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

**Table of Contents**

[1. Introduction 3](#_Toc96225378)

[1.1. Background 3](#_Toc96225379)

[1.2. Research Purpose 3](#_Toc96225380)

[1.3. Report Structure 4](#_Toc96225381)

[2. Motivation & Objectives 4](#_Toc96225382)

[2.1. Research Motivation 4](#_Toc96225383)

[2.2. Research Objectives 4](#_Toc96225384)

[3. Literature Review 5](#_Toc96225385)

[3.1. Machine Learning Approach 5](#_Toc96225386)

[3.2. Deep Hybrid Approach 5](#_Toc96225387)

[4. Data Processing 6](#_Toc96225388)

[4.1. Data Collection 6](#_Toc96225389)

[4.2. Data Description 6](#_Toc96225390)

[4.3. Null Analysis 7](#_Toc96225391)

[4.4. Data Pre-processing 8](#_Toc96225392)

[5. Data Exploration 9](#_Toc96225393)

[5.1. Feature Distribution 9](#_Toc96225394)

[5.1.1. Maximum and Minimum Temperature Distribution 9](#_Toc96225395)

[5.1.2. Wind Gust Distribution 9](#_Toc96225396)

[5.1.3. Wind Speed Distribution 10](#_Toc96225397)

[5.1.4. Humidity Distribution 11](#_Toc96225398)

[5.1.5. Pressure Distribution 11](#_Toc96225399)

[5.1.6. Cloud Distribution 12](#_Toc96225400)

[5.1.7. Temperature Distribution 12](#_Toc96225401)

[5.2. Exploratory Data Analysis 13](#_Toc96225402)

[5.2.1. Average Wind Speed Analysis 13](#_Toc96225403)

[5.2.2. Average Humidity Analysis 14](#_Toc96225404)

[5.2.3. Average Pressure Analysis 15](#_Toc96225405)

[5.2.4. Average Temperature Analysis 16](#_Toc96225406)

[6. Predictive Analysis 17](#_Toc96225407)

[6.1. Feature Extraction 17](#_Toc96225408)

[6.2. Correlation Analysis 18](#_Toc96225409)

[6.3. Training, Testing and Validation Data 20](#_Toc96225410)

[6.4. Machine Learning Classifiers 20](#_Toc96225411)

[6.4.1. Logistic Regression 20](#_Toc96225412)

[6.4.2. K Nearest Neighbors 21](#_Toc96225413)

[6.4.3. Decision Tree 22](#_Toc96225414)

[6.4.4. AdaBoost Tree 23](#_Toc96225415)

[6.4.5. Random Forest 24](#_Toc96225416)

[6.5. Weather Prediction on Unknown Data 25](#_Toc96225417)

[7. Conclusion & Future Scope 25](#_Toc96225418)

[References 27](#_Toc96225419)

# 1. Introduction

## 1.1. Background

In this modern era of technology, everybody wants to start their day with a good plan following which they move towards their goals. Most of the people often used to check the weather report before preparing their daily plan as several times the challenging weather condition makes life miserable. Unsurprisingly it can be said that the weather conditions are analyzed in everyday life generated using the forecasting method based on the historical data. Machine learning algorithm-based data mining techniques allow predicting the future trend from the historical data regarding any particular event. However different parameters have been considered to predict the weather in upcoming days. In order to predict the weather different mechanisms can be applied including the applications of mathematical integration for post-processing of renewable energy and machine learning methods for testing the data (Haupt et al. 2018). Olaiya and Adeyemo (2012) have trained the neural network and decision tree classifiers on weather data of Ibadan, Nigeria from 2000 to 2009 to predict the future trend of the wind speed, evaporation, radiation, rainfall, minimum and maximum temperature. During these research, the different machine learning models have been compared with each other based on the same feature set. A comparative analysis between the deep learning and machine learning classifiers represents that the deep learning models slightly outperform different machine learning classifiers to classify the weather condition for predicting the weather (Schultz et al. 2021). However, it has been proved that the performance of the models main depends on the relevant feature selection based on their correlations. The weather forecasting classifier needs to be trained based on the related features from the respective dataset.

## 1.2. Research Purpose

The report represents the machine learning-based predictive analysis to predict the weather of a day ahead based on the given data. The proposed system will help to predict the future trend of the weather considering the historical data. Through this model, the probability of the rainfall can be analyzed so that all the people can make their plans without any confusion. The system will also help to monitor the other parameters for which the chances of rainfall become affected.

## 1.3. Report Structure

The report has been structured as follows –

* Section 2 contains the objectives and the motivations of the research.
* The detailed literature review has been presented in section 3.
* Section 4 represents the data collection, description, and pre-processing.
* The feature distribution and exploratory data analysis have been presented in Section 5.
* Section 6 represents the predictive analysis based on different machine learning algorithms.
* Finally, Section 7 represents the conclusion and future scope of the project considering the findings throughout the research work.

# 2. Motivation & Objectives

## 2.1. Research Motivation

The main motivation behind this work is to build an intelligent system to predict future weather conditions based on the given data. The system will further help the people to understand the upcoming rainy days so that they can plan their weeks accordingly. This will help them to manage their work and maintain balance in life. If they have advanced information that the day after tomorrow will be a rainy day, then they will try to finish their important work within specified day for which they need to go outside otherwise they will do their work after that rainy day.

## 2.2. Research Objectives

The research work has several objectives from which some major objectives are as follows –

* Exploratory data analysis will be performed to find the location-wise average trend of wind speed, humidity, pressure, maximum temperature, minimum temperature rainfall.
* Correlation analysis will help to select the relevant features from the weather data.
* Predictive analysis will help to predict the future rainy days from the weather data. Different machine learning classifiers have been proposed to train the data based on the extracted features.
* Comparative analysis will be performed to find the different trained classifiers to select the best classifier.
* The best classifier will be used to predict future rainy days based on the unknown weather data.

# 3. Literature Review

## 3.1. Machine Learning Approach

According to the research work conducted by Merenti-Välimäki and Laininen (2002), the selection of best feature variables can affect the model performance during the analysis. During their research, they have used the logistic regression model for the classification of the weather codes based on the different atmospheric features. Finally, the model used a monotonic regression function to successfully classify the weather type variables. Holmstrom, Liu and Vo (2016) developed a predictive system to forecast the weather for each day from 2011 to 2015 from Stanford, California. During the feature selection phase, the researchers considered maximum and minimum temperature along with the mean of humidity and pressure. Three regression models were used to predict the temperatures based on the linear relation between the selected features. During the analysis, they found that the professional forecasting model achieved better prediction accuracy than other models. A research work represents the methodology for short term weather forecasting by selecting the features based on their correlation (Moon and Kim 2020). The researchers used the weather data of different locations in South Korea collected from the European Centre for Medium-Range Weather Forecasts. During the research, the coefficient of the multinomial regression was extracted. In order to predict the precipitation type, the authors used these coefficients. Their model achieved better accuracy compared to the other models of the respective data.

