Table of Contents

[1. Introduction 2](#_Toc155175237)

[2. Methodology 2](#_Toc155175238)

[2.1 Dataset 2](#_Toc155175239)

[2.2 Library 3](#_Toc155175240)

[2.3 Exploratory Data Analysis 3](#_Toc155175241)

[2.3.1 Dimension of Data 3](#_Toc155175242)

[2.3.2 Type of Data 3](#_Toc155175243)

[2.3.3 Descriptive Statistic 4](#_Toc155175244)

[2.3.4 Class Distribution 5](#_Toc155175245)

[2.3.5 Data Correlations 5](#_Toc155175246)

[2.3.6 Skewness 5](#_Toc155175247)

[2.4 Data Visualization 6](#_Toc155175248)

[2.4.1 Univariate Plots 6](#_Toc155175249)

[2.4.1.1 Histogram 6](#_Toc155175250)

[2.4.1.2 Density Plots 7](#_Toc155175251)

[2.4.2 Multivariate Plots 7](#_Toc155175252)

[2.4.2.1 Correlation Matrix Plot 7](#_Toc155175253)

[2.4.2.2 Scatter Plot Matrix 8](#_Toc155175254)

[2.5 Data Preprocessing 8](#_Toc155175255)

[2.5.1 Standardize Data 9](#_Toc155175256)

[2.6 Feature Selection 9](#_Toc155175257)

[2.6.1 Univariate Selection 10](#_Toc155175258)

[2.6.2 Recursive Feature Elimination (RFE) 10](#_Toc155175259)

[2.6.3 Feature Importance 10](#_Toc155175260)

[3. Conclusion 11](#_Toc155175261)

[4. Future Works 11](#_Toc155175262)

[References 11](#_Toc155175263)

# Introduction

Weather forecast can help people to make adaptation and mitigation for future plans (Schacher, 2022). This project will try to predict a weather based on several steps. First is analysing the dataset and visualizing the data to determine whether the data need a data transformation. Second is to transform the data using preprocessing method. Third is feature selection to choose which feature is important for the performance of the model. More steps will be done in the future works.

# Methodology

## 2.1 Dataset

The dataset used in this project is collected from Kaggle which is weather prediction dataset for Seattle. The dataset contains weather data from January 1, 2012 – December 31, 2015. This dataset will be used to predict between 5 weathers which are rain, drizzle, snow, sun, and fog by using the other aspects namely, precipitation, maximum temperature, minimum temperature, and wind speed. The ethical consideration for this case is that the data is collected from Seattle, and the result might be inaccurate for other area, therefore, it might be irrelevant for global used due to geographical issues, therefore, more data need to be collected from over the world, so it can be valid for other area as well.

A screenshot of a computer

Description automatically generated with low confidenceBelow is the first 20 rows of the dataset:

## Library

A screen shot of a computer

Description automatically generated

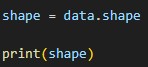
Here is the library needed to import all component that will be used for data analysation and visualization, preprocessing, and feature selection.

## 2.3 Exploratory Data Analysis

EDA is the step in this project to analyse and understand the data.

### 2.3.1 Dimension of Data

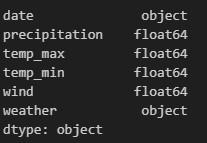
First is finding the dimension of the data to interpret the training time by looking at the rows and performance by looking at the column. Below is the code, followed by the result:



The dataset has 1461 rows and 6 columns.

### 2.3.2 Type of Data

Second is to find the data type of each attribute. Below is the code followed by the result:



A picture containing text, screenshot, font

Description automatically generatedThe data types for weather are string value, the type need to be changed into integer to make a classification by replacing weather with an integer from 1 – 5 using this code below:

A screenshot of a computer

Description automatically generated Below is the result showing the weather column is already integer.

### 2.3.3 Descriptive Statistic

Next, descriptive statistic will be performed to make better understanding about the data. Below is the code showing with the result of 3 decimal behind the number.

