**London Weather Forecast Report**

**Module: Programming for machine learning**

**Cohort: BCNS/24A/FT1**

**KISSOON Dhruv (2403\_27574),**

**HONG-LIN Severine (2403\_27569)**

**GHOORUN Vriddhi (2403\_27563),**

**HARRAH Bhevisha (2403\_27568)**

**GOPEE Shahista (2403\_27566),**

**KHODABOCUS Rehaan (2403\_27573)**

# **Introduction**

Weather forecasting has traditionally been done by physical models of the atmosphere, which are unstable to perturbations, and thus leading to an inaccuracy in a long time run. Machine learning techniques are more robust to perturbations and can therefore make better predictions on weather conditions.

In this report, a machine learning approach will be implemented to showcase how the company, Climate Insights Ltd has used the machine learning algorithm based on a pre-stored dataset to make weather related predictions by analysing and providing a better insight for local businesses in London, especially those who are dependent on weather conditions. But first the dataset has to be vetted and analysed for errors or duplicates after which, exploratory data analysis (EDA) must be performed to find the relationships between the features of this dataset following a general conclusion on the EDA the most suitable machine algorithm will be picked. Finally, the model suitability will be judged by some performance metrics.

# **Analyzing datasets and identifying any missing features or additional ones**

**Existing weather features:**

Light rain, heavy rain, snow, wind, precipitations, fog, temp\_min and temp\_max

**Additional Features**

1. **Visibility**

* Visibility can be predicted from foggy weather conditions. For example, when fog level is too high, the visibility will be at its highest. From the fog column, visibility can be useful in explaining atmospheric conditions and how fog, visibility, humidity and precipitation are correlated and may lead to rain.

1. **Humidity**

* Humidity can help explain certain weather phenomena. For example, high humidity levels are often accompanied precipitation. Even though the dataset has a precipitation column, humidity could provide more insight into the atmospheric conditions leading to rain or fog. Humidity, temperature, and wind are all interrelated. For instance, the combination of high humidity and high temperature can make the weather feel hotter due to the heat index, and variations in humidity levels can influence how temperature changes are experienced.

1. **Extreme Weather**

* From all the data set available, an additional column named ‘Extreme Weather’ can be added. This column will be of great use for further predictions on extreme weather conditions such as flash floods, hurricanes or any other weather phenomena. Outliers from all the associated data set can be classified in this column, with their possible associated weather conditions, this can help to better predict extreme weather conditions over time.

1. **Solar Radiation**

* Solar radiation (Sunlight) is the electromagnetic energy emitted by the sun. When sunlight reaches the Earth's atmosphere and surface, it warms the Earth thus absorbing solar radiation which in turn heats the land, oceans, and air, impacting the temperature patterns. Solar energy evaporates water from water and land areas, leading to precipitation. It directly influences temperature circulation, creating temperature differences that lead to wind patterns and atmospheric circulation. SST is a crucial factor influencing humidity, air temperature, and the formation of cyclones, it is measured in Watts per square meter

1. **Wind direction**

* From the wind column dataset, wind direction may be predicted to have a better insight into how wind can have an impact on the direction of rain.

**For DATE Column**

* Extract days, months and years
* Extract by day-of-week
* Deduce SEASON\*

**Calculations/Predictions**

* ***Count occurrences of different weather conditions => see which type is more common***
* ***Compare weather across different months => calculate mean***
* Average daily wind
* To verify the datatypes used for manipulations
* ***Graph plot to analyze the data for any gaps or superfluous data***

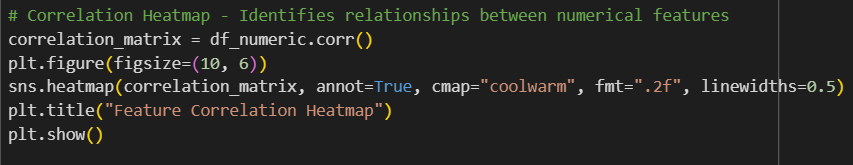
# **Determination of stability by analysing the dataset’s dimensionality**

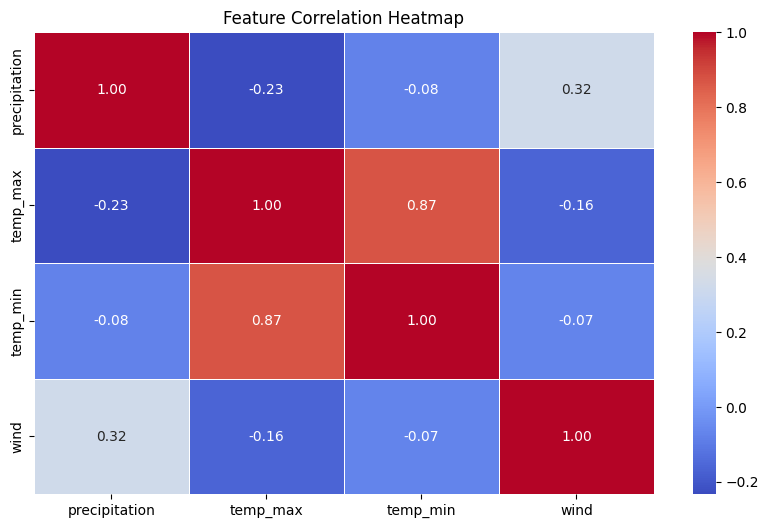
The dataset is a low-dimensional dataset as the dataset has only 6 columns of data. It is easier to interpret relationships between a small set of variables. With fewer features, there will be a lower risk of overfitting, especially if the sample size is large relative to the number of features, leading to the creation of simpler models. The current data set available captures only the basic weather information. However, if the study requires a deeper analysis such as predicting subtle weather patterns or capturing other climatic effects, additional variables like humidity or wind direction might be required as mentioned above. The date column can be broken down to create extra temporal features (day, month and year). While this increases dimensionality, it can be beneficial if any additional analysis is important for the study.

