491 and p-value is 0.0.Therefore, we are rejecting the null hypothesis.

Classification Model:

The analysis of Rain Prediction leads to the condition whether it is going to rain tomorrow or not.The target variable under this predictive model is categorical variable, and so target here is divided between yes or no

classes.Basically we have labels here so it is a supervised machine learning Classification model.
The objective is to make a model that classifies the problem by learning the patterns of data to find the best correlation among all the features to predict the outcome of the target variable i.e., whether it is going to rain tomorrow or not.

Training and testing the Supervised machine learning on algorithms:

To train the predictive model we have split the dataset into a training and validation set as 80% data for training and 20% data for validation and then started training the predictive model on algorithms such as:

1. RandomForestClassifier

- By using multiple features we are building a forest and this will help in training
- multiple decision trees.
- In this model, the ensemble helps to train multiple models at a time and the majority of votes leads to the final prediction.
- Bagging is helping this model to pick the observations for training a model.
- Function criterion as gini is helping the trees to split and remove the impurities from the features to split smoothly for prediction.
- We have tuned the model with the hyperparameters such as n_estimator with 100 decision trees for prediction and max_depth as 17 to build the height of a tree as 17 from root node to leaf node.

2. GradientBoostingClassifier

- Gradient boosting algorithm is one of the popular algorithms for its prediction speed and accuracy, it suits well for large and complex datasets.
- With the help of log-likelihood as loss function our model is minimising the errors.
- This Gradient boosting algorithm is helping our model to minimise the bias error when the model is building a tree.
- We have developed a decision tree with hyperparameters such as 200 trees and the length of this tree is 11.
- The hyperparameter learning rate taken by us is 0.07 which is lower than 0.1 and this rate is helping our model to learn the patterns in data fastly and this produces the improvements that are helpful in generalising the model.

3. DummyClassifier

- DummyClassifier is taken from scikit-learn, it helps in providing the strategies to the model for generating the predictions.
- Dummy Classifiers make predictions that ignore the input prediction.
- This classifier acts as a baseline model and with this classifier model we are comparing it against our other models.
- Using the strategy parameter we are selecting the behaviour of a specific baseline.
- We have fitted the model with y parameters as features are ignored for this model.

Hyper parameter tuning functions

Randomised Search Cross Validation for Hyperparameters Tuning:

- RandomizedSearchCV is a function that is performing hyperparameter tuning for our machine learning model. It is performing searches over the specified parameters in Random Forest Classifier and returning the best parameters for our model. The parameter distribution used here is grid parameter as it is a dictionary of hyperparameters and their possible values.
- Trying with 10 numbers of combinations and performing cross validation as 5 folds during the optimization process.
- Using the number of jobs as -1 which is making our model run parallel.
- It then repeats this process multiple times and selects the best set of hyperparameters and this is helping our model to get the highest accuracy score.

Receiver Operating Characteristic (ROC) Curve:

This algorithm is used to evaluate the performance of a binary classifier at different classification thresholds.

- We are plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different threshold values.
- The graph is highlighting the relationship between going to rain tomorrow and not going to rain tomorrow.
- As the graph is moving from zero to towards the left hand border and then straight across horizontally, this indicates the test is accurate.
- The area under the ROC curve (AUC-ROC) is a performance metric that ranges from 0.5 to 1. A higher AUC-ROC indicates better model performance.

