**Weather Forcast with Machine Learning**

Gia Luong Vu

Table of Contents

[Analysis 3](#_Toc174468284)

[Project Introduction 3](#_Toc174468285)

[Definitions 3](#_Toc174468286)

[Weather Forecast software 4](#_Toc174468287)

[Data analysis 7](#_Toc174468288)

[Data collection 7](#_Toc174468289)

[Exploratory Data Analysis (EDA) 8](#_Toc174468290)

[Data Cleaning 9](#_Toc174468291)

[Feature Engineering 9](#_Toc174468292)

[Outlier removing 10](#_Toc174468293)

[Data Imputing 10](#_Toc174468294)

[Data Scaling 11](#_Toc174468295)

[Neural networks with deep learning 12](#_Toc174468296)

[Interviews 13](#_Toc174468297)

[Objectives 14](#_Toc174468298)

[Main objectives 14](#_Toc174468299)

# Analysis

## Project Introduction

Weather forecast has been an essential aspect of modern society, impacting hugely on daily life, agriculture, transportation. It is also extremely sophisticated and complicated when it comes to Uk weather, where the condition can change rapidly. Therefore, I want to replicate the weather forecasst, but adjust to make it specific to Coventry’s residents – where NatMatSci students are living and extend its capability.

I’m building this project towards data analysis and deep learning, which are my main interests. I’ll be collecting data from the Internet via API, using a method of data analysis called EDA to alter it to make it useful and processable. Then, I’m planning to build a neural network from scratch and train it so that it can predict the weather in the future. A simple GUI will also be create using tkinter allow user to interact with. The details will be mentioned in the research section.

### Definitions

* **Humidity**: The amount of moisture present in the air.
* **UV Index**: A measure of the strength of ultraviolet radiation from the sun that can cause skin damage.
* **Precipitation**: Any form of water, such as rain, snow, sleet, or hail, that falls from the atmosphere to the ground.
* **Wind gust speed**: The maximum speed of a brief and sudden increase in wind intensity, typically lasting for a few seconds.
* **Dew point**: The temperature at which air becomes saturated with water vapor, causing condensation to occur, often leading to the formation of dew or fog.