A screen shot of a computer code

Description automatically generated with low confidence

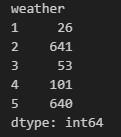
A picture containing text, screenshot, font, number

Description automatically generated

By looking at the result, the range between the value is vary, for instance precipitation from 0 – 55.9, temp\_max from -1.6 – 35, etc. This will need some preprocessing method to make the data in equal range.

### 2.3.4 Class Distribution

Because this is a classification problem, class distribution is needed to check how balanced is the data. Below is the code and the result.



The data is highly imbalanced as the difference between the weather is considerable. In order to get the best result later on, it need a resampling method which is oversampling to balance the data by making all attributes have the same value as the maximum value which is in this case is 641.

### 2.3.5 Data Correlations

Next step is to find the correlations between attributes using:

A picture containing text, screenshot, font

Description automatically generated

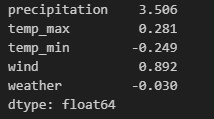
A picture containing text, screenshot, font, black

Description automatically generated

From the result, only temp\_max and temp\_min has a high correlation, the others are normal.

### 2.3.6 Skewness

Now, to identify whether the data has a gaussian distribution, skew method will be performed by using this code:



Precipitation has a very right skew because as shown in the descriptive statistic, the difference between the 75% and the maximum value is a lot from 2.8 to 55.9. This needs some pre-processing method to correct the skew.

## 2.4 Data Visualization

After doing the exploratory data analysis, the data need to be analysed more deeply by using a univariate plots and multivariate plots to get more information or insights from the data.

### 2.4.1 Univariate Plots

This technique is to get a better understanding of the data for each attributes independently.

### 2.4.1.1 Histogram

A picture containing text, font, screenshot, line

Description automatically generatedHistogram is the fastest way to tell whether the data have a gaussian distribution, exponential distribution, or skewed. The library that will be used is matplotlib. Below is the code followed by the histogram.

A picture containing diagram, text, plot, line

Description automatically generated

By looking at the histogram, attributes precipitation has an exponential distribution, and temp\_max, temp\_min, and wind may have nearly gaussian distribution. It is important to spot a gaussian distribution because many machine learning algorithms assume gaussian distribution for the input.

### 2.4.1.2 Density Plots

Density plots is the other method to visualize the data. It gives clearer view of the distribution of the data. It also utilises the matplotlib library. Below is the code followed by the result:

A picture containing text, diagram, line, plot

Description automatically generated

### 2.4.2 Multivariate Plots

This method will be used to show the correlation between each attribute in the dataset.

### 2.4.2.1 Correlation Matrix Plot

Plotting with correlation matrix plot can be useful to spot a highly correlated variables with the other variables. Assuming this project will use a logistic or linear regression later on, a highly correlated variable can impact the performance of the model and result in a poor performance. Here is the code below and the result:

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, screenshot, square

Description automatically generated

By looking at the result, temp\_max and temp\_min is almost highly positive correlated and might need to be removed in feature selection later.

### 2.4.2.2 Scatter Plot Matrix

This plot technique is used for checking the structure between attributes. Below is the code and result:

A graph of different weather conditions

Description automatically generatedA close up of a black background

Description automatically generated

Based on the result, temp\_max and temp\_min almost have structured relationship and one might need to be removed later.

## 2.5 Data Preprocessing

Machine learning method often assume the data have gaussian distribution, as a result, the dataset used in this project need a transformation because the data have exponential distribution and only close to gaussian distribution. The library used for this preprocessing method will be scikit-learn and the method that will be used for preprocessing is standardize.

### 2.5.1 Standardize Data

Standardizing is a method of transforming data to standard gaussian distribution with mean value of 0 and standard deviations of 1 in order to make the data into a closer scale (Hale, 2019). The reason why this project use standardize is the data has a close gaussian distribution and according to Brownlee (2016), standardize works better for logistic regression and this project is a classification problem, so logistic regression will be used later. To begin standardizing the data, StandardScaler will be imported from scikit-learn library. Below is the code followed by the result.