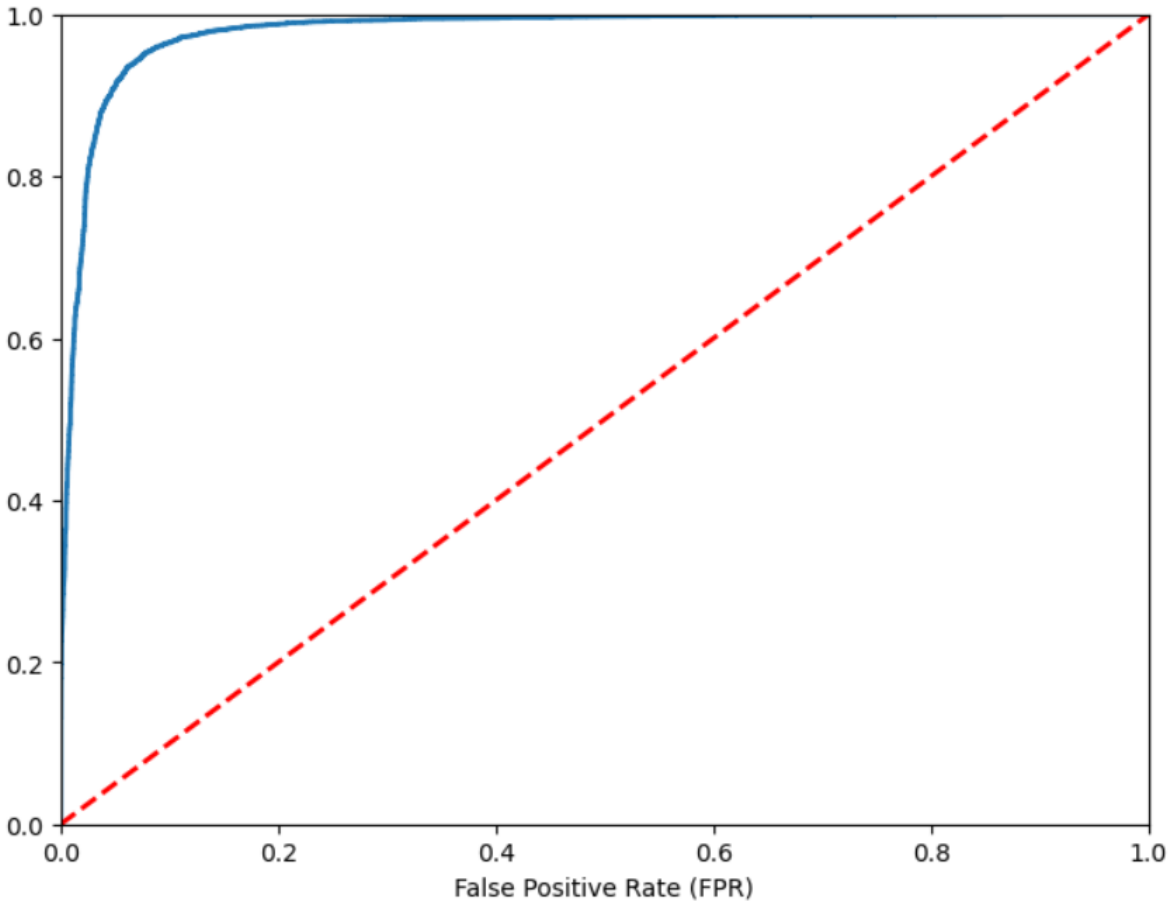

Figure : ROC Curve

4. Results

1. RandomForestClassifier

The obtained train score of Random forest model is 0.9804006278711425 and test score is 0.9274235355106273, where as the accuracy sore and F1 scores are: 0.9274235355106273 and 0.9293464547060308. As the accuracy score is more than 50% so we can say here that the model has performed predictions accurately.

Confusion matrix:

It is evident from the matrix below that data is well distributed among the classes.

