IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Machine Learning-Based Approach for Weather Prediction based on Weather Data using Data Mining

Submitted By

Arpit Dubey

INTRODUCTION

- Weather reports helps the people to prepare their daily plan.
- These reports are prepared based on the weather forecasting techniques.
- Data mining techniques are used to predict futuristic values, trends etc.
- Machine Learning algorithms are used to predict the weather condition.
- Atmospheric features are used for classification of weather data.

USE CASE

- Identification of location-wise atmospheric trend.
- Correlation analysis to select the relevant features.
- Predictive analysis to predict the future rainy days.
- Comparative analysis will be performed to find the best classifier.
- Hyperparameter optimization of best performing classifier.
- The best classifier will be used to predict future rainy days.

DATA SET

- The dataset based on the weather records of different cities in Australia along with atmospheric parameters.
- The data has been collected from Kaggle.
- There are mainly two weather datasets i.e.,
  - Weather Dataset [Having the target attribute “RainTomorrow”]
  - Unknown Weather Dataset [Without having target attribute “RainTomorrow”]
- The Weather dataset consists of 21 features and 1 target variable.

DATA QUALITY ASSESSMENTS

- The dataset contains several null values which will affect the experiments as these values have no analytical meaning.
- Also some outliers have been found in several columns during the primary analysis.
- Empty fields of character and numerical variables have been identified which will be resolved during data cleaning and pre-processing step.

DATA EXPLORATION

- The correlation has been analyzed between the numerical features.

- During the analysis, some high correlations have been found.

- Although due to the various features of the dataset the matrix is quite large in size.

DATA VISUALIZATION

- Maximum and Minimum Temperature Distribution
- Wind Gust Distribution
- Wind Speed Distribution
- Humidity Distribution
- Pressure Distribution
- Cloud Distribution
- Temperature Distribution

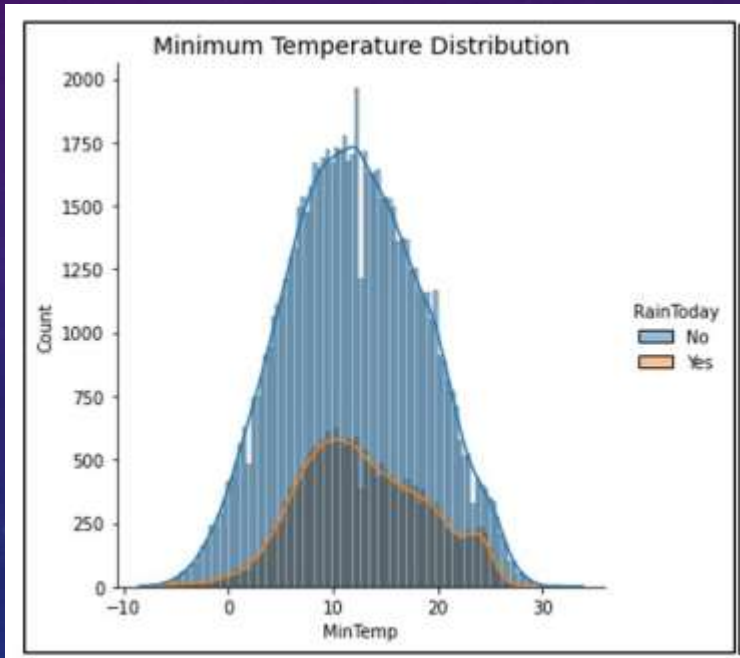

Fig. 1. Distribution of minimum temperature.

