PREDICTION OF OCCURRENCE OF RAINFALL THE NEXT DAY IN AUSTRALIA USING LOGISTIC REGRESSION, K-NEAREST NEIGHBORS, AND RANDOM FOREST CLASSIFIER

CVL 736 – Soft Computing Technique in Water Resources

Project Report

INDIAN INSTITUTE OF TECHNOLOGY DELHI

Submitted to: -

Manabendra Saharia, Assistant Professor,

Department of Civil Engineering, IIT Delhi

Submitted by: -

Gaurav Mali (2023CEW2465),

M.Tech in Water Resource Engineering,

Department of Civil Engineering, IIT Delhi

INTRODUCTION

A crucial component of decision-making in many industries is weather forecast. Particularly for sectors like agriculture, transportation, and emergency management, accurate rainfall forecasting is crucial. The complexity of meteorological data is attributed to its dynamic patterns and interdependencies between different variables. Machine learning is an appealing alternative to traditional forecasting techniques as they frequently fail to catch these small details. In this project, different classification models are being used to predict the occurrence of rainfall tomorrow based on today's weather data. Input variables considered for predicting the rainfall occurrence are maximum and minimum temperature, wind speed, wind direction, humidity, pressure, location, and rainfall today. Logistic Regression, K-nearest neighbors, and Random Forest Classifiers are different classification models used to predict the occurrence of rainfall. Their performance is compared using the accuracy score and confusion matrix to find the most accurate model. The dataset contains missing values and outliers which are handled by performing data pre-processing to clean the data before training the models.

METHODOLOGY

The dataset was initially obtained from Australia's Bureau of Meteorology, where about ten years' worth of daily weather observations from several places in Australia are included in this dataset. It consists of 145460 data and 22 input columns. A wide range of meteorological factors, such as temperature, humidity, wind speed, location, rainfall, evaporation, sunshine, cloud, and atmospheric pressure, are included in the dataset and used during this study. The long-term collection of data ensures a thorough understanding of weather patterns. However, after performing data pre-processing, it was obtained that the dataset contains missing values and outliers that need to be removed. The sunshine and evaporation columns contain more than 40% missing values and the cloud column contains about 39% missing values, hence these columns are removed from the data-frame using python pandas. The dataset is then divided into training set will be used to train the models and testing set which will be used to ensure the effectiveness of the models in real world scenarios in 70 % and 30 % respectively. The dataset is split into training and testing set before imputing the remaining missing values in order to prevent data leakage. The remaining missing values are imputed using most frequent strategy for the categorical input variables and for the numerical input variables, median strategy is used. Distribution for numerical input features are plotted using distribution graph

and it is obtained that wind speed follows right skewed and humidity is left skewed. In contrast, the rest follows a normal distribution. Boxplot is used to detect the outliers present in the input data. It is essential to handle the outliers as it may affect the performance of some models significantly. The Z-score method is used to detect and remove the outliers for the input variable following normal distribution, the z-score is calculated for each data point in the training and testing set and if the z-score exceeds a certain threshold (threshold = 2.5 in our case), then it is considered as outliers. These outliers are replaced using the training and testing set's median, respectively, because the median is a more reliable indicator of central tendency than the mean and is less susceptible to extreme extremes. The interquartile range (IQR) method is used to detect and remove the outliers for input data following skewed distribution. The interquartile range is found for both training and testing sets by subtracting the first quartile (Q1 = 25th percentile) from the third quartile (Q3 = 75th percentile) of the data (IQR = Q3 – Q1). Lower and upper bounds are used to identify outliers, lower bound = Q1 – k * IQR and upper bound = Q3 + k * IQR, where k = 1.5 for our data (user-defined constant). These bounds act as clipping thresholds, where data points that fall outside the defined clipping thresholds are outliers, and are truncated to the nearest threshold value. Clipping outliers using IQR is effective in handling skewed or non-normally distributed data.