## 3.2. Deep Hybrid Approach

Another research work represents the Bayesian Enhanced Approach (BEA) to prepare the time-series weather predictor based on Artificial Neural Network (Rivero et al. 2016). In this research, the time series were considered based on the short- and long-term dependencies to forecast the rainfall. The researchers compared their proposed model with different predictors selected from the literature based on their performance and found the BEA method outperforms several algorithms. Grover, Kapoor and Horvitz (2015) proposed a deep hybrid model to forecast the weather from the historical data based on the feature set variables e.g., temperature, pressure, winds and dew points. The data-centric kernel was developed which can incorporate over space considering the turbulence. The long- and short-term features were used to conduct the analysis using a gradient-tree model. Their deep hybrid network achieved better results compared to NOAA (National Oceanic and Atmospheric Administration) weather forecast.

# 4. Data Processing

## 4.1. Data Collection

In order to prepare the dataset, the data has been collected based on the weather records of different cities in Australia along with atmospheric parameters. This dataset has been available at Kaggle, one of the largest data repositories. The dataset has been downloaded from the website to perform the experiments on it. There are mainly two weather datasets i.e., the weather and unknown data. Both of the datasets contain location-wise weather data but the dataset of weather data has an additional attribute to present whether tomorrow will be rain or not. In this research, the machine learning-based approach will be used to predict the values for that attribute of the dataset of unknown data.

## 4.2. Data Description

In this section, the detailed description of the dataset along with the attribute names and corresponding data types have been presented. The data description will help to understand the characteristics of the data for extracting important insights from it. During the analysis, the purpose and appropriateness of each attribute has been analyzed for further experimentation. Table 1 represents the data description of the weather dataset.

Table 1. Data description of weather data.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Data Type | Description |
| Location | Object | Name of the city from Australia. |
| MinTemp | Float | The Minimum temperature during a particular day. (degree Celsius) |
| MaxTemp | Float | The maximum temperature during a particular day. (degree Celsius) |
| Rainfall | Float | Rainfall during a particular day. (millimeters) |
| Evaporation | Float | Evaporation during a particular day. (millimeters) |
| Sunshine | Float | Bright sunshine during a particular day. (hours) |
| WindGusDir | Object | The direction of the strongest gust during a particular day. (16 compass points) |
| WindGuSpeed | Float | Speed of strongest gust during a particular day. (kilometers per hour) |
| WindDir9am | Object | The direction of the wind for 10 min prior to 9 am. (compass points) |
| WindDir3pm | Object | The direction of the wind for 10 min prior to 3 pm. (compass points) |
| WindSpeed9am | Float | Speed of the wind for 10 min prior to 9 am. (kilometers per hour) |
| WindSpeed3pm | Float | Speed of the wind for 10 min prior to 3 pm. (kilometers per hour) |
| Humidity9am | Float | The humidity of the wind at 9 am. (percent) |
| Humidity3pm | Float | The humidity of the wind at 3 pm. (percent) |
| Pressure9am | Float | Atmospheric pressure at 9 am. (hectopascals) |
| Pressure3pm | Float | Atmospheric pressure at 3 pm. (hectopascals) |
| Cloud9am | Float | Cloud-obscured portions of the sky at 9 am. (eighths) |
| Cloud3pm | Float | Cloud-obscured portions of the sky at 3 pm. (eighths) |
| Temp9am | Float | The temperature at 9 am. (degree Celsius) |
| Temp3pm | Float | The temperature at 3 pm. (degree Celsius) |
| RainToday | Object | If today is rainy then ‘Yes’. If today is not rainy then ‘No’. |
| RainTomorrow | Integer | If tomorrow is rainy then 1 (Yes). If tomorrow is not rainy then 0 (No). |

## 4.3. Null Analysis

According to the initial observation, the dataset contains several null values which will affect the experiments as these values have no analytical meaning. During the analysis total number of null values for each attribute have been evaluated. Table 2 represent the null value analysis of both the weather and unknown datasets.

Table 2. Null analysis of weather and unknown data.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Weather Data | Unknown Data |
| Location | 0 | 0 |
| MinTemp | 443 | 194 |
| MaxTemp | 230 | 92 |
| Rainfall | 979 | 427 |
| Evaporation | 42531 | 18312 |
| Sunshine | 47317 | 20499 |
| WindGusDir | 6521 | 2809 |
| WindGuSpeed | 6480 | 2790 |
| WindDir9am | 7006 | 3007 |
| WindDir3pm | 2648 | 1130 |
| WindSpeed9am | 935 | 413 |
| WindSpeed3pm | 1835 | 795 |
| Humidity9am | 1233 | 541 |
| Humidity3pm | 2506 | 1104 |
| Pressure9am | 9748 | 4266 |
| Pressure3pm | 9736 | 4245 |
| Cloud9am | 37572 | 16085 |
| Cloud3pm | 40002 | 17092 |
| Temp9am | 614 | 290 |
| Temp3pm | 1904 | 822 |
| RainToday | 979 | 427 |
| RainTomorrow | 0 | - |

## 4.4. Data Pre-processing

In this section, the step-by-step process for the pre-processing of the data has been mentioned.

1. In order to fill the null values of the numerical columns, the mean of corresponding values has been calculated and the empty fields will be filled using the mean values.
2. The empty character fields have been filled with the most frequent values of the respective attributes (except RainToday).
3. The records having the empty values for the RainToday attribute have been dropped from the dataset since it is quite difficult to decide whether the day was rainy or not.

The cleaned and pre-processed datasets will be used for data explorations for further analysis.

# 5. Data Exploration

## 5.1. Feature Distribution

In this section, the data has been analyzed to identify the distribution of each of the feature variables from the dataset. This analysis provides a brief idea about all the features along with the frequency of the corresponding values. The value with the highest frequency has occurred for maximum time in the respective attribute from the database and vice versa.