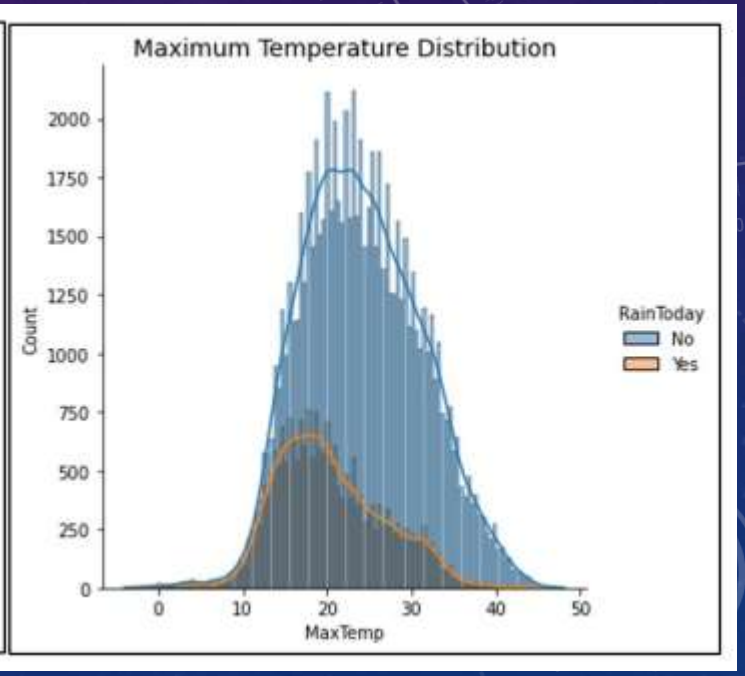

Fig. 2. Distribution of minimum temperature.

FEATURE ENGINEERING

- Null Analysis helps to identify the null values.
- Mean values of the numerical columns were used to fill empty and null fields.
- Most frequent values of the character columns were used to fill empty fields.
- Remove the outliers from the dataset.

EXPLORATORY DATA ANALYSIS - I

- Average Wind Speed Analysis :
  - Highest wind speed at 9 am – Melbourne (20.29 kmph)
  - Highest wind speed at 3 pm – Gold Coast (25.77 kmph)