If the company focuses only on general weather trends or simple analysis, the six columns’ data set may be entirely adequate, especially when temporal features from the date have been extracted, thus making it straightforward for exploration analysis and initial modeling. Its suitability depends entirely on the study’s goals. For basic weather analysis, the columns are sufficient. However, if the study demands a more detailed understanding of other weather conditions, additional relevant variables may be required to have a better interpretation of the data set.

***ANALYSIS OF EDA***

***1) Correlation Heatmap***





**Type of Analysis:** **Multivariate Analysis** (Multiple numerical features analysed together)

**Goal**:

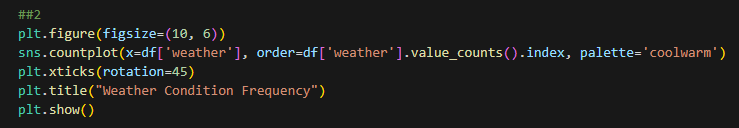
* Identify **relationships between numerical features** in the dataset**.**
* Determine **which variables most affect each other.**
* Help **feature selection** for machine learning.

**Trends observed**:

* **Strong positive correlation (0.87) between** temp\_max **and** temp\_min → Hotdays mostly have hot nights**.**
* **Moderate positive correlation (0.32) between wind speed and precipitation** → Heavy rainfalls are followed by strong winds, but not always.
* **Weak negative correlation (-0.23) between precipitation and max temperature** → Rainfall happens more frequently on cooler days but is not temperature driven.

**Deduction**:

* **A strong positive correlation between temp\_max and temp\_min shows that there is a constant daily temperature range, days with higher min\_temps have higher max\_temps**
* **Thanks to the weak negative correlation between precipitation and max temperature it can be said that days with higher precipitation tends to have lower max\_temps.**
* Due to the weak negative correlation between wind and maximum temperature could indicate that days with lower wind might tend to be slightly warmer

***2) Weather Condition Frequency***

A graph of different colors

Description automatically generated

**Type of Analysis:** **Univariate Analysis** (Single categorical variable: weather)

**Goal:**

* **Comprehend the distribution of various weather types** within the dataset.
* Determine **the most common and rarest weather conditions.**

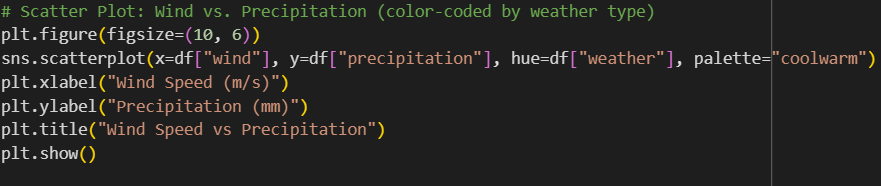
**Trends Observed:**

* **Light Rain is the most frequent occurring weather condition.**
* **Heavy Rain is less frequent but still appears regularly.**
* **Sunny days are frequent but happen in specific seasonal patterns.**
* **Fog and snow show up the least and hence are rare occurrences.**

**Deduction**:

* **Sunny and rainy days occur about equally frequently**
* **Snowy weather is the least likely weather to occur.**
* **Rare weather conditions (fog, snow) may have fewer training samples,** affecting classification model performance.

***3) Wind vs. Precipitation (Color-Coded by Weather Type)***



A graph of different colored dots

Description automatically generated with medium confidence

**Type of Analysis:** **Bivariate Analysis** (Wind Speed & Precipitation)

**Goal:**

* Examine the link **between wind speed and precipitation.**
* Determine if **certain weather types are associated with high wind speeds and heavy rainfall**.

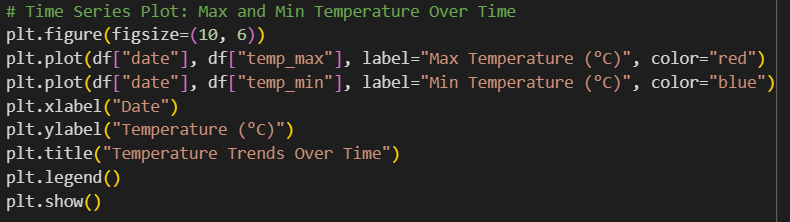
**Trends Observed**:

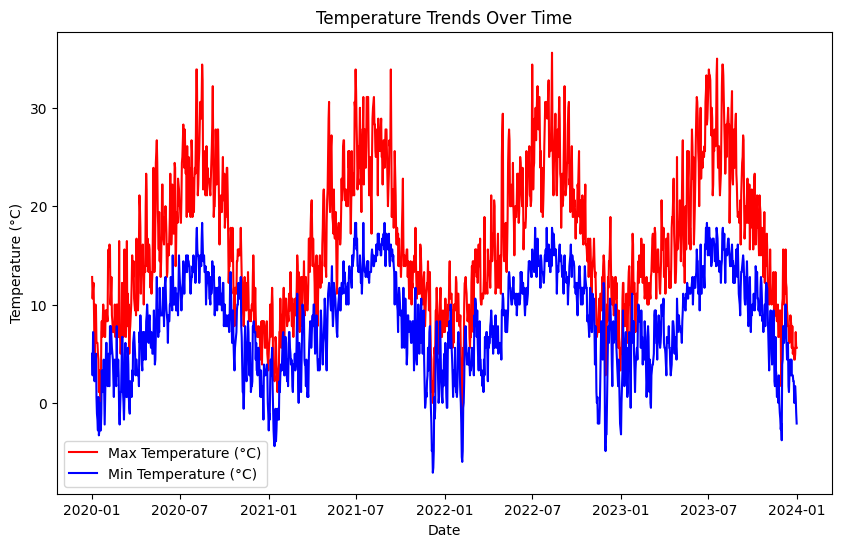
* A weak relationship is observed between wind speed and precipitation, suggesting they are **not strongly correlated** in this dataset.
* Most data points show **low wind speeds (0-15 km/h) and low precipitation rates (0-5 mm).**
* There are certain points of extreme with **heavy precipitation (above 10 mm) and relatively high wind velocities (above 20 km/h)** exist.
* These could correspond to **strong weather conditions**, such as storms or heavy rain accompanied by gusty winds.
* A few scattered points have **very low wind velocities but heavy rainfall,** possibly indicating **calm rain events** or measurement inconsistencies.