### 5.1.1. Maximum and Minimum Temperature Distribution

In this section, the distribution of minimum and maximum temperature has been analyzed to identify the temperature with the highest frequency. Figure 1 and Figure 2 represent the distribution of the minimum and maximum temperature on the basis of the RainToday values.

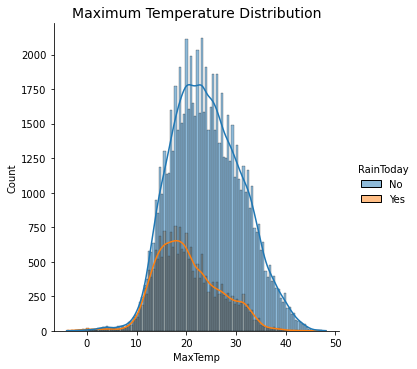
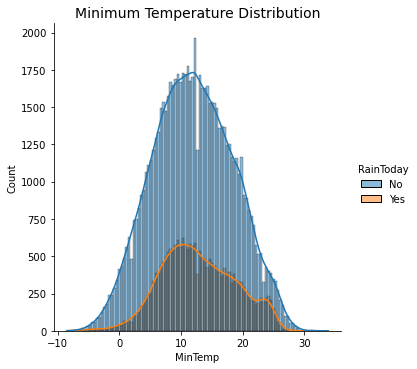


Fig. 1. Distribution of minimum temperature. Fig. 2. Distribution of minimum temperature.

During the analysis it has been found that the range of minimum temperature is from -8.5oC to 33.9oC and 11oC minimum temperature has the highest frequency in the dataset. On the other hand, the range of maximum temperature is from -4.1oC to 48.10C and 20oC maximum temperature has the highest frequency in the dataset.

### 5.1.2. Wind Gust Distribution

In this section, the distribution of strong wind gusts has been analyzed based. In Figure 3, the wind gust distribution has been presented based on the RainToday values.

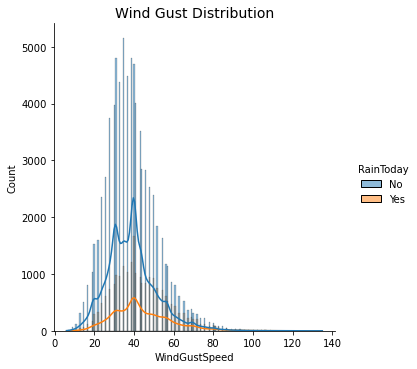


Fig. 3. Distribution of wind gusts.

During the analysis, it has been found that the range of strong wind gusts is from 6 compass points to 135 compass points and 39.98 compass points of wind gusts have the highest frequency in the dataset.

### 5.1.3. Wind Speed Distribution

Here the distribution of wind speed has been analyzed based on two different times in a particular day. Figure 4 and Figure 5 represent the distribution of the wind speed at 9 am and 3 pm.

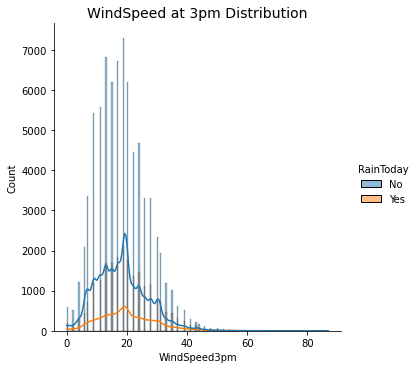
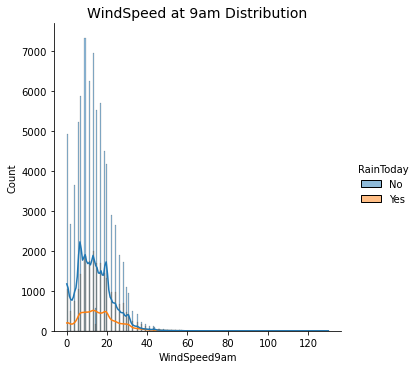


Fig. 4. Distribution of wind speed at 9 am. Fig. 5. Distribution of wind speed at 3 pm.

During the analysis, it has been found that the range of wind speed at 9 am is from 0 kmph to 130 kmph and 9 kmph of wind speed has the highest frequency in the dataset. On the other hand, the range of wind speed at 3 pm is from 0 kmph to 87 kmph and 17 kmph of wind speed has the highest frequency in the dataset.

### 5.1.4. Humidity Distribution

In this section, the distribution of humidity has been analyzed based of two different times in a particular day. Figure 6 and Figure 7 represent the distribution of the humidity at 9 am and 3 pm.

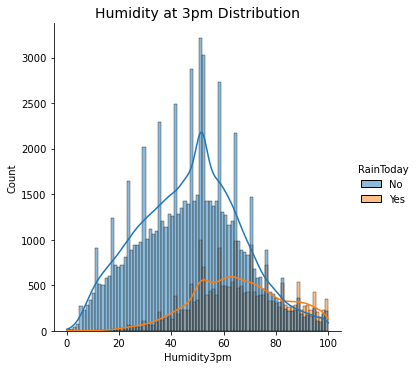
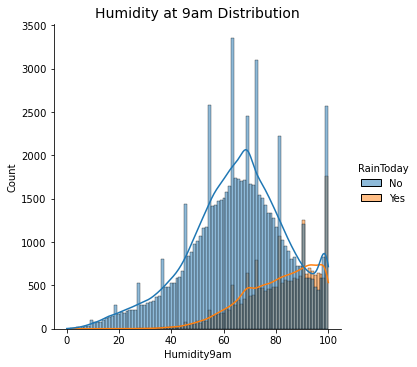


Fig. 6. Distribution of humidity at 9 am. Fig. 7. Distribution of humidity at 3 pm.

During the analysis, it has been found that the range of humidity at 9 am and 3 pm is from 0% to 100% and 99% of humidity at 9 am has the highest frequency in the dataset. On the other hand, 54.43% of humidity at 3 pm has the highest frequency in the dataset.

### 5.1.5. Pressure Distribution

In this section, the distribution of pressure has been analyzed based of two different times in a particular day. Figure 8 and Figure 9 represent the distribution of the pressure at 9 am and 3 pm.