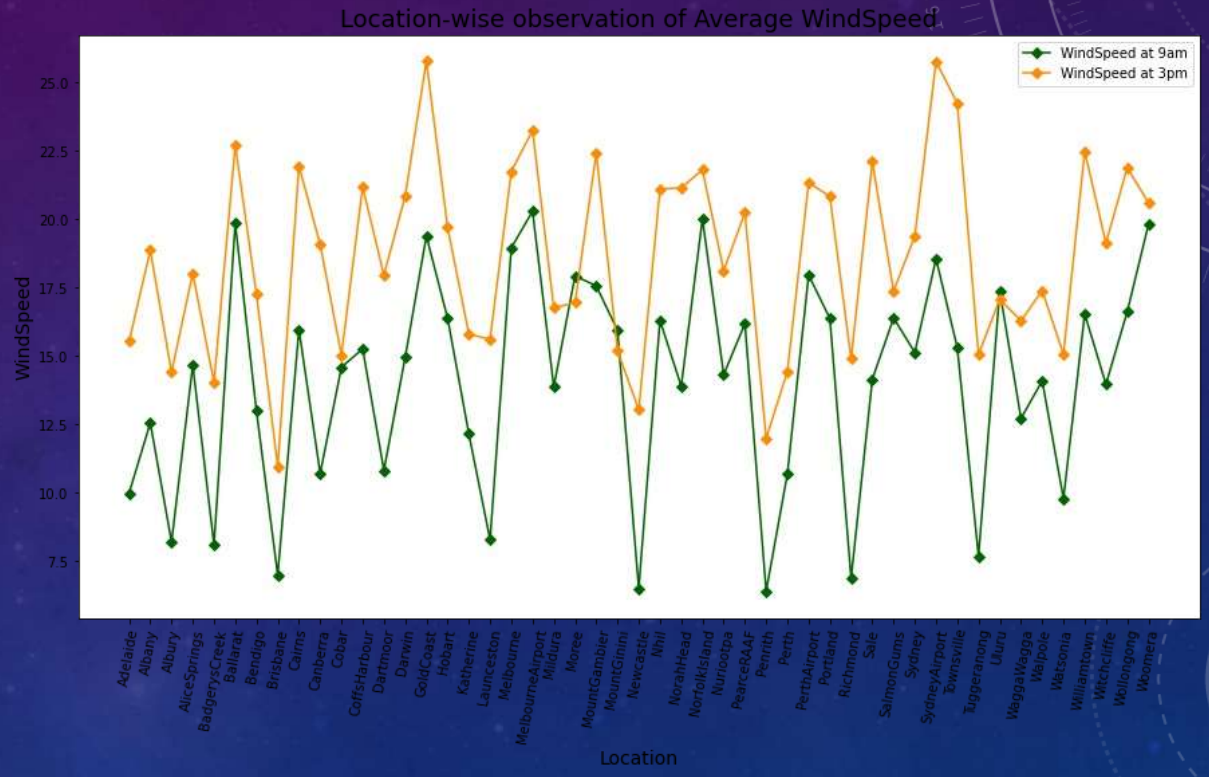

Fig. 3. Location-wise observation of average wind speed.

EXPLORATORY DATA ANALYSIS - II

- Average Humidity Analysis:
  - Highest humidity at 9 am – Dartmoor (84.38%)
  - Highest humidity at 3 pm – MountGinini (68.24%)

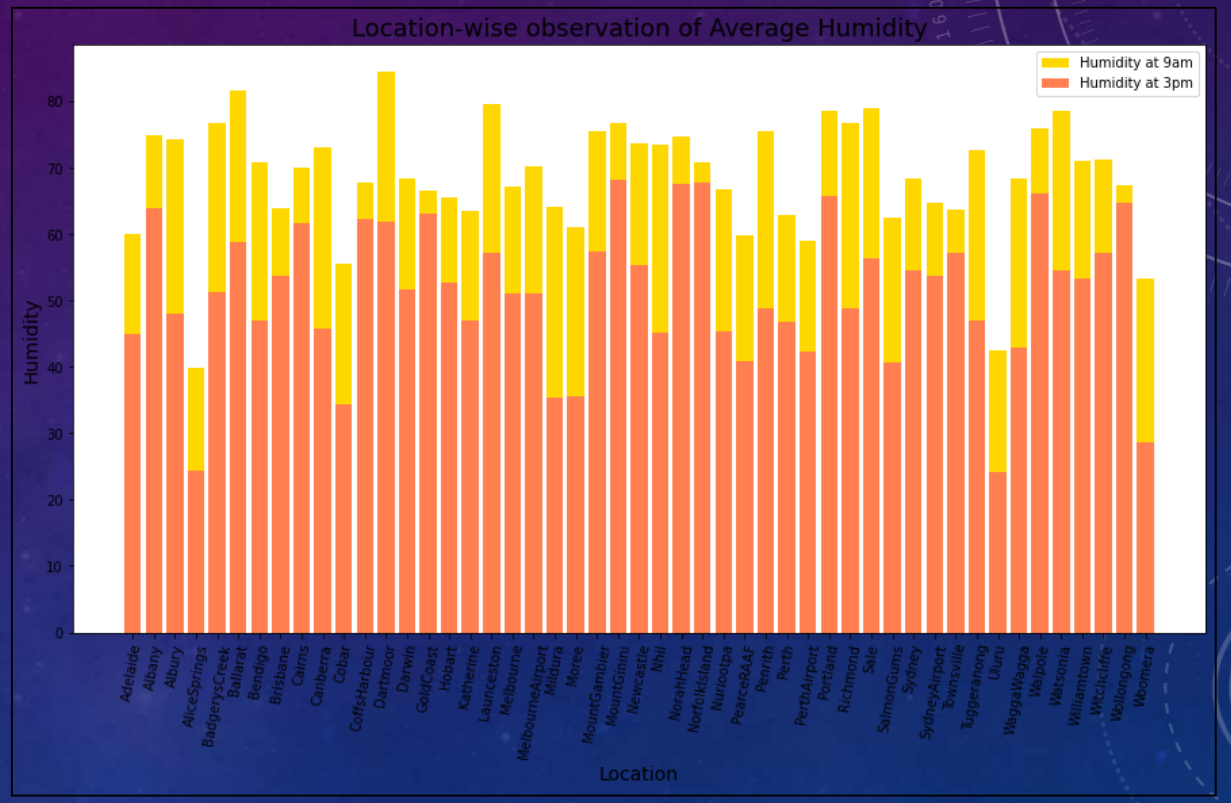

Fig. 4. Location-wise observation of average humidity.