**Deduction:**

* The scatter plot does **not have a significant trend** of wind speed versus precipitation. This suggests that **high wind speeds are not always associated with more precipitation** and conversely.
* A few **high-precipitation and high-wind-speed points** likely represent storms or heavy rainfall events.
* Some **low-wind yet high-precipitation** points suggest calm rain events.
* The presence of outliers could **skew predictive models,** especially if using regression techniques. A classification model should **account for these extreme events separately** rather than treating them as noise.

***4) Time Series Plot: Max and Min Temperature Over Time***





**Type of Analysis:** **Bivariate Analysis** (Date vs. Temperature)

**Goal:**

* To analyze temperature fluctuations over time.
* Identify trends, seasonal variations, and possible anomalies.

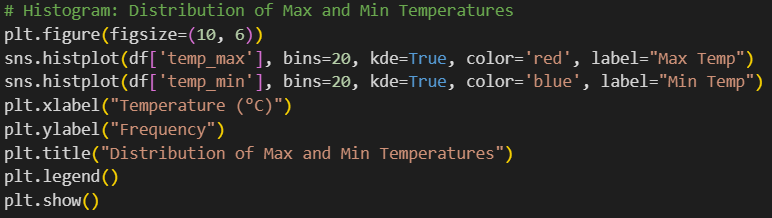
**Trends Observed**:

* Max and min temperature change daily, with obvious peaks and drops.
* Some of the **outliers** (unusually high or low temperatures) **present** may indicate extreme weather conditions.
* There seems to be an obvious seasonal trend affecting min and max temperatures

**Deduction:**

* Like mentioned before in the previous graphs it can also be seen here that min and max temps have a positive correlation. Also, the difference between the min and max curves are relatively constant meaning a positive correlation.
* **Outliers indicate extreme temperature events,** worth investigating**.**
* **Trendlines (moving averages) may enhance clarity** and assist in identifying patterns like heatwaves or cold spells.
* **There seems to be not much variation in the temperature pattern showing that each year this points to a stable and predictable seasonal cycle**
* **A prediction of temp\_max likely also means a prediction of temp\_min can be deduced and vice versa**

***5) Distribution of Max and Min Temperatures***



A graph of different colored bars

Description automatically generated

**Type of Analysis: Univariate Analysis** (Max and Min Temperature separately)

**Goal:**

* To understand how temperatures are distributed.
* To find the **most common** temperature ranges.
* Detect **skewness** or anomalies in the data.

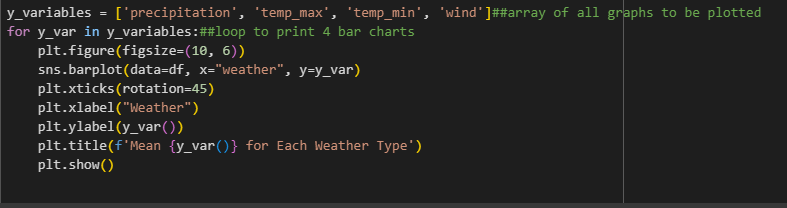
**Trends Observed:**

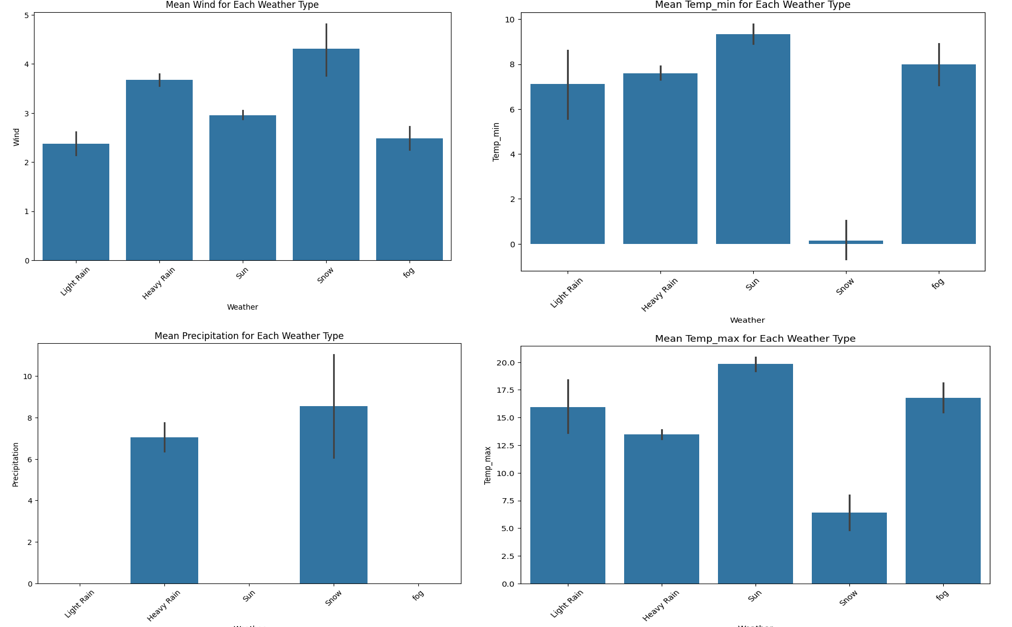
* The max temperature follows an approximately **normal distribution**, with most values grouped around a central range.
* The min temperature distribution may be somewhat skewed.
* The two distributions **may overlap,** but max temperatures tend to be higher.