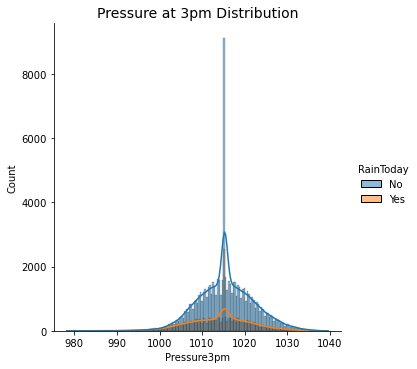
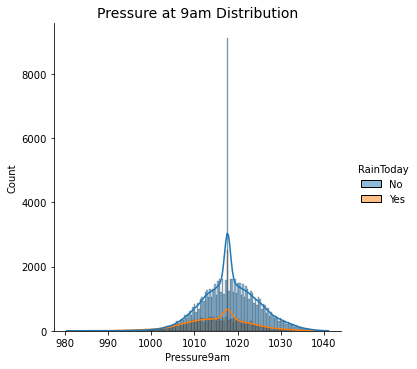


Fig. 8. Distribution of pressure at 9 am. Fig. 9. Distribution of pressure at 3 pm.

During the analysis, it has been found that the range of wind pressure at 9 am is from 980.5 hPa to 1042 hPa and 1017.68 hPa of pressure has the highest frequency in the dataset. On the other hand, the range of pressure at 3 pm is from 978.2 hPa to 1039.6 hPa and 1015.28 hPa of pressure has the highest frequency in the dataset.

### 5.1.6. Cloud Distribution

In this section, the distribution of cloud has been analyzed based of two different times in a particular day. Figure 10 and Figure 11 represent the distribution of the cloud at 9 am and 3 pm.

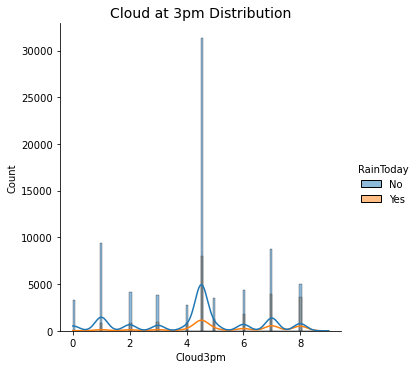
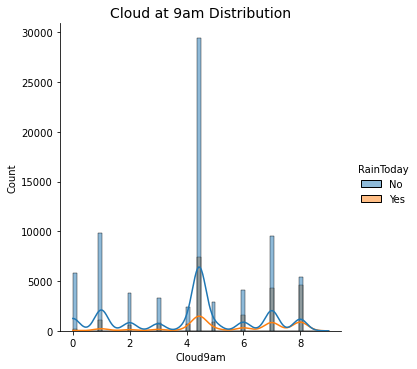


Fig. 10. Distribution of cloud at 9 am. Fig. 11. Distribution of cloud at 3 pm.

During the analysis, it has been found that the range of cloud at 9 am and 3 pm is from 0 eighths to 9 eighths and 4.44 eighths of cloud at 9 am has the highest frequency in the dataset. On the other hand, 4.52 eighths of cloud at 3 pm has the highest frequency in the dataset.

### 5.1.7. Temperature Distribution

In this section, the distribution of temperature has been analyzed based of two different times in a particular day. Figure 12 and Figure 13 represent the distribution of the temperature at 9 am and 3 pm.

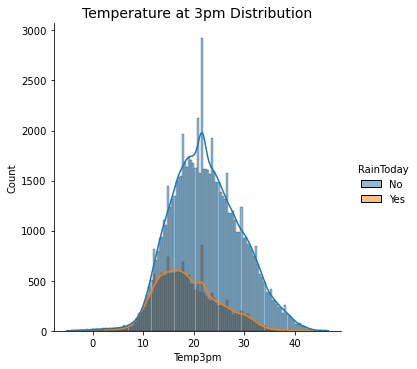
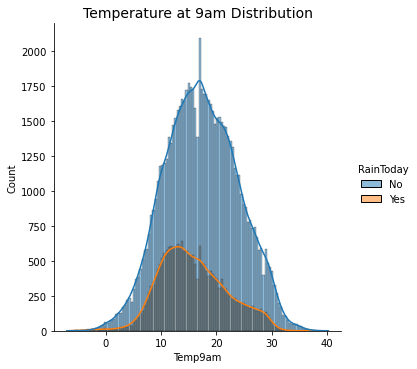


Fig. 12. Distribution of temperature at 9 am. Fig. 13. Distribution of temperature at 3 pm.

During the analysis, it has been found that the range of wind temperature at 9 am is from -7oC to 40.2oC and 17oC of temperature has the highest frequency in the dataset. On the other hand, the range of pressure at 3 pm is from -5.1oC to 46.7oC and 27.68oC of temperature has the highest frequency in the dataset.

## 5.2. Exploratory Data Analysis

In this phase of analysis, the average trends of several features have been analyzed based on different locations along with their comparisons. This analysis will help to identify how the values of these features have changed over the locations. Although from this analysis, it can be identified how the changes in different features affect the overall temperature of the corresponding locations.

### 5.2.1. Average Wind Speed Analysis

During this analysis, the average wind speed has been analyzed based on the different locations in Australia at two different times during a particular day. Figure 14 represents the line plot to visualize the location-wise trend of wind speed.

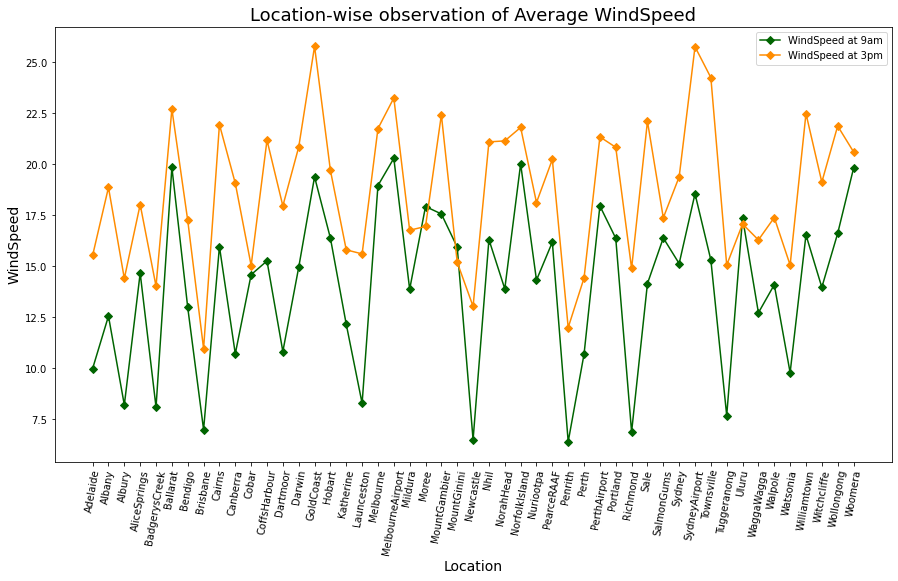


Fig. 14. Location-wise observation of average wind speed.