EXPLORATORY DATA ANALYSIS III

- Average Pressure Analysis :
  - Highest Pressure at 9 am – Canberra (1018.93 hPa)
  - Highest Pressure at 3 pm – Adelaide (1016.79 hPa)

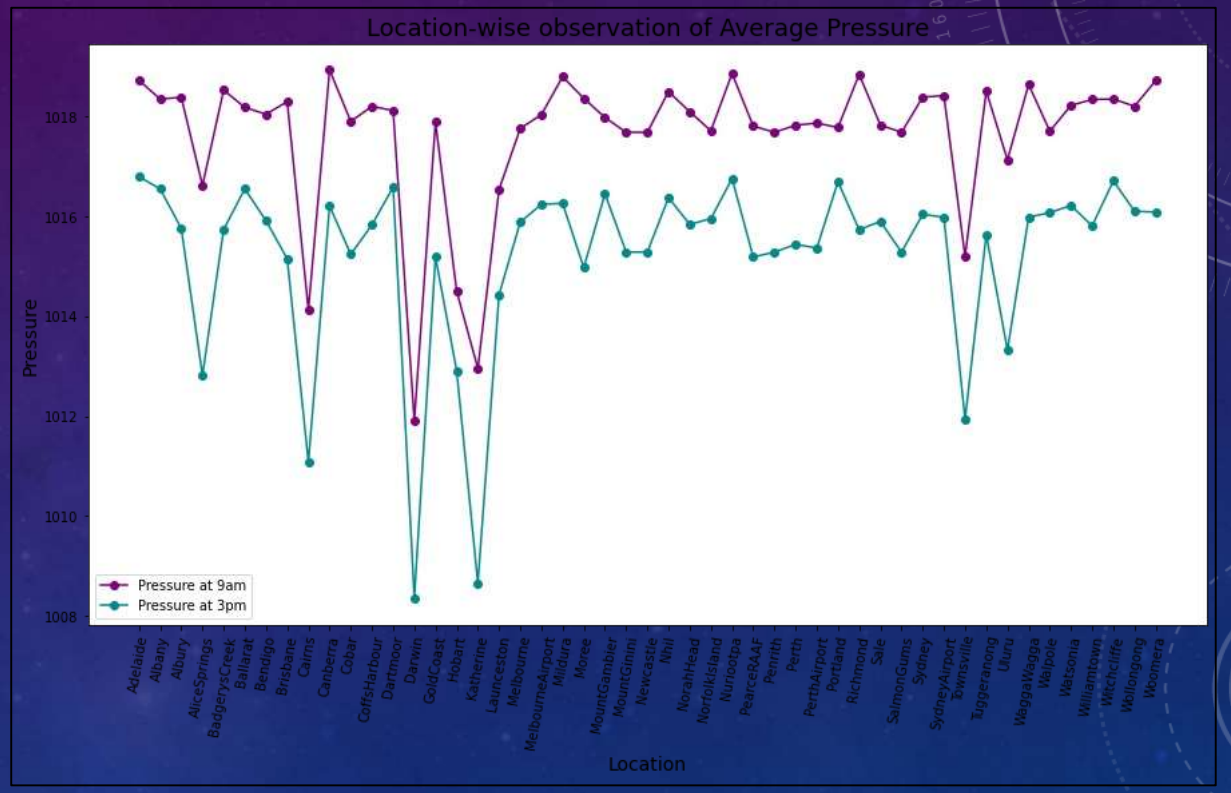

Fig. 5. Location-wise observation of average Pressure.