**Deduction:**

* **The dataset has a reasonable temperature spread** with no extreme skewness.
* Most frequent maximum temperatures fall between approximately 20°C and 25°C. Higher maximums are less frequent, but the distribution extends to higher temperatures than the minimum temperature distribution extends to lower temperatures.
* Most frequent minimum temperatures are approximately between 10°C and 15°C. Lower minimums are less frequent.
* **Outliers in the tail regions** could represent heatwaves or cold snaps but these are rare.
* Adding mean/median labels could provide better insights into the central tendency.

***6) MULTIPLE MEAN VARIABLES ACROSS WEATHER TYPES***



****

**Type of Analysis:** **Multiple** **Bivariate Analysis** (Categorical: Weather, Numerical: Max Temperature, Precipitation, Min temperature, wind)

**Goal:**

* Find underlying patterns or factors that explain the variation in precipitation, temperature and wind across different weather types?
* Try to predict the weather type based on the values of precipitation, temperature, and wind.

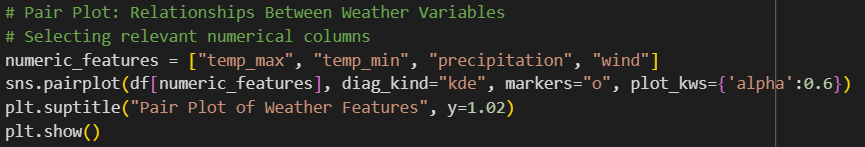
**Trends Observed:**

* Snow tends to have higher precipitation than heavy rain and also precipitation tends to be very low to none in weathers apart from snow and heavy rain.
* Heavy rain and snow tend to have higher mean wind temperature
* “sun” weather tends to have a higher max temperature compared to other weather types while snow the lowest, surprisingly fog tends to have the second highest mean temperature.
* The temperature range between max and min appears to be largest for sunny weather, suggesting a greater temperature variation. Whereas Snow shows a very small temperature range indicating consistently cold conditions.

**Deduction:**

* Sunny weather in this data is associated with the warmest temperatures, both in terms of maximum and minimum daily temperatures. It also experiences the lowest wind speeds and negligible precipitation. This suggests sunny days are typically calm, warm, and dry.
* Snowy weather has the lowest temp, highest precipitation and highest windspeed pointing to the fact that snowy days are usually colder, more windy and wet
* Precipitation, especially in the forms of heavy rain and snow, tends to be associated with higher wind speeds.
* Temperature range is most significant during sunny weather and minimal during snowy weather.

***7) Relationships Between Weather Variables***



A group of blue dots

Description automatically generated

**Type of Analysis: Multivariate Analysis** (Multiple numerical variables)

**Goal:**

* To explore **relationships between numerical weather variables** (e.g., temperature, wind, precipitation).
* Detect correlations and potential **patterns useful for predictive modelling.**

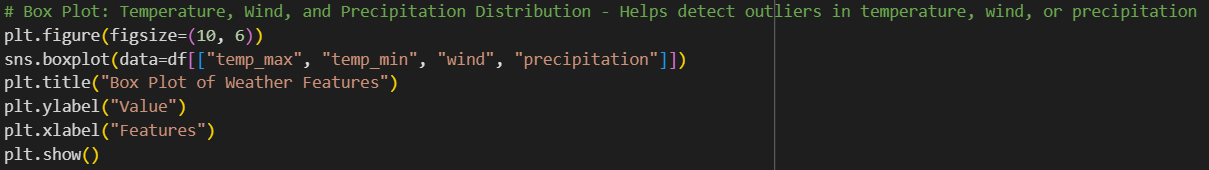
**Trends Observed:**

* **Temp\_max and temp\_min is strongly correlated** (expected).
* **Wind speed does not strongly correlate with temperature**, but extreme winds might coincide with extreme temperatures.
* **Precipitation seems independent of temperature**, suggesting rain events occur at varied temperatures.

**Deduction:**

* **Temp\_max and temp\_min move together**, confirming seasonal trends.
* Wind and precipitation do not show strong patterns with temperature.
* Most of the variables are unrelated to each other this makes it harder to predict the weather as its harder to notice a pattern potentially extra features are to be considered…

***8) Temperature, Wind, and Precipitation Distribution***



A diagram of different colored rectangular shapes

Description automatically generated

**Type of Analysis: Univariate Analysis** (for each variable)

**Goal**:

* To **detect outliers** in temperature, wind, or precipitation.
* To compare the distribution of numerical weather variables.

**Trends Observed:**

* **Max & Min Temperature**: Likely have some **outliers** (extremely high or low values).
* **Wind Speed:** May have **a few extreme values**, possibly from storms or high-wind events.
* **Precipitation**: Expected to show **skewed distribution**, with rare heavy rainfall events appearing as outliers.

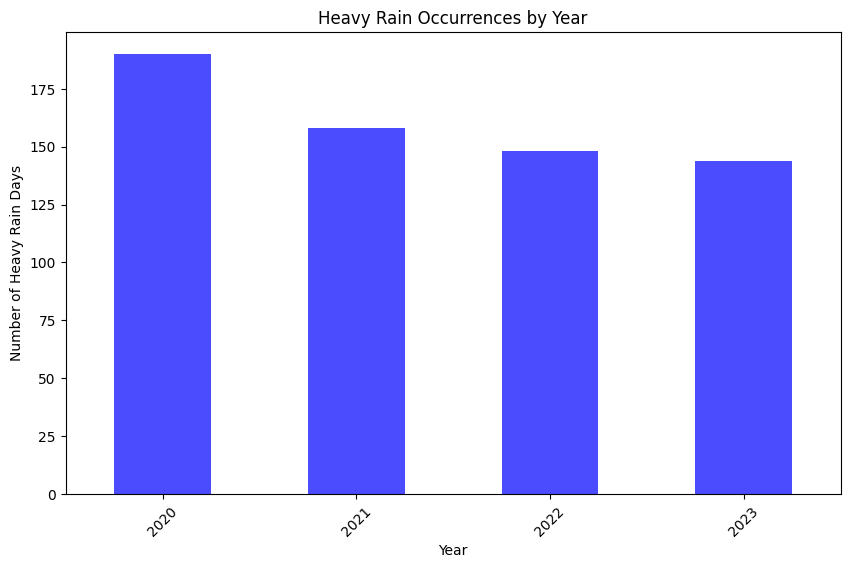
**Deduction:**

* Temperature distributions are consistent with mostly no outliers.
* **Some of the outliers from the wind data may indicate extreme events like Storms.**
* **Precipitation shows extreme outliers**, meaning as can be seen from graph (3) above, heavy rain makes up for the bulk of these outliers possibly also likely due to events distinct from the norm.
* **Further investigation is needed into extreme precipitation events (**did they coincide with storms?).
* Precipitation may be harder to predict as there is a large fluctuation in raw data