During this analysis, it has been found that the wind speed of the Melbourne Airport at 9 am has the highest wind speed at 20.29 kmph. On the other hand, at 3 pm the Gold Coast of Australia has the highest wind speed at 25.77 kmph. Finally, it can be concluded that the wind speed at 3 pm is much higher than the wind speed at 9 am.

### 5.2.2. Average Humidity Analysis

During this analysis, the average humidity has been analyzed based on the different locations in Australia for two different times in a particular day. Figure 15 represents the bar plot to visualize the location-wise trend of humidity.

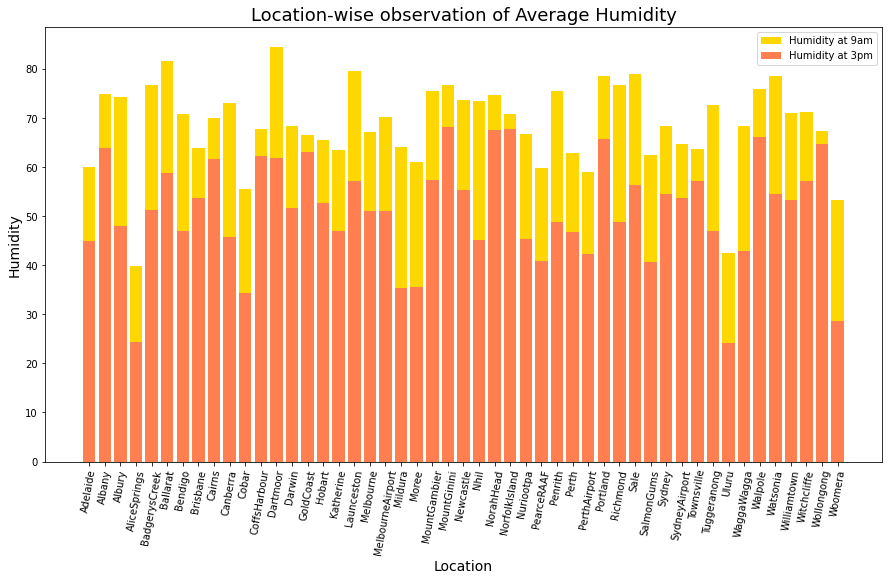


Fig. 15. Location-wise observation of average humidity.

During this analysis, it has been found that the humidity of the Dartmoor at 9 am has the highest humidity as 84.38%. On the other hand, at 3 pm the MountGinini of Australia has the highest humidity as 68.24%. Finally, it can be concluded that humidity at 9 am is much higher than the wind speed at 3 pm.

### 5.2.3. Average Pressure Analysis

During this analysis, the average pressure has been analyzed based on the different locations in Australia for two different times in a particular day. Figure 16 represents the line plot to visualize the location-wise trend of pressure.

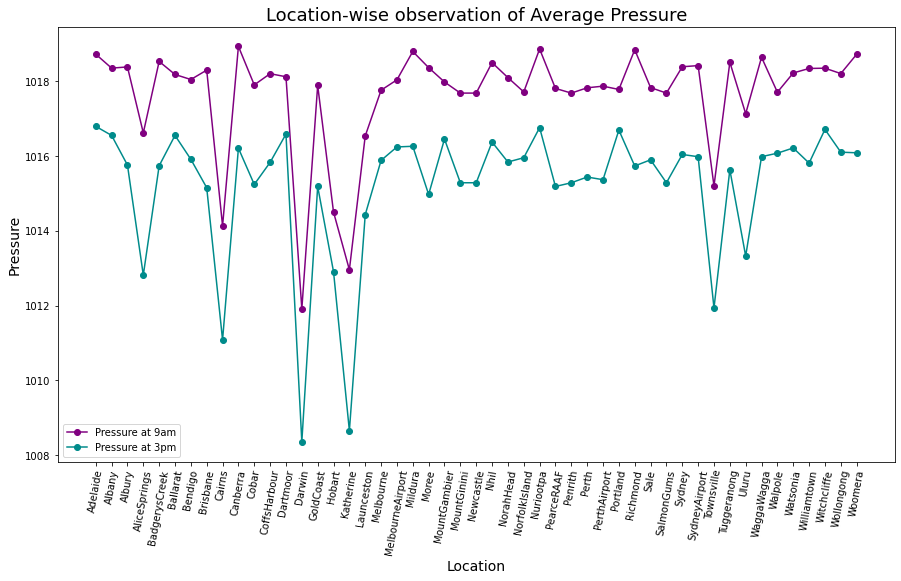


Fig. 16. Location-wise observation of average pressure.

During this analysis, it has been found that the pressure of Canberra at 9 am has the highest pressure as 1018.93 hPa. On the other hand, at 3 pm the Adelaide of Australia has the highest pressure as 1016.79 hPa. Finally, it can be concluded that pressure at 9 am is much higher than the wind speed at 3 pm.

### 5.2.4. Average Temperature Analysis

During this analysis, the average temperature has been analyzed based on the different locations in Australia for two different time on a particular day. Figure 17 represents the line plot to visualize the location-wise trend of temperature.

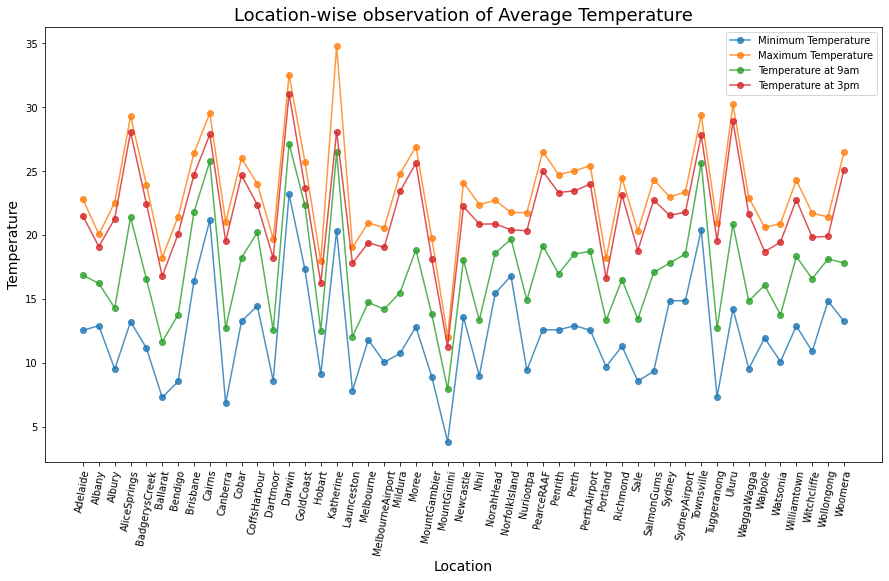


Fig. 17. Location-wise observation of average temperature.