Exploratory Data Analysis - IV

- Average Temperature Analysis :

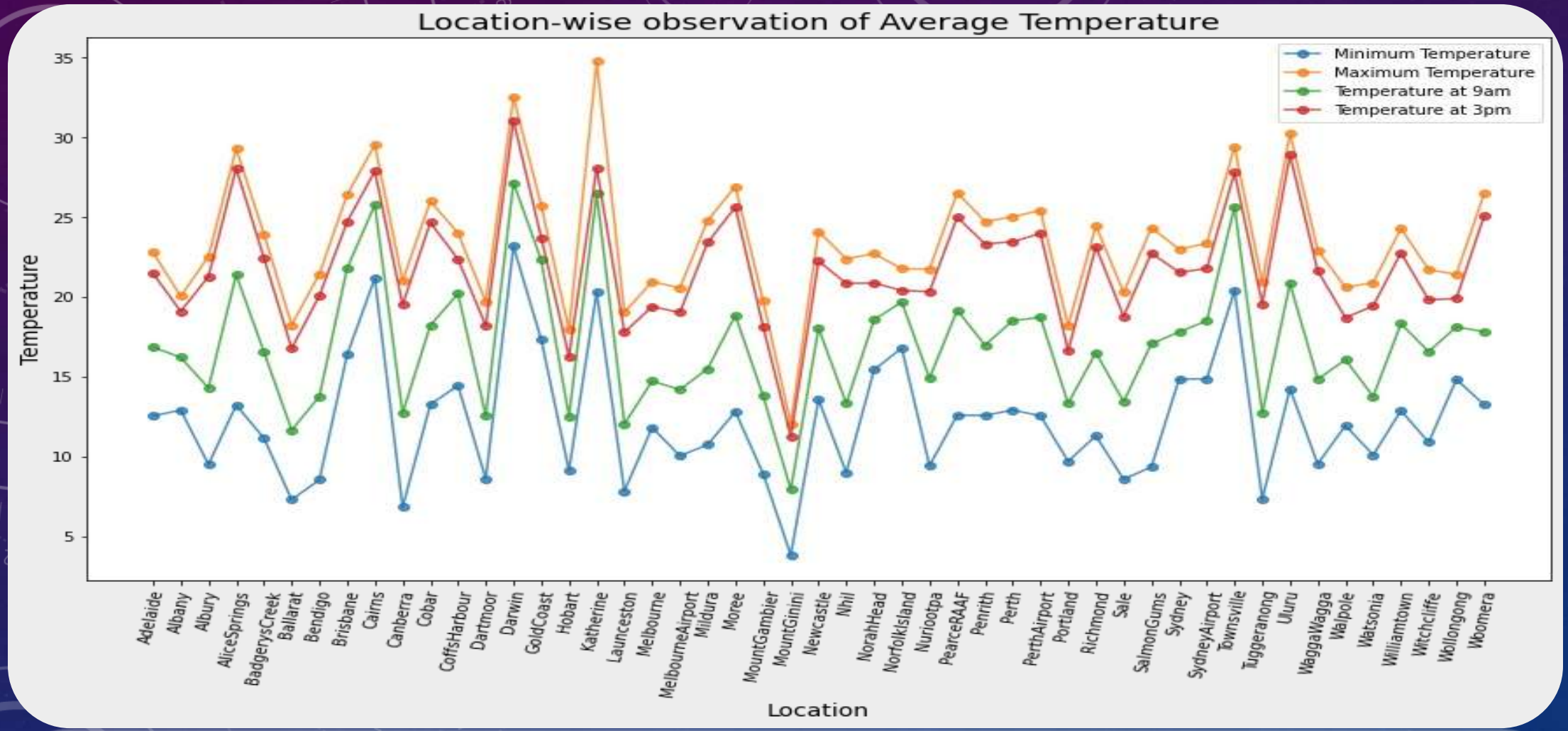

Fig. 5. Location-wise observation of average temperature.

FEATURE ENGINEERING

- Feature Extraction :
  - Conversion of character variables into numerical variables.
  - Character variables labeled as encoded using One-Hot Encoding.
- Feature Selection :
  - Select the features having correlations at least 0.20 or above.