***9) Heavy Rain Occurrences per Year***

***A screen shot of a computer code

Description automatically generated***



**Type of Analysis: Univariate Time-Series Analysis**

**Goal**:

* To analyze **how often heavy rain occurs** on a yearly basis.
* Identify **trends** in heavy rain frequency over time.

**Trends Observed:**

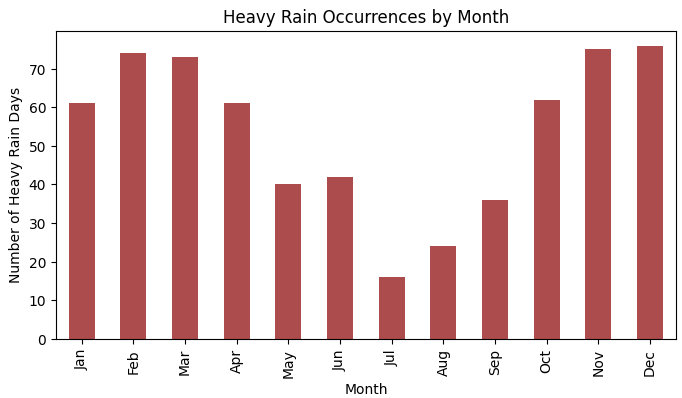
* The **number of heavy rain days fluctuates** from year to year.
* **The Trends** indicate climate shifting towards drier years

**Deduction:**

* From 2020-2021 there was a sudden decrease after which the decrease has been incremental but consistent
* considering the trend is decreasing, it may indicate drier years

***10) Heavy Rain Occurrences per Month***

****

****

**Type of Analysis: Univariate Seasonal Analysis**

**Goal:**

* To determine **which months, experience the heaviest rain**.
* Identify **seasonal trends in heavy rainfall**.

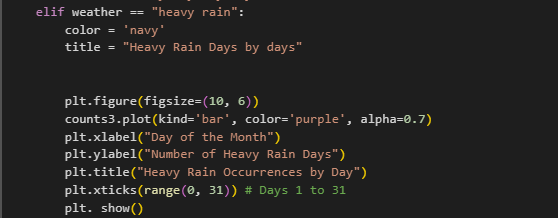
**Trends Observed:**

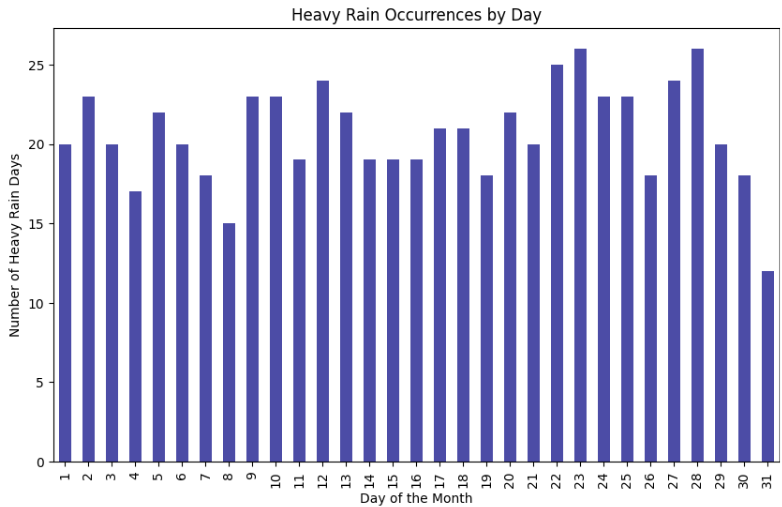
* **Certain months may show consistently higher occurrences** of heavy rain (e.g., monsoon season).
* April to august on comparison have the least heavy rain days with august and July being the two lowest ones by a more significant amount
* But apart from the drier months of May to September the number of heavy rain days are more or less consistent

**Deduction:**

* Comparing with wind data may indicate if storms are linked to heavy rain.
* The amount of heavy rain fluctuates with the seasonal changes (winter, summer, autumn etc..)
* This confirms that Machine learning predictions should depend on seasons also not on just pure raw data as mentioned in the introductory part.

***11) Heavy Rain Occurrences by Day of the Month***

******

****

**Type of Analysis: Univariate Analysis**

**Goal**:

* To check if heavy rain occurs **randomly or follows a pattern** within each month.