# 6. Predictive Analysis

In this section, the machine learning classifiers have been developed to forecast the weather condition based on the selected features. The correlations between the feature variables have been analyzed to understand their relevance and identify the best features for the predictive analysis. The classifier developed using the best features will always produce a better outcome compared to others. During the analysis, the character attributes of the dataset have been labelled as encoded and divided into numerous columns using One-hot encoding. Finally, some popular machine learning classifiers have been introduced to learn the knowledge from the weather dataset and predict the rainy days from the unknown dataset by analyzing the given feature set.

## 6.1. Feature Extraction

In order to extract the features from the dataset initially the numerical and character attributes have been identified. The character variables cannot be used as the input to the machine learning classifiers so these variables need to be converted into numerical variables. The identified character variables are – Location, WindGustDir, WindDir9am, WindDir3pm and RainToday. In these variables, the RainToday have binary values (Yes and No) so this variable can be converted into 0 (No) and 1 (Yes). Other variables need to be labelled as encoded using one-hot encoding so that the final feature set has only contained the numerical data. During one-hot encoding new columns are created to equal the unique value of an attribute. The same approach has been applied to both the weather and unknown datasets to generate the final datasets.

## 6.2. Correlation Analysis

The correlation has been analyzed between the numerical data from both of the datasets. Correlation defines the statistical measure of the linear relationship between two variables (Lindley 1990). There are three types of correlations i.e., positive, negative and neutral. Figure 18 represents the heatmap for the correlation matrix of the weather data.

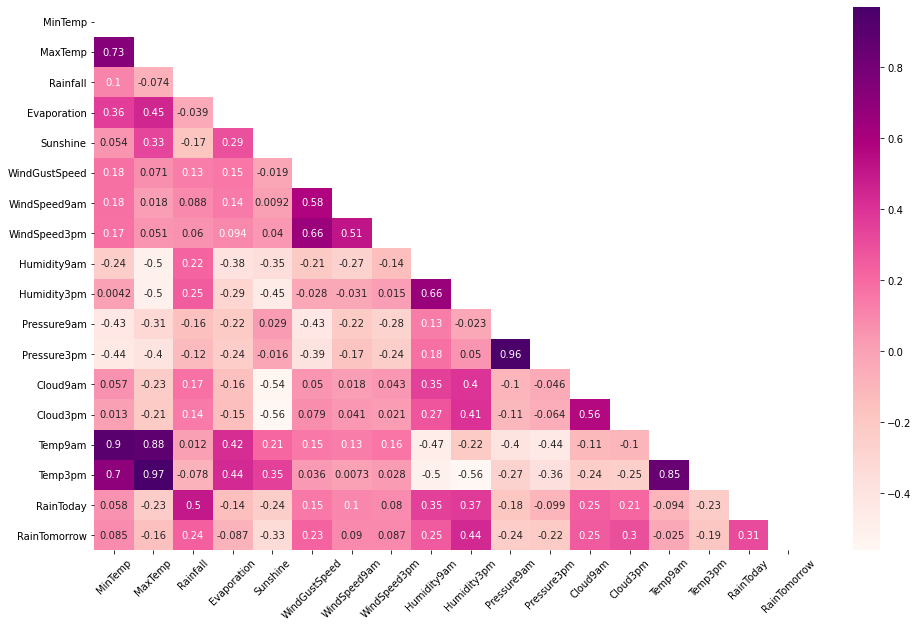


Fig. 18. Correlation matrix for the numerical features of weather dataset.

The numerical features from the unknown data have been considered for this analysis. Figure 19 represents the heatmap for the correlation matrix of the weather data.

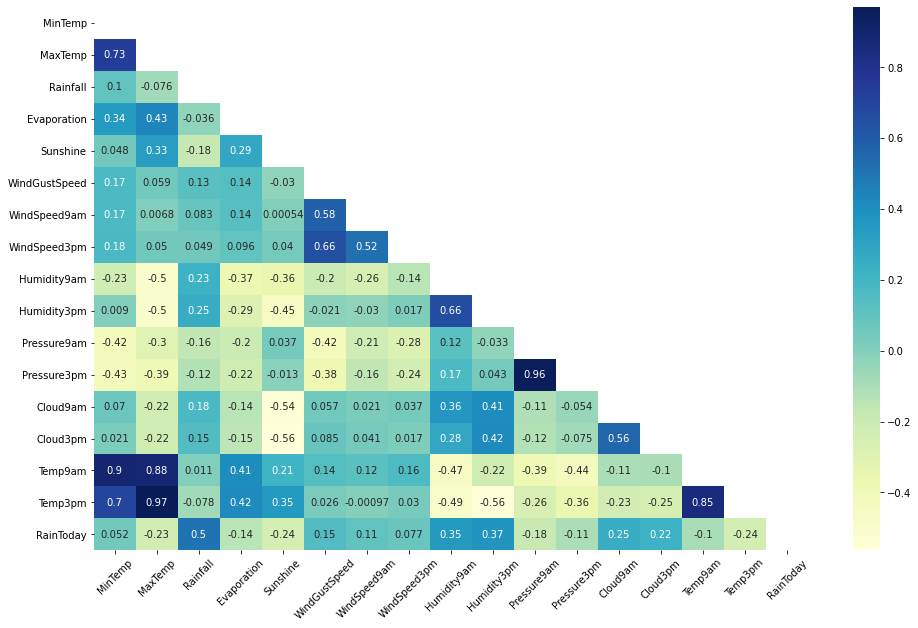


Fig. 19. Correlation matrix for the numerical features of the unknown dataset.

During the analysis, some good correlations have been found from both of the datasets. Table 3 represents the features along with their correlational measures.