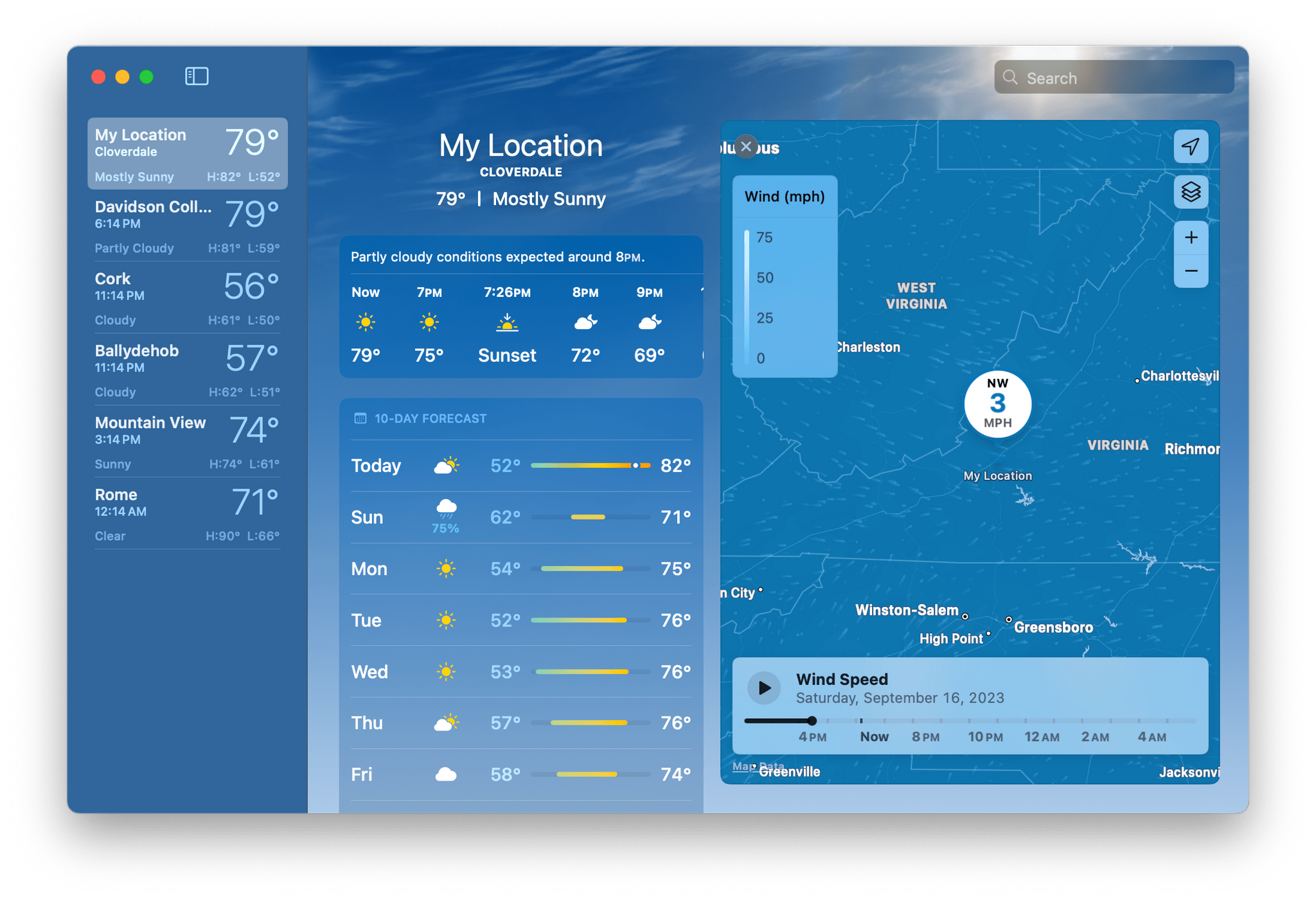
### Weather Forecast software

1. **Apple Weather App**

The Apple Weather app is a versatile and user-friendly application designed to provide users with up-to-date and fairly accurate weather forecasts directly on their Apple devices for places all over the world. The app offers a comprehensive overview of current weather conditions, hourly forecasts, up to 10-day forecasts along with detailed data namely Air Polution, UV Index, Humidity, Wind Speed, etc.

The app automatically detects the user's current location, personalized weather information tailored to their specific area. Users can also add multiple locations to track weather conditions in different places of interest, making it convenient for travelers, commuters, and anyone with diverse weather-related needs.

Beyond basic weather information, the Apple Weather app offers additional features such as severe weather alerts, radar maps, and customizable weather widgets for quick access to important forecast details. With the integration across Apple devices, users can access the Weather app from their iPhones, iPads, and Apple Watches, ensuring consistent and convenient access to weather updates wherever they go.



*The Interface of the Apple Weather App*

1. **BBC Weather Website**

The BBC Weather Website is a trusted source for up-to-date and accurate weather forecasts across the United Kingdom and internationally. It delivers comprehensive forecasts tailored to specific regions, providing users with essential information to plan their day effectively. The BBC Weather service offers detailed insights into current conditions, hourly updates, and extended forecasts for up to 15 days.

Despitenot being an app, the BBC Weather Website are more accurate and it can cover a smaller area, making the data more precise. It also have lots of features such as the news, Weather Watchers, warnings, etc.