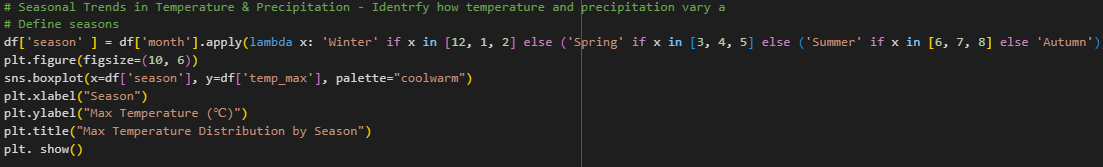
**Trends Observed:**

* A uniform distribution likely means heavy rain is **random.**

**Deduction:**

* **There tends to be no specific pattern as a more or less even spread of “heavy rain” cases can be seen across the graph**
* **Heavy rain is distributed across the month the graph shows that heavy rain events are properly distributed in the 31 days of the month**

***12) Max Temperature Distribution by Season***

****

**A diagram of different colored squares

Description automatically generated**

**Type of Analysis: Bivariate Analysis** (Season vs. Max Temperature)

**Goal:**

* To analyze how **maximum temperature varies** across the four seasons**.**
* Identify **seasonal extremes** and **temperature ranges.**

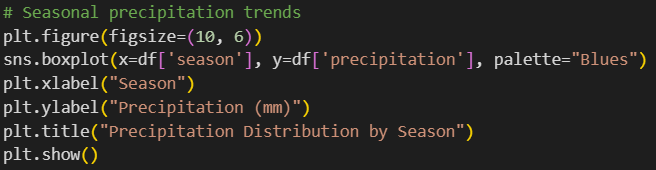
**Trends Observed:**

* **Summer** shows the highest median max temperatures.
* **Winter** has the lowest median temperatures, with a smaller spread. But some outliers
* **Spring & Autumn** exhibit intermediate temperatures, with slightly overlapping distributions.
* **Outliers (**extremely high/low max temperatures) are present in all seasons.

**Deduction:**

* Summer has the highest max temperatures, as expected.
* Autumn & Spring transition smoothly between temperature extremes.
* The presence of extreme outliers suggests occasional heatwaves or cold spells.

***13) Precipitation Distribution by Season***

******

***A graph showing the number of seasons

Description automatically generated***

**Type of Analysis: Bivariate Analysis** (Season vs. Precipitation)

**Goal:**

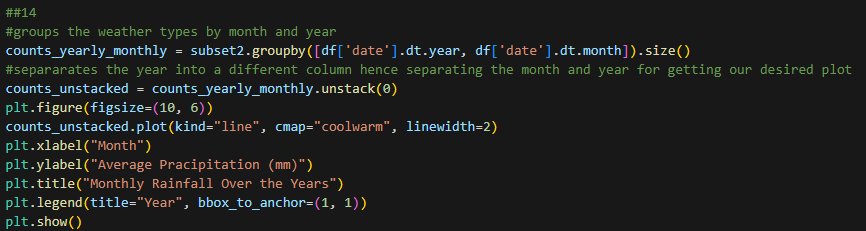
* To determine **which seasons, receive the most precipitation** and observe variations.

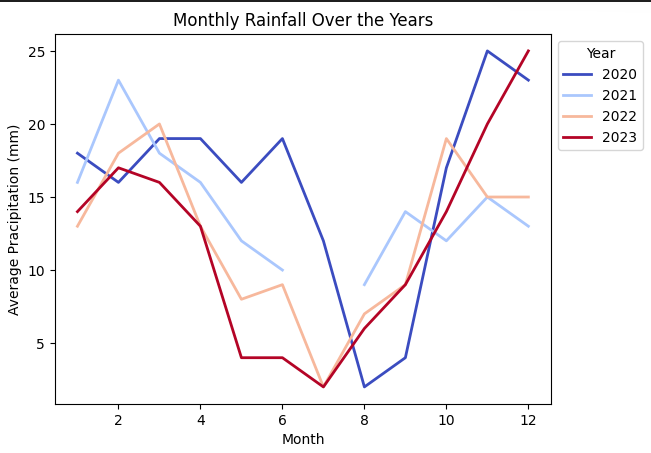
**Trends Observed:**

* **Winter and Autumn** have more rainfall.
* **Summer** generally has lower precipitation but may contain extreme outliers (storms).
* **Spring is comprised of average rainfall**, acting as a transition period.
* Presence of **outliers in all seasons**,

**Deduction:**

* Winter & Autumn are the wettest seasons meaning precipitation is to be expected more.
* **Summer may experience occasional heavy rainfall due to thunderstorms.**
* **Precipitation will be tougher to predict as there are lots of outliers in every seasons.**
* Outliers may signal extreme weather events such as storms or flash floods.
* Consistent with graph 6 it can be seen that winter has the highest average precipitation more than likely, snow.

***14) Monthly Rainfall Trends Over the Years***

****

**Type of Analysis: Multivariate Analysis** (Month vs. Year vs. Precipitation)

**Goal:**

* To analyze how **average monthly rainfall changes over several years.**
* Identify long-term trends in precipitation.

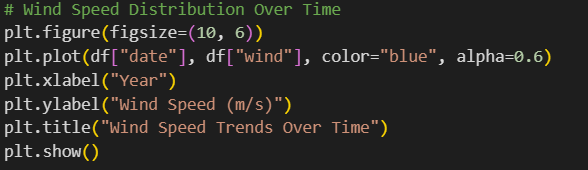
**Trends Observed:**

* Steady rainfall in some months, with variation from year to year in other months**.**
* Certain years might have seen **abnormal rainfall trends** (e.g., exceptionally wet, or dry periods).
* **Long-term climate trends** may emerge if some months show a persistent increase/decrease in rainfall.
* While the overall seasonal pattern is consistent, there is significant year-to-year variability in the *amount* of rainfall.
* **Across all 3 years It can be seen that around month 6 there is a clear dip in precipitation and then it increases again towards the end and start of the year (consistent with graph 10)**
* While the overall seasonal pattern is consistent, there is significant year-to-year variability in the *amount* of rainfall.

**Deduction:**

* **If some months are showing increasing rainfall over the years, climate change or shifting weather patterns may be the reason.**
* **Across all 3 years It can be seen that around month 6 there is a clear dip in precipitation, and it increases towards the end and start of the year**
* Wet and dry periods in different years represent natural variability.

***15) Wind Speed Trends Over Time***

******

*A blue line graph with white text

Description automatically generated*

**Type of Analysis: Univariate Analysis** (Wind Speed Over Time)

**Goal:**

* To track the variation **wind speed** over the period of the dataset.