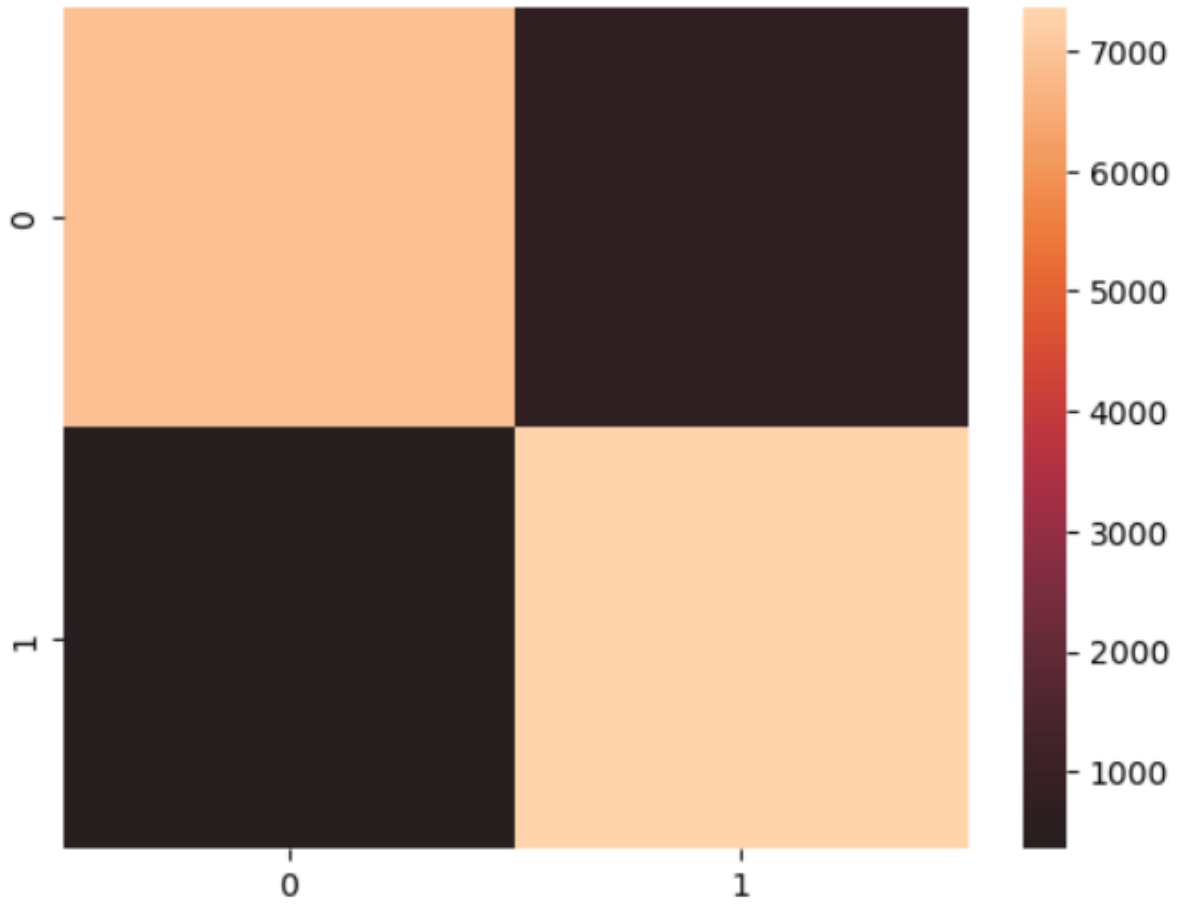

Figure: Confusion matrix for Random Forest Classifier

2. GradientBoostingClassifier

The train and test scores achieved from the Gradient Boosting Classifier are 0.9900923085785055, 0.9331907724209435. And the accuracy and F1 scores are 0.9331907724209435, 0.93528340970435.

Though there is a slight difference between train and test results, still we can say that the model is performing significantly better.

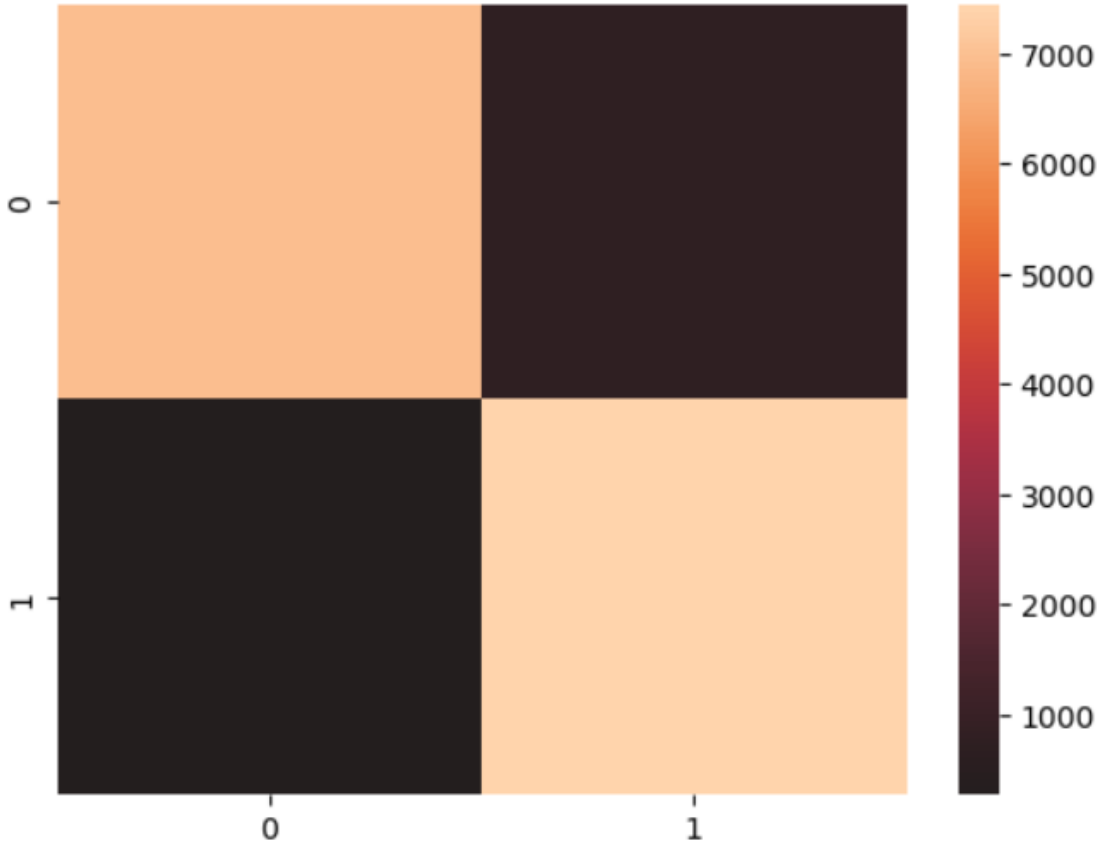

Figure: The confusion matrix for Gradient Boosting Classifier

The confusion matrix shown above gives the conclusion that the model has distributed the data well among the classes.