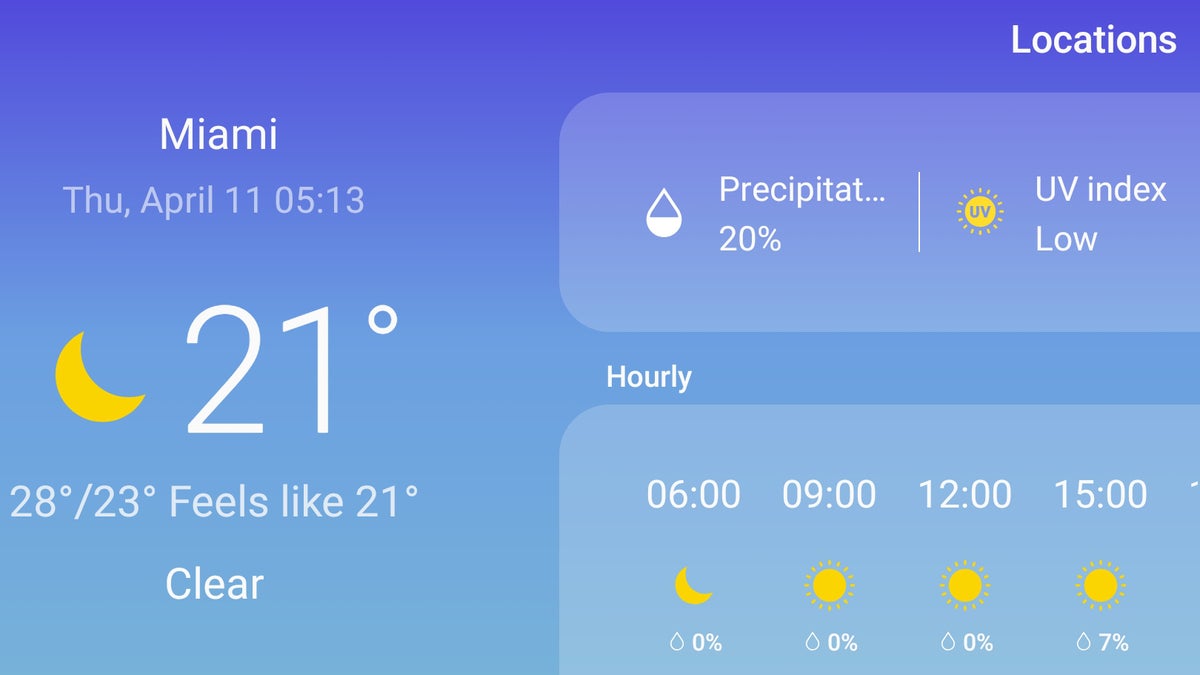


*The Interface of the BBC Weather website*

1. **Samsung Weather App**

The Samsung Weather app is a pre-installed feature on many Samsung smartphones and tablets, seamlessly integrated into the One UI user interface. Offering users a convenient way to access up-to-date weather forecasts and conditions, the app serves as a reliable companion for daily weather tracking.

At its core, the Samsung Weather app provides users with accurate weather forecasts tailored to their location. Whether users are interested in current conditions or future predictions, the app delivers comprehensive insights into temperature fluctuations, humidity levels, wind speeds, and precipitation forecasts. By leveraging data from reputable weather sources, the app strives to offer users reliable and precise weather information.



1. **Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Apple Weather App | BBC Weather Website | Samsung Weather App |
| Accuracy | Relative high | High | Relative high |
| Lead time[[1]](#footnote-1) | 10 days | Up to 15 days | 7 days |
| Geographic coverage | Cover a large area | Cover small area |  |
| Granularity[[2]](#footnote-2) | Detailed | Not detailed |  |
| Reliability | Fairly | Reliable | Fairly |
| Accessibility | Very convenient | Not too convenient | Very convenient |

### Data Analysis

#### Data collection

In the developed society, most of the data can be found on the Internet, and weather’s data is not an exception. There are a lot of sources in the Internet, but many of which are either not free, or lack of data. Then, I came across <https://weather.visualcrossing.com/> , a really useful website for collecting data that provide about 50 years data. It has API that support Python so I can extract daily weather forecast easily.

A screenshot of a computer

Description automatically generated

Interface and the API

Grid form of the data



Grid form of data

#### Exploratory Data Analysis (EDA)

Exploratory Data Analysis, or EDA, is a process of analyzing data with an aim to summarize the characteristics of a dataset. It consists of a number of techniques, and I’ll be using some of them, namely data visualization, statistics summary, univariate/bivariate analysis.