**Trends Observed:**

* **Small Fluctuations in wind speed** can be seen from time to time
* The wind speeds seem to have more or less constant variations throughout the timeframe with a few outliers

**Deduction:**

* **If wind speeds increase over time, it may indicate that storms happen more frequently, or atmospheric patterns shift.**
* **Wind speed is easier to predict as there Is not much variation in its dataset.**
* Abnormally low or high wind speeds over a period may be due to extreme weather conditions.

**GENERAL CONCLUSION**

* From all the graphs combined it can be seen that most variables aren’t very well correlated to each other(apart from temp\_min and temp\_max) (graph 1,3,7)
* However some good deductions have also been made as in graph 6 (mean ppt vs weather type) if precipitation is present it is either rain or snow and (on average) if ppt is of higher values it is more likely to be snow, It can also be seen that amount of precipitation could be hard to predict(graph 8) as it has many outliers across all seasons and not only that but it can also be seen that in graph 14 that it even varies by year. However some trends are still present.as seen in some graphs(graph **14,10,9)**
* Wind seems to only have one weak positive correlation to precipitation and is not really correlated with a concrete foothold to anything else.(graph 1,3,7)

Generally, it can be concluded that some more features would really help to have a better machine learning model as most of the features of this dataset has weak correlations to each other or have too many outliers.

## **Why label encoding was chosen over hot encoding?**

Since Weather is the target variable, (the one we're trying to predict) it would be best if label encoding were to be used, as one hot encoding is used for features do not target variables. since all features are all numerical also, one hot encoding is not necessary. Even if weather is a categorical variable with no order the other conditions supersede this statement, however it is important to note as the number 1 chosen AI model is classification it is important to make sure a false sense of order is not introduced as in (sunny>foggy) due to the nature of label encoding.

**Chosen Machine Learning Model: Classification**

The goal is to determine whether a Classification or Regression Model is more suitable for predicting weather conditions based on the given dataset: London\_Weather.csv.

**Why is Regression not used?**

Regression is a kind of predictive modeling where the variable to be estimated is of continuous nature. This helps to predict a range of values based on previously taken measurement. For example: price or temperature, their values never stay the same, they continue to change, hence the term “continuous”.

The output type for Regression is Continuous, that is; numbers that may be decimal or integer.

**Why could we have chosen Regression?**

As Regression is a type of Supervised Learning that deals with predicting continuous values or numeric outcomes, it would have been enough to use it for the dataset that was given as it contains numeric continuous values such as Maximum Temperature, Minimum Temperature, Wind, and Precipitation with are all continuous values that changes over time.

If we used Regression and we have the output, it would look like:

* Predicting Temperature
  + Output will be Numerical: 23.5°C, instead of “Sun”
* Predicting Precipitation Amount
  + Output will be Decimal: 8.2mm instead of “Heavy Rain”
* Predicting Wind Speed:
  + Output will be Numerical: 15.7 km/h.

Business would want to know the Weather instead of the actual temperature.

Example: It will be Sunny it better instead of saying its 23.5°.

**Classification**

Classification is a type of Supervised Learning where the model learns in an environment with labelled data. Classification is used for predicting a discrete output, for example in this case: Sunny, Rainy, Snow.

Classification predictive modeling is the task of approximating a mapping function, f from input variables, X to discrete output variables, y.

The output variables are often called labels or categories, The mapping function predicts the class or category for a given observation.

Classification models are used when the outcome variable is categorical or binary, meaning it has a limited number of options.

Types of Classification:

* Binary Classification which can have only two possible classes; Example: Spam or Not Spam.
* Multi-class Classification where more than two possible classes. For Example: Identifying a picture which can be cats, dogs, or birds. In our case, it can be Sunny, Rainy, Snow, Heavy Rain or different outputs.

**Why have we chosen Classification?**

As the output target variable is categorical (Sunny, Rain, Snow, etc....), having discrete classes, it has to classify instances into pre-determined classes.

The data set used in this study comes from “London\_Weather.csv” which was provided. The data in the resolution of one day spans the period from 1st January 2020 to 31st December 2023. A total of 1463 attributes are present in the raw dataset.

With Classification, we have Evaluation Metrics.

* Accuracy.
* Precision.
* Recall.
* F1-Score.
* MSE.
* MAE (not sure need to recheck).

1. Algorithm Suitability

* With Multi-Class Classification, the dataset’s target has multiple classes making algorithms like Decision Trees, Random Forest ideal for capturing non-linear relationships between features, (e.g., Temperature, Wind) and weather types.

1. Limitations of Regression

* Regression Models, e.g., Linear Regression, predicts numerical values, which are irrelevant here since the target is categorical.
* Converting the categorical “weather” variable into numerical labels, e.g., “0” for “Sun” and “1” for “Wind”, would force regression models to treat these classes as ordered values, leading to inaccurate interpretations.
  + Ordered Values here mean that regression models would inherently assume that such relationship exists between labels, that is the values are not identifiers but carry a mathematical relationship. Example: 0<1<2. This can lead to meaningless interpretations if the categories are not inherently ordered.
  + If we had:
    - Sun = 0
    - Rain = 1
    - Snow = 2
    - If we had a prediction of 1.5, then it would imply something meaningless like “Half Rain Half Snow” which is not ideal.

**Choosing the Best Classification Model for Weather Prediction**

When selecting a machine learning model for weather classification, it’s essential to consider accuracy, interpretability, computational cost, and how well the model handles complex patterns. Below, we analyze three popular classification models: Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) that are best suited for our dataset.

1. Random Forest Classifier

Random Forest is an ensemble learning method that creates multiple decision trees during training. Each tree votes on the final classification, and the majority vote determines the prediction.

Why Use It for Weather Prediction?

* It handles missing data well.
* Works effectively with both numerical and categorical weather features.
* Reduces overfitting because it combines multiple trees.
* Provides feature importance, helping us understand which weather factors impact predictions the most.

Limitations:

* Slower than simpler models because it builds multiple trees.
* Less interpretable than a single decision tree.