3. DummyClassifier

The test score of Dummy classifier is 0.49954639709694143

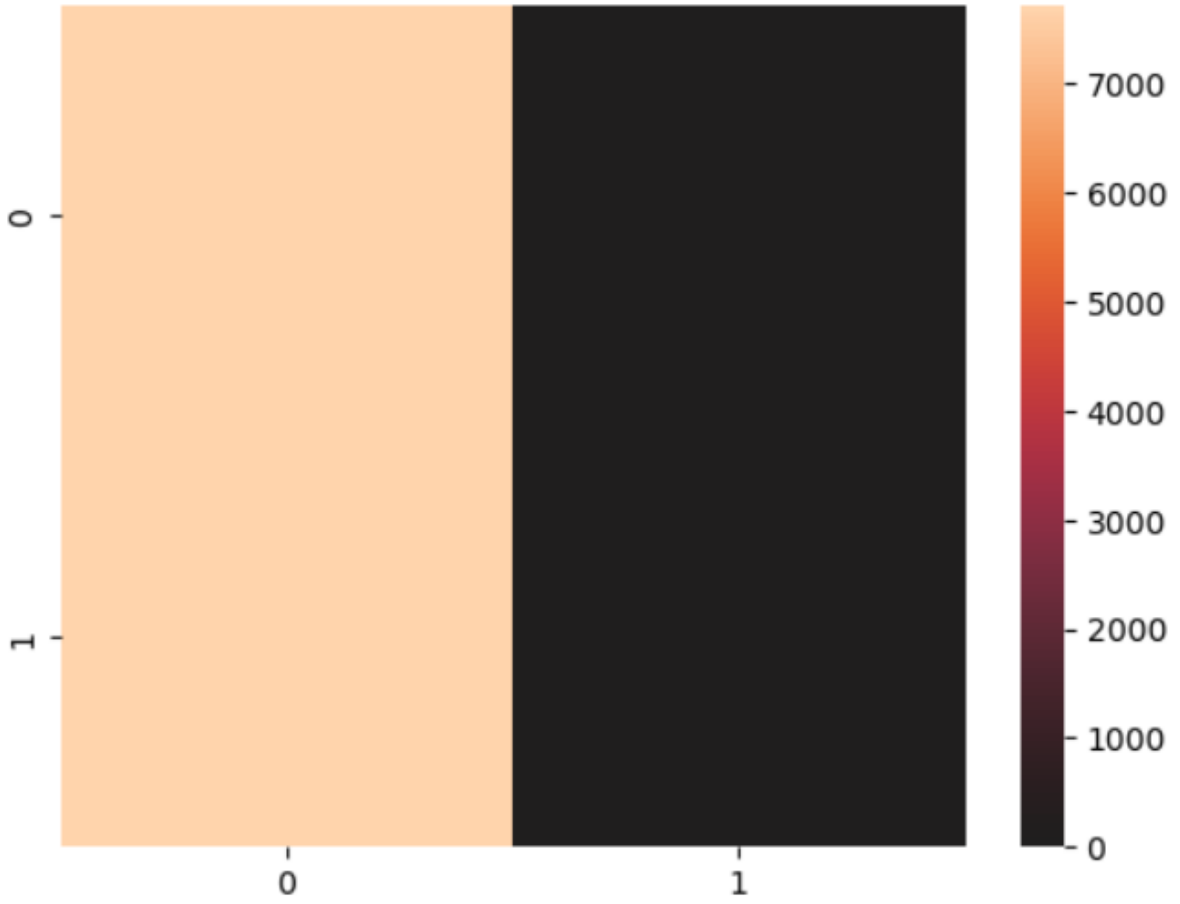

Figure: Confusion matrix for Dummy Classifier

Answer to the research questions

1. What are the chances of rainfall in Australia tomorrow?

The models have given the prediction that there is a high probability that there will be no rain tomorrow.

2. What is the accuracy of machine learning models in predicting rainfall?

With a 0.9% accuracy rate, therefore we can say that the model has learned the patterns of data very well and has given us the prediction that there is a 0.9 percent chance that there will be no rain on the next day.

5. Conclusion

We have analysed the performance of various predictive models and came to the conclusion that the model has learned all the patterns of data fitted to the model and has produced the accurate prediction that there will be no rain the next day. We also have saved our trained model and this will help us in accurately predicting the next day's weather condition any time in the future. We don’t have to train the model multiple times or execute the whole code again.

6. References

- Observations were drawn from numerous weather stationsThe daily observations are available from http://www.bom.gov.au/climate/data
- Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml
- Data sourcehttp://www.bom.gov.au/climate/data
- https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data
- https://github.com/hemangsharma/Assignmnet2_DSI_36100