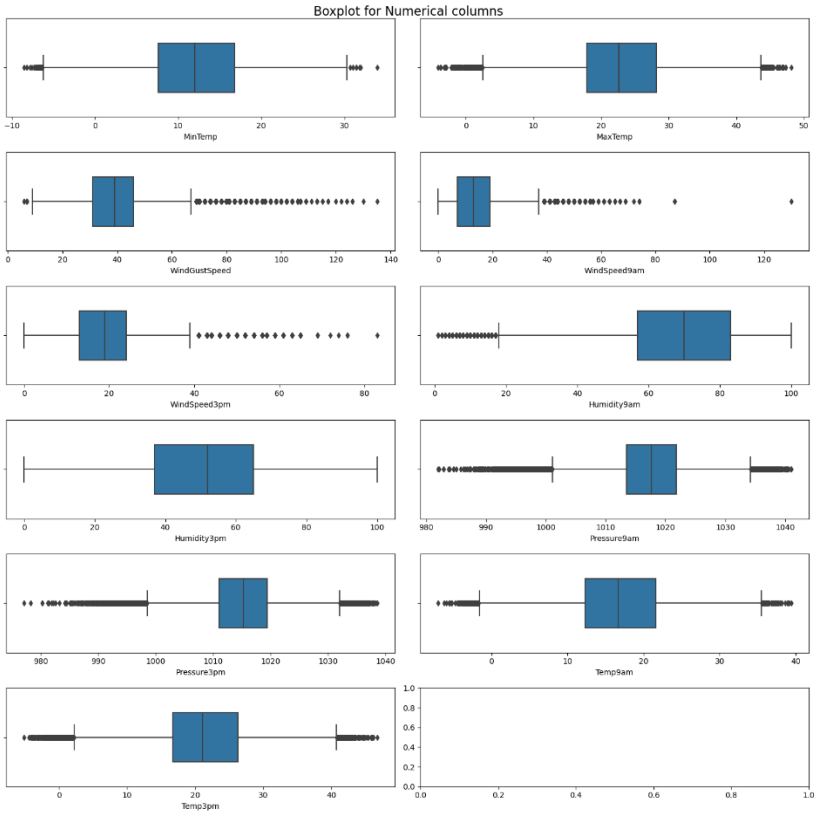

Fig 1. Boxplot of numerical inputs before removing outliers.

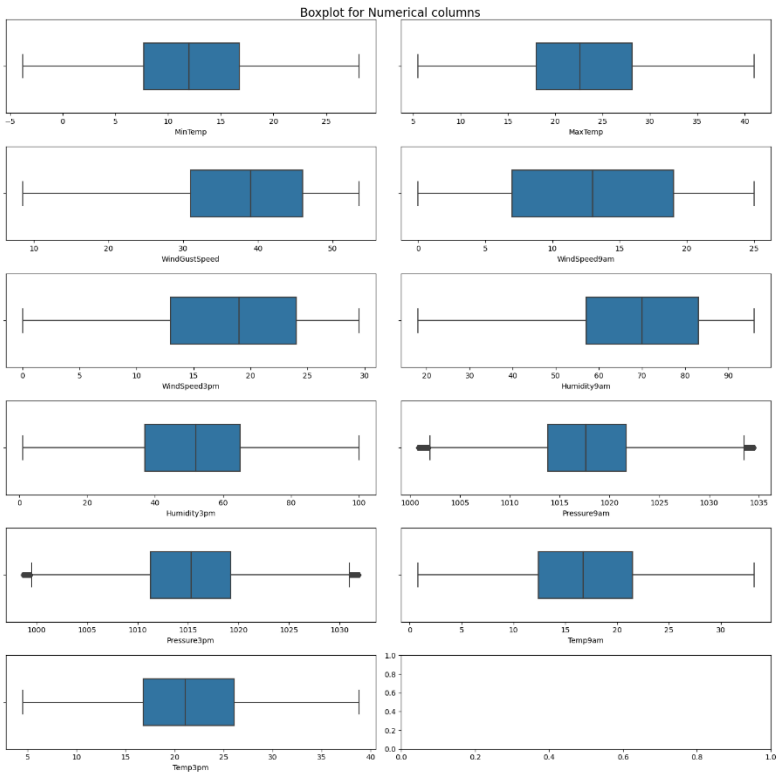

Fig 2. Boxplot of numerical inputs after handling of outliers.

Categorical input variables are then encoded to convert them into numerical representations to ensure compatibility with various algorithms as machine learning models usually operate on numerical data. Training and testing datasets are then scaled using MinMaxScalar to ensure that every numerical characteristic in a classification issue is on a consistent scale typically between 0 and 1. Because of scaling, the model performs better and has more stable convergence since some characteristics are prevented from dominating the training process based on their magnitudes. Min-max scaling is especially useful for distance-based methods like the KNN algorithm.

1) Logistic Regression:

An approach for binary classification that is often used is called logistic regression. On the basis of the input features, it estimates the likelihood of a binary result, in this example, whether or not it will rain tomorrow. The model is trained on the training set and it is evaluated on the testing set. The model's performance is evaluated with respect to the testing set by using measures like accuracy, and using a confusion matrix to determine how well the model can classify instances of rainfall and non-rainfall.

2) K-Nearest Neighbors (KNN):

KNN is an instance-based, non-parametric learning method. A data point is classified by comparing it to the training set's k-nearest neighbors. Cross-validation is used to identify the ideal value for k, and the testing set is used to evaluate the model's performance. K value used here is 15. Similarly, as previous model, the performance is evaluated using measures like accuracy score and confusion matrix, and also analysing how changes in the number of neighbors impact the model performance.

3) Random Forest Classifier:

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Here, the parameters used are max depth = 11, number of trees = 40, and criteria = 'gini'. It is excellent at identifying complex relationships within the data. As with the previous algorithms, the Random Forest model is assessed, and trained, and its performance is evaluated using an accuracy score and confusion matrix.