A **statistics summary** typically refers to a concise overview or summary of key statistical measures and characteristics of a dataset. The components of it are:

* Central Tendency: Mean, Median, Mode
* Variability: Standard deviation, Variance, Range
* Distribution: Frequency

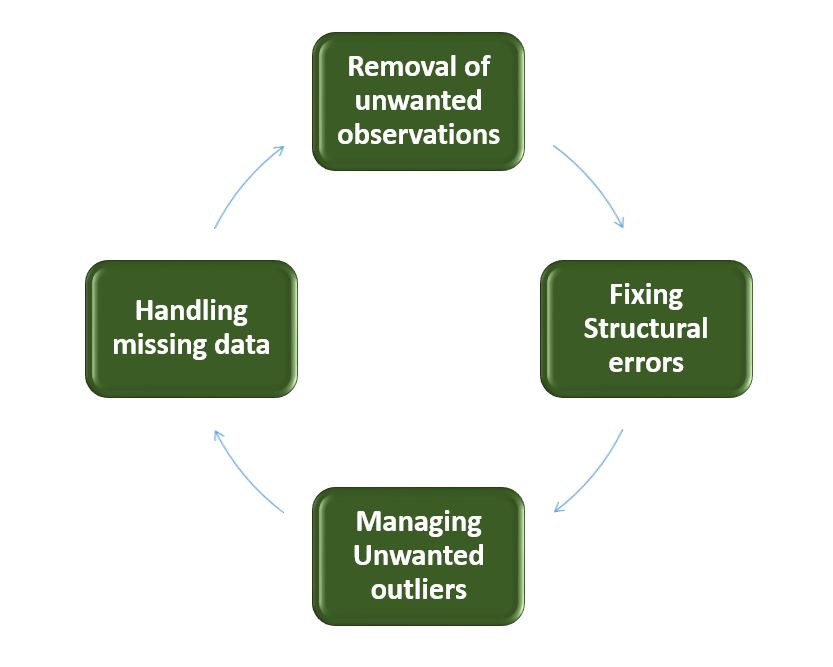
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Types of data visualization are

* **Histogram**: Shows the distribution of a numerical dataset.
* **Bar Chart**: Represents categorical data with rectangular bars
* **Pie Chart**: Displays proportions of a whole.
* **Scatter Plot**: Displays the relationship between two numerical variables.
* **Line Graph**: Shows trends over time
* **Box Plot**: Visualizes the distribution and identifies outliers
* **Heat Map**: Represents data density on a geographical map

#### Data Cleaning

I’m going to implementing this in 3 steps:

* Feature Engineering
* Outlier removing
* Imputing missing data
* Data scaling



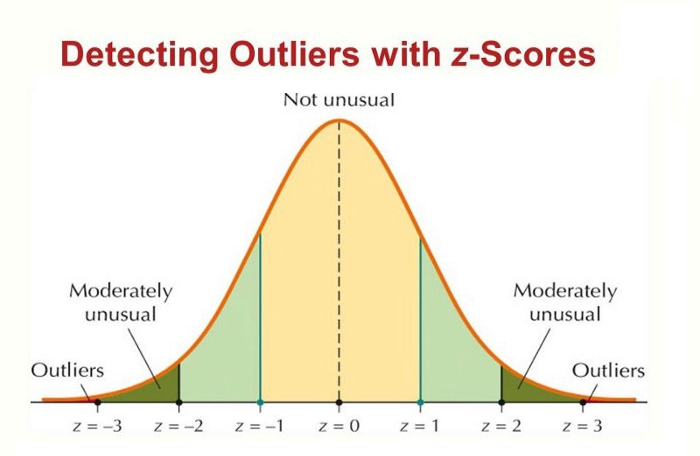
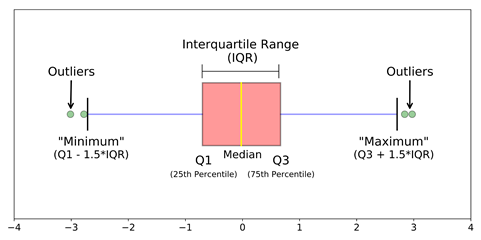
##### Feature Engineering

Feature engineering is the process of selecting, creating, or transforming features (variables) in your dataset to improve the performance of machine learning models. In my project specifically, it includes encoding and data binning and data extraction.