Table 3. Correlation Analysis.

|  |  |  |
| --- | --- | --- |
| Feature Selection | Weather Dataset | Unknown Dataset |
| MinTemp vs. MaxTemp | 0.73 | 0.73 |
| WindSpeed9am vs. WindGustSpeed | 0.58 | 0.58 |
| WindSpeed3pm vs. WindGustSpeed | 0.66 | 0.66 |
| WindSpeed3pm vs. WindSpeed9am | 0.51 | 0.52 |
| Humidity3pm vs. Humidity9am | 0.66 | 0.66 |
| Presure3pm vs. Presure9am | 0.96 | 0.96 |
| Cloud3pm vs. Cloud9am | 0.56 | 0.56 |
| Temp9am vs. MinTemp | 0.9 | 0.9 |
| Temp9am vs. MaxTemp | 0.88 | 0.88 |
| Temp3pm vs. MinTemp | 0.7 | 0.7 |
| Temp3pm vs. MaxTemp | 0.97 | 0.97 |
| Temp3pm vs. Temp9am | 0.85 | 0.85 |
| RainToday vs. Rainfall | 0.5 | 0.5 |

## 6.3. Training, Testing and Validation Data

The weather data has been divided into X and y sets based on the feature variables and the target variable. The X set contains all the independent variables whereas the y set contains the dependent variable “RainTomorrow”. The X and y sets have been divided into 80:20 ratio i.e. 80% of data can be used for training purpose and 20% of data can be used for testing purpose (Jordan and Mitchell 2015). The unknown data has been considered as the validation set which will be used for validating the trained classifiers.

## 6.4. Machine Learning Classifiers

The machine learning-based classifiers have been trained based on the training data and tested the performance using the test data. The best classifier will be selected on the basis of their classification performance. Finally, the best performing classifier has been used on the validation data to forecast the rainy days based on the feature set. The sklearn[[1]](#footnote-1) library has been used to import these classifiers. The models have been trained using the X\_train and y\_train sets and tested on the X\_test data for predicting whether the day will be rainy or not.

### 6.4.1. Logistic Regression

Logistic Regression is the most popular supervised machine learning algorithm. This classification model mainly classifies the dependent variable “RainTomorrow” based on the feature set of independent variables. In this scenario, the RainTomorrow attribute is a binary attribute so the binomial logistic regression algorithm will be used for preparing the classifier. The model achieved 83.97% prediction accuracy on the test data. In Figure 20 and Figure 21, the classification report and the confusion matrix have been presented for the Logistic Regression classifier.

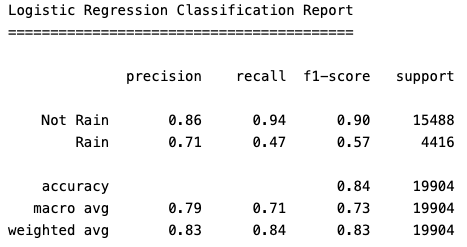


Fig. 20. Classification Report for Logistic Regression classifier.

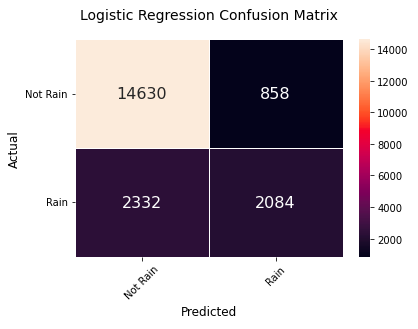


Fig. 21. Confusion matrix for Logistic Regression classifier.

### 6.4.2. K Nearest Neighbors

K Nearest Neighbors is a supervised machine learning algorithm that mainly identifies the similarity between the new and the available data points to categorize them into the most similar classes. This non-parametric algorithm can be used for regression as well as for classification tasks but it is widely used to perform classification tasks. During the experimentation, this classifier has been used to predict the chance to be rain based on the feature variables. This algorithm performs slower than others because it learns the knowledge during the classification on the test set. The model achieved 82.96% prediction accuracy on the test data. In Figure 22 and Figure 23 the classification report and the confusion matrix have been presented for the KNN classifier.

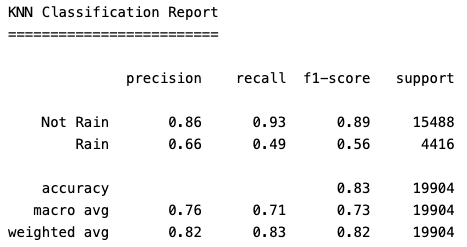


Fig. 22. Classification Report for KNN classifier.

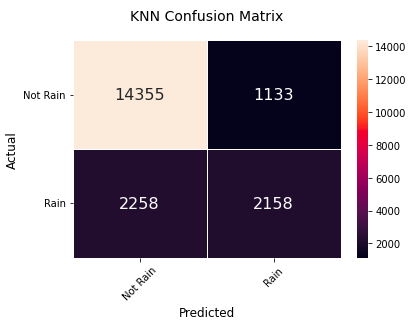


Fig. 23. Confusion matrix for KNN classifier.

### 6.4.3. Decision Tree

Decision Tree is a supervised machine learning algorithm mainly used for regression and classification tasks. This classifier has been used to classify the RainTomorrow attribute based on different atmospheric features. It divides a problem into several small problems to generate a tree structure based on the decisions. Then it will traverse all the paths of the tree to find the solution based on the decision rules. The algorithm performs the classification tasks faster but the outcomes may be erroneous. However, the model achieved 77.59% prediction accuracy on the test data. In Figure 24 and Figure 25 the classification report and the confusion matrix have been presented for the Decision Tree classifier.

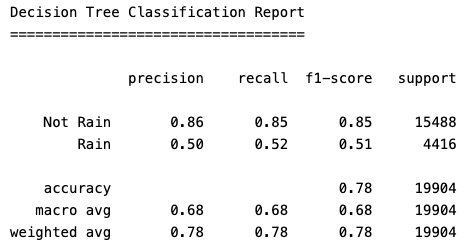


Fig. 24. Classification Report for Decision Tree classifier.

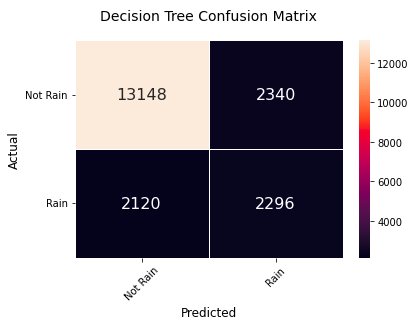


Fig. 25. Confusion matrix for Decision Tree classifier.