1. Gradient Boosting Machines (GBM)

GBM is also an ensemble learning method, but instead of training multiple trees independently (like Random Forest), it builds trees sequentially, each correcting the errors of the previous tree. This leads to better accuracy but requires more tuning.

Why Use It for Weather Prediction?

* Often more accurate than Random Forest due to sequential learning.
* Can detect complex patterns in the data, making it effective for weather classification.
* Reduces both bias and variance, leading to better generalization.

Limitations:

* Computationally expensive training takes longer, especially with large datasets.
* Requires hyperparameter tuning (e.g., learning rate, number of trees).
* More sensitive to outliers than Random Forest.

1. Support Vector Machines (SVM)

SVM is a supervised learning algorithm that finds the optimal decision boundary (or hyperplane) that best separates different weather categories. It works well when there’s a clear separation between classes.

Why Use It for Weather Prediction?

* Works well for small datasets where clear decision boundaries exist.
* Effective when there are a few irrelevant features in the dataset.
* Can use different kernels (linear, polynomial, RBF) to fit different types of data.

Limitations:

* Slow for large datasets (ours has ~1,462 rows).
* Requires careful feature scaling (normalization).
* Difficult to tune hyperparameters like kernel type, C, and gamma.

Which Model Is Best for This Weather Dataset?

Considering that our dataset has weather-related numerical data, we need a model that:

* Handles missing values well
* Works with mixed data types (temperature, precipitation, wind speed, etc.)
* Provides good accuracy and interpretability

Final Recommendation:

* If speed and interpretability matter → Use Random Forest.
* If accuracy is the priority and computational cost is not an issue → Use GBM.
* If dataset is small and decision boundaries are clear → Use SVM (but not ideal here).

The London Weather Analysis Assignment is designed to predict weather types based on pre-recorded weather data values. This assignment utilizes a machine learning model to classify the weather into its correct categories.

Features

* Input Parameters: The code uses a variety of weather-related data such as Date, Precipitation, Maximum Temperature, Minimum Temperature and Wind Speed.
* Prediction: Based on the input features, the machine learning model classifies the weather into its own category, for instance: Sunny, Heavy Rain, Snow, Light Rain, Fog.
* Data Exploration: The dataset includes outliers and a range of values.

Technologies Used

* Backend: Python.
* Data Handling: pandas, NumPy.
* Data Visualization: seaborn, matplotlib.

Dataset

This assignment is provided with the weather data set that will be fed into the machine learning algorithm. The dataset includes:

* date: The date of when the data was recorded.
* precipitation: Decimal values representing precipitation.
* temp\_max: Decimal values presenting the maximum temperature.
* temp\_min: Decimal values presenting the minimum temperature.
* wind: Decimal values representing wind speed.
* weather: Categorical target variable for classification.

**Best Classification Models for this case**.

1. Random Forest Implementation
   1. Handles non-linear relationships and feature interactions (e.g., how wind and temperature jointly affect snow likelihood)
2. K-Nearest Neighbors (KNN)
   1. Classifies based on the majority class of the nearest points.
3. Support Vector Machine Implementation
   1. Suitable for high-dimensional data after encoding categorical features.
4. **Random Forest**

Random Forest (RF) is a commonly utilized supervised learning algorithm used for classification as well as regression tasks. It is a collective learning technique that merges various decision trees to create a strong and precise model. A decision tree is a forecasting model that uses a tree-shaped diagram to outline different potential results and the choices that result them, see Figure 2. In this structure, every internal node represents a choice, whereas each leaf node indicates a predicted result.

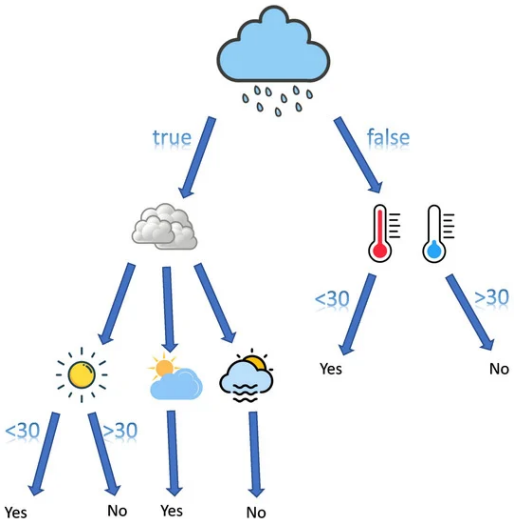


Figure 2

RF models provide multiple benefits, such as the ability to manage high-dimensional data, non-linear correlations, and absent values.

1. **KNNs**

KNN is a widely used supervised learning algorithm. It works by locating the k nearest training samples to a specific test sample and categorizing the test sample into the most frequent class withing its k nearest neighbors, see Figure 3.

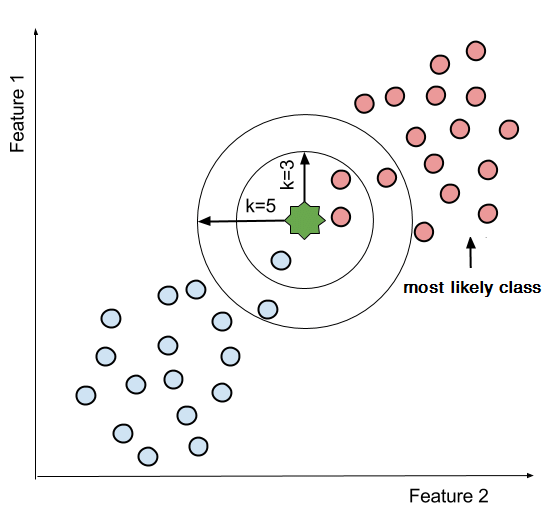


Figure 3

A typical example of a KNN classification for two class problems (e.g., Pink and Blue Circles) when the k parameter is set to 3 and 5. The green star represents a unseen sample point.

Nonetheless, KNN has certain drawbacks, such as being sensitive to outliers and facing challenges when mapping high