* Encoding: Data encoding is the process of converting categorical or textual data into a numerical format that can be used by machine learning algorithms. Useful encoding algorithms are
  + **Label Encoding:** Assigns a unique integer to each category. For example, "Red," "Green," and "Blue" could be encoded as 1, 2, and 3
  + **One-Hot Encoding:** Converts each category into a binary vector, where each category is represented by a vector of zeros with a single one in the position corresponding to that category. For instance, "Red," "Green," and "Blue" would be represented as [1, 0, 0], [0, 1, 0], and [0, 0, 1].
* Data binning: Data binning, also known as discretization, is the process of transforming continuous data into discrete intervals or "bins." This technique is often used to reduce the impact of small observation errors and to make models less sensitive to noise in the data
* **Data extraction** refers to the process of selecting and creating relevant features from raw data that will be used as input for machine learning models

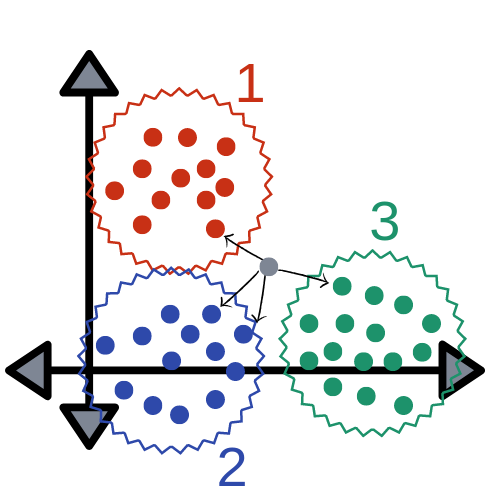
##### Outlier removing

Outlier removal is the process of identifying and eliminating data points that significantly differ from the majority of the data. These outliers can distort statistical analyses and machine learning models, leading to inaccurate predictions or insights. I typically use 2 types of outlier removing:

* The **Z-score** is a statistical measure that indicates how many standard deviations a data point is from the mean of the dataset. It’s a useful method for detecting outliers in normally distributed data.
  + 
* The **Interquartile Range (IQR)** is a measure of statistical dispersion, which is the range within which the central 50% of the data lies. It’s effective for identifying outliers in skewed or non-normally distributed data
  + 

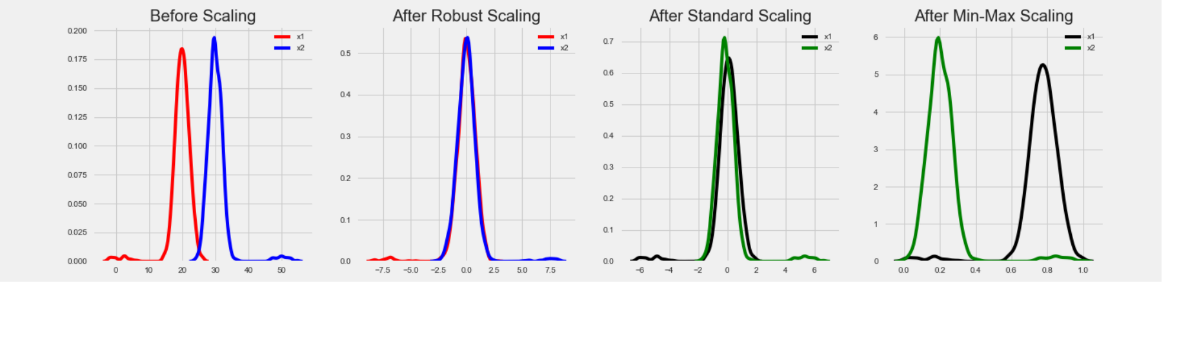
##### Data Imputing

**Data imputing** is the process of filling in missing or incomplete data within a dataset. Missing data can occur for various reasons, such as errors in data collection, entry mistakes, or unavailable information. Imputing is essential because most machine learning algorithms cannot handle missing data directly, and ignoring or dropping incomplete records can lead to biased or less accurate models. In my model, I use 2 types of data imputing:

* **Standard Imputer**: This method fills missing values using basic statistical measures. Common strategies include:
  + **Mean Imputation**: Replacing missing values with the mean of the observed values for that feature.
  + **Median Imputation**: Using the median of the observed values for a feature.
  + **Most Frequent Imputation**: Filling missing values with the most frequent value (mode) in the feature.
* KNN Imputer: The K-Nearest Neighbors (KNN) imputer fills missing values based on the values of the nearest neighbors. It uses the distances between data points to find the closest samples and then imputes missing values based on a weighted average (or other aggregate measure) of these neighbors.
  + 

##### Data Scaling

Data scaling refers to the process of transforming features to ensure they are on a similar scale, which can improve the performance and convergence speed of machine learning algorithms. Common methods for data scaling include:

* **Normalization**: Adjusts the values to fit within a specific range, usually [0, 1] or [-1, 1]. This can be done using MinMaxScaling that rescales the data based on the minimum and maximum values of the feature.
* **Standardization**: Transforms the data to have a mean of 0 and a standard deviation of 1. This is done using Z-score Standardization where each feature value is scaled based on the mean and standard deviation of the feature.
* 

### Feedforward Neural networks with deep learning

#### Overall

### Interviews

**Q: How often do you rely on weather forecast applications for planning your day or activities?**

A: I usually check the weather forecast when I’m planning to go outside to Tesco or going to football. I usually check in the morning to decide what to where for that day

**Q: What features do you find most useful in a weather forecast application?**

A: I often check for temperature and the weather condition like raining, snowing or sunny

**Q: Have you used multiple weather forecast applications? If so, what are some differences you've noticed between them?**

A: Apart from my devices, I sometimes watch weather forecast on the news. The biggest difference I’ve noticed is the fact that my app are more personalized and the news only tell me the overall condition of the whole country.

**Q: How satisfied are you with the accuracy of weather predictions provided by the application you use?**

A: I’m quite pleased with the current accuracy of predictions. However, I would like to have some sense about how accurate the data is.

**Q: Do you have any preferences for the user interface or design of weather forecast applications?**

A: I prefer a simple but concise interface that can provide me with useful data. Some graph might make the interface more interesting

**Q: Have you encountered any challenges or difficulties when using weather forecast applications? If so, what were they?**

A: There aren’t any difficulties in using my app, but I want it to be more customizable.

**Q: Do you typically use additional features such as radar maps, severe weather alerts, or long-term forecasts in weather forecast applications?**

A: I rarely use them, I will just look through them when I’m curious or when I have free time.

**Q: How important is it for you to receive timely updates and notifications from the weather forecast application?**

A: It is fairly important to me have timely updates and notifications from weather forecast application as I usually need them when I plan to go outside.

**Q: Have you ever recommended a weather forecast application to others? If yes, what influenced your recommendation?**

A: I think that the default app in both IOS or Android can be helpful, but some laptops does not have really useful weather forecast.

## Objectives

### Main objectives

1. Model
2. The model has to be able to predict 3 days in the future
3. The prediction has to have at least 3 weather conditions including
   * Humidity
   * Temperature
   * Wind speed/ Gust speed
4. The error of the model, measured by RMSE has to be less than all the weather condition by more than 20%
   * RMSE (Root Mean Square Error) is calculated by taking the square root of the mean of the square of the risiduals
   * RMSE =
     + n is the number of observations or data points.
     + is the actual value for the observation.
     + is the predicted value for the observation.
5. The model can be train by itself after loading data for each day
6. Designing
7. There have to be at least 5 weather conditions including
   * Temperature
   * Wind speed/ Gust speed
   * Humidity
   * UV Index
   * Precipitation
8. There has to be a Graphic User Interface that that show all all weather condition and prediction
9. The GUI should not be
10. Users can search weather data for the date they want to
11. The application can update by itself without re-running
12. Should be larger than 600x400
13. There has to be at least some
14. It takes less than 5 seconds to update for new day and load new data
15. User can be able to customize with image

1. Evaluate how far in advance the forecast predicts weather conditions accurately [↑](#footnote-ref-1)
2. Evaluate how detailed it is [↑](#footnote-ref-2)