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, font, typography

Description automatically generated

First, the data will be split into input and output. X represents the input and Y represents the output. As a result of the 0 value for the mean and 1 for the standard deviation, the range between the data is on a similar scale, therefore, no attributes will outweigh each other.

## Feature Selection

Feature Selection is a method to select the attribute that contributed the most on the prediction. Doing feature selection bring some benefits, for example:

* Reducing overfitting.
* Improving accuracy of the model.
* Reducing the time needed to train the model. (McCombe, 2019)

The techniques that will be used for feature selection are univariate selection, recursive feature elimination, and feature importance. The reason why this project has used more 1 technique is because it gives more explanation and clarity about the most important features that will have a high contribution in the prediction.

### 2.6.1 Univariate Selection

This is a statistical test for selecting the features that have the strongest relationship with the output variable (Thailappan, 2021). The reason using this technique is because it suitable for classification problem that has numerical inputs and categorical data (Brownlee, 2016). Below is the code and the result:

A picture containing text, font, screenshot

Description automatically generated

A screenshot of a computer screen

Description automatically generated with low confidence

The result showing the best feature based on the score is precipitation (119.086), maximum temperature (94.839), and wind (40.779).

### 2.6.2 Recursive Feature Elimination (RFE)

A picture containing text, screenshot, font

Description automatically generatedRFE is using model accuracy which in this case logistic regression to spot the features that contributed the most by ranking the feature. For this project, the number of top features will be 3. Below is the code and the result:

A black background with white text

Description automatically generated with low confidence

The result show exactly like univariate selection with precipitation, maximum temperature, and wind as the best features.

### 2.6.3 Feature Importance

Feature importance works similarly with RFE, but this technique utilised another model which are ExtraTressClassifier. Below is the code and the result:

A picture containing text, screenshot, line

Description automatically generated



The result showing the importance score of the features, and the higher the score means more important. Therefore, the result is also similar to univariate and RFE.

# Conclusion

In conclusion, the dataset used in this project is weather prediction from Kaggle. The preprocessing method used is standardized to make all the data on a similar scale. Furthermore, the best features according to feature selections are precipitation, maximum temperature, and wind. Minimum temperature will be removed since it is less important compared to others and it has high correlation with maximum temperature, but maximum temperature has higher score, so minimum temperature will be the one to be removed.

# Future Works

* Evaluating algorithm by using classification metrics.
* Spot checking using machine learning algorithm and compare the performance between each model.
* Fine tuning the model.
* Finalizing the model.

# References

1. Brownlee, J. (2016). Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end. Machine Learning Mastery.
2. Hale, J. (2019, March 4). *Scale, Standardize, or Normalize with Scikit-Learn*. Retrieved from towardsdatascience: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02#:~:text=Think%20about%20how%20a%20scale,distribution's%20standard%20deviation%20equals%20one.
3. McCombe, M. (2019, May 24). *Intro to Feature Selection Methods for Data Science*. Retrieved from towardsdatascience: https://towardsdatascience.com/intro-to-feature-selection-methods-for-data-science-4cae2178a00a#:~:text=There%20are%20three%20types%20of,%2C%20Ridge%2C%20Decision%20Tree).
4. Schacher, J. (2022, January 25). *The ethical dilemmas of data-driven climate action*. Retrieved from Global Partnership for Sustainable Development Data: https://www.data4sdgs.org/blog/ethical-dilemmas-data-driven-climate-action
5. Siva, S.(2023). seattle-weather-predection[Data Set]. https://www.kaggle.com/code/sivaneshansiva/seattle-weather-predection/input
6. Thailappan, D. (2021, June 11). *Feature Selection Techniques in Machine Learning*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/feature-selection-techniques-in-machine-learning-2/