### 6.4.4. AdaBoost Tree

The AdaBoost is an adaptive boosting technique mainly used as an ensemble method in machine learning. In order to improve the performance of other models, this tree-based linear model is used. The model used the random probability method to perform the classification of the RainTomorrow attribute. This model generates the single level decision tree with a single classified decision. The model achieved 84.42% prediction accuracy on the test data. In Figure 26 and Figure 27, the classification report and the confusion matrix have been presented for the AdaBoost classifier.

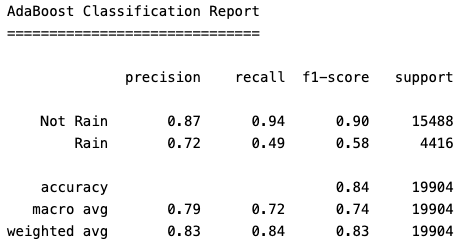


Fig. 26. Classification Report for AdaBoost classifier.

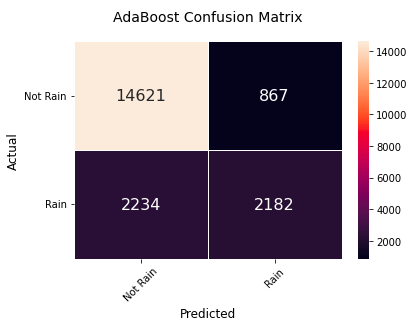


Fig. 27. Confusion matrix for AdaBoost classifier.

### 6.4.5. Random Forest

Random Forest is a supervised machine learning technique that can be used for both regression and classification tasks. This is also an ensemble model which is developed to solve the complex problem based on the combined principle of different classifiers. This classifier will take less time to classify the RainTomorrow variable by creating several decision trees based on the selected feature set. The model achieved 84.66% prediction accuracy on the test data. In Figure 28 and Figure 29, the classification report and the confusion matrix have been presented for the AdaBoost classifier.

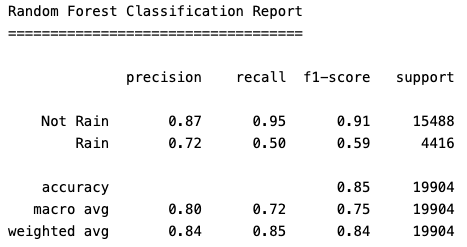


Fig. 28. Classification Report for Random Forest classifier.

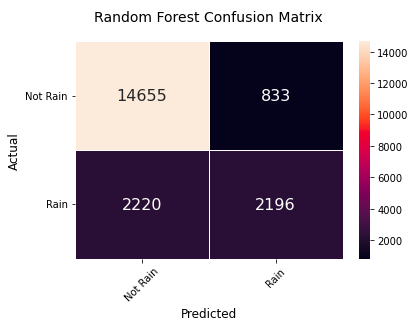


Fig. 29. Confusion matrix for Random Forest classifier.

## 6.5. Weather Prediction on Unknown Data

During the analysis, it has been found that the Random Forest classifier has achieved the highest classification accuracy on the test data. The model has been used to predict future weather conditions from unknown weather data. In order to perform the predictive analysis on the unknown dataset, the trained Random Forest classifier takes the features from the validation dataset. The model has been predicted the outcome for the RainTomorrow attribute and prepared a new dataset with the final predicted outcome.

# 7. Conclusion & Future Scope

In this report, the weather data from different locations in Australia was analyzed to forecast the future weather condition using machine learning-based classifier. The main challenge was to clean and preprocess the dataset with a huge number of null values in various attributes. Although that problem was resolved by using the central tendency theory. The mean values were calculated for the numerical attributes and to remove the null values from character attributes the value having the highest frequency was considered. During the exploratory data analysis, several insights were extracted to identify the important features for predictive analysis. The distribution of the features helped to analyze the frequency of the values for each attribute. The average trend for wind speed, humidity, pressure and temperature were computed based on different locations and time periods. The important features have been identified during the correlation analysis based on a correlational score between two variables. This research work will help to understand the usage of several machine learning classifiers along with their working principles. In order to perform the experiments, different machine learning models were introduced to classify the RainTomorrow attribute based on the feature variables. Among all the classifiers the Random Forest classifier achieved the highest classification accuracy on the test data and is considered as the best classifier. Finally, the Random Forest classifier was used on the unknown data to predict whether tomorrow will be rain or not. After successful prediction, a new dataset was generated with the predicted outcomes.

In future, this analysis can be improved by applying advanced feature selection techniques so that the classifiers will produce more accurate outcomes. Furthermore, the hyperparameter optimization will help to improve the proposed classifiers to perform well during the classification. Although the deep learning-based approach can be used to improve the quality of the predicted outcome.

# References

Grover, A., Kapoor, A. and Horvitz, E., 2015, August. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 379-386).

Haupt, S.E., Cowie, J., Linden, S., McCandless, T., Kosovic, B. and Alessandrini, S., 2018, October. Machine learning for applied weather prediction. In *2018 IEEE 14th international conference on e-science (e-Science)* (pp. 276-277). IEEE.

Holmstrom, M., Liu, D. and Vo, C., 2016. Machine learning applied to weather forecasting. *Meteorol. Appl*, pp.1-5.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), pp.255-260.

Lindley, D.V., 1990. Regression and correlation analysis. In *Time series and statistics* (pp. 237-243). Palgrave Macmillan, London.

Merenti-Välimäki, H.L. and Laininen, P., 2002. Analysing effects of meteorological variables on weather codes by logistic regression. *Meteorological Applications*, *9*(2), pp.191-197.

Moon, S.H. and Kim, Y.H., 2020. An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression. *Atmospheric Research*, *240*, p.104928.

Olaiya, F. and Adeyemo, A.B., 2012. Application of data mining techniques in weather prediction and climate change studies. *International Journal of Information Engineering and Electronic Business*, *4*(1), p.51.

Rivero, C.R., Patiño, D., Pucheta, J. and Sauchelli, V., 2016. A new approach for time series forecasting: bayesian enhanced by fractional brownian motion with application to rainfall series. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *7*(3).

Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A. and Stadtler, S., 2021. Can deep learning beat numerical weather prediction?. *Philosophical Transactions of the Royal Society A*, *379*(2194), p.20200097.

1. https://scikit-learn.org/ [↑](#footnote-ref-1)