RESULTS

The performance of these three algorithms is evaluated based on multiple metrics to provide a comprehensive understanding of their strengths and weaknesses. These measures consist of the accuracy score and confusion matrix. The best rainfall prediction method is chosen after a comparative analysis of the models' performances. The performance evaluation parameter of the three models are shown in Table 1.

Table 1. Accuracy score of training and testing dataset using the three models.

	Accuracy Score	
Models	Training dataset	Testing dataset
Logistic Regression	0.8415	0.8388
KNN	0.8261	0.8045
Random Forest Classifier	0.8655	0.8448

The confusion matrix of the three models are shown in figures below.

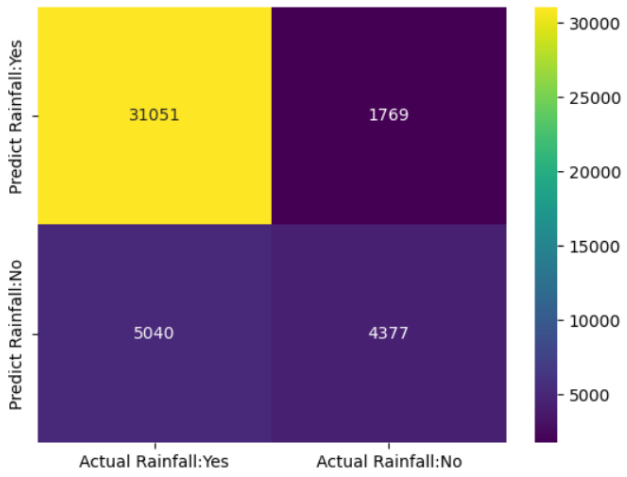

Fig 3. Confusion matrix for Logistic Regression

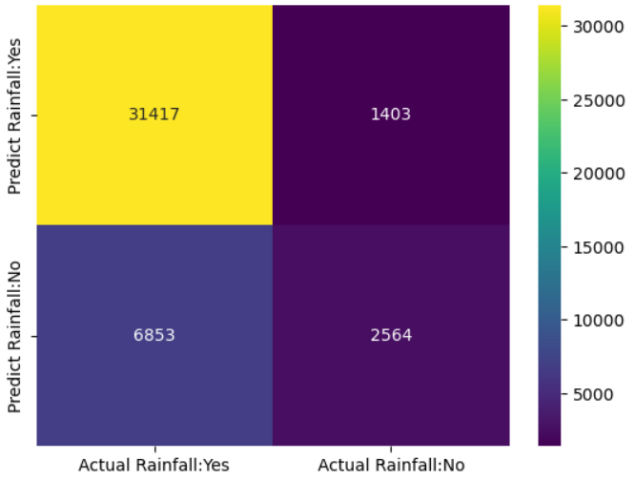

Fig 4. Confusion matrix for KNN.

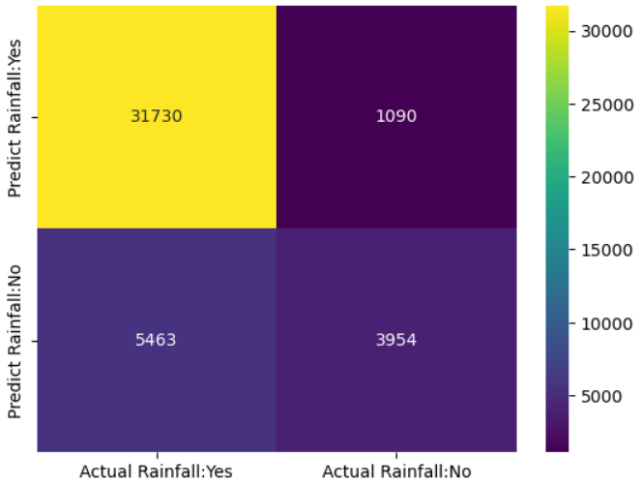

Fig 5. Confusion matrix for Random Forest Classifier.

CONCLUSION

From the results, after comparing the accuracy scores of the three models, we obtained that the accuracy score for the training and testing set is highest in the Random Forest Classifier model. From the confusion matrix, it is obtained that the logistic regression model has made 31051 + 4377 = 35427 accurate predictions and 5040 + 1769 = 6809 wrong predictions, and the KNN model has made 31417 + 2564 = 33981 accurate predictions and 6853 + 1403 = 8256 wrong predictions, and the Random Forest Classifier model has made 31730 + 3954 = 35684 accurate predictions and 5463 + 1090 = 6553 wrong predictions.

From the above comparison, we can say that the Random Forest Classifier and Logistic Regression model performed almost similar for testing dataset but overall Random Forest algorithm performed better compared to the KNN and Logistic Regression models.

References

1. Sarasa-Cabezuelo, A. (2022). Prediction of Rainfall in Australia Using Machine Learning. *Information*, *13*(4), 163. *MDPI AG*. http://dx.doi.org/10.3390/info13040163