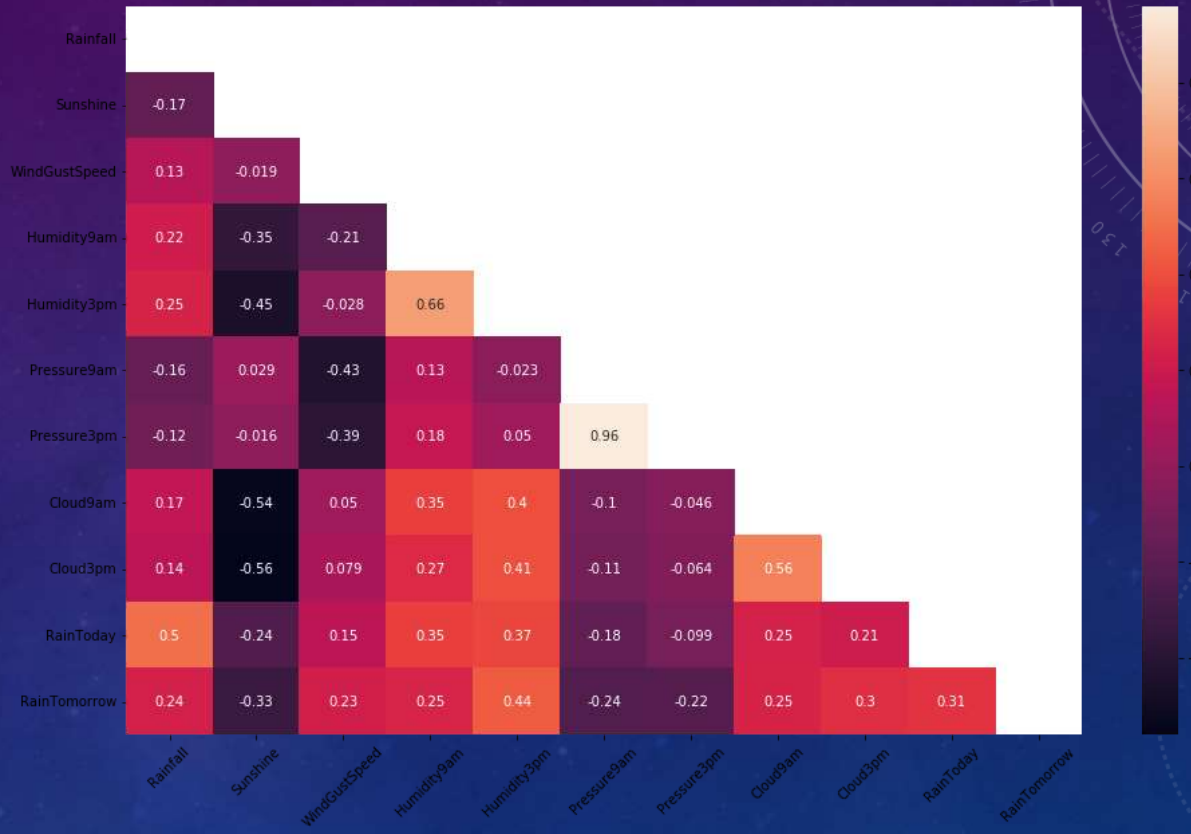

Fig. 6. Correlation matrix for the numerical features of weather dataset.

MODEL TRAINING & EVALUATION

- Training, Testing and Validation Data :
  - Train Data [80% of the Weather Data]
  - Validation Data [20% of the Weather Data]
  - Test Data [Unknown Weather Data]

- Machine Learning & Deep Learning Classifiers :
  - Logistic Regression [Classification Accuracy - 84.45%]
  - K Nearest Neighbors [Classification Accuracy - 83.14%]
  - Decision Tree [Classification Accuracy - 78.28%]
  - AdaBoost Tree [Classification Accuracy - 84.67%]
  - Random Forest [Classification Accuracy – 85.03%]
  - Deep Sequential [Classification Accuracy – 84.04%]

Best Model Performance Analysis

- Random Forest Model outperforms all other classifiers :

Logistic Regression Classification Report

	precision	recall	f1-score	support
Not Rain	0.87	0.95	0.91	15442
Rain	0.72	0.47	0.56	4266
accuracy			0.84	19708
macro avg	0.79	0.71	0.74	19708
weighted avg	0.83	0.84	0.83	19708

Fig. 7. Classification Report for Logistic Regression classifier.

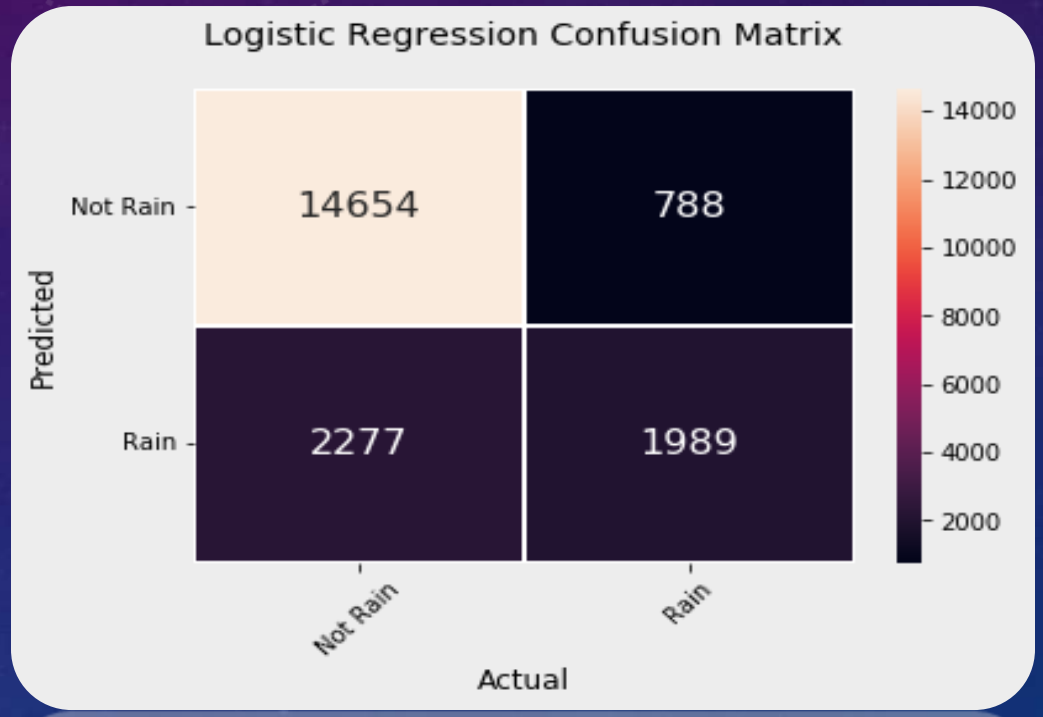

Fig. 8. Confusion matrix for Logistic Regression classifier.

HYPERPARAMETER OPTIMIZATION

- Randomized Search method has been used to find the best hyperparameters.
- Random Forest classifier achieved better validation accuracy.
- Random Forest model has been trained with the hyperparameters having optimized values.
- The model achieved **85.38%** prediction accuracy.
- Hyperparameter optimization of Random Forest Classifier helps to produce better performance.

FINAL PREDICTIVE ANALYSIS

- Weather Prediction on Unknown Data
  - Random Forest classifier predicts future weather conditions.
  - Classifier analyzed the features from the validation dataset.
  - Prepared a new dataset with the predicted the outcome for the RainTomorrow attribute.

CONCLUSION

- The project work represents machine learning and deep learning based approach for weather forecasting.
- Main challenge was to pre-process the datasets having huge null values.
- During Exploratory Data Analysis several insights were extracted.
- Important features were identified using correlation analysis.
- Machine Learning and Deep Learning classifiers used to predict the weather condition.
- Random Forest model has been selected as the best performing model.

FUTURE SCOPE

- Advanced feature selection techniques will be applied to improve the outcome.
- Hyperparameter optimization will help to improve the performance of Deep Learning classifier.
- Complex Deep Learning-based approach can be used to improve the predictive analysis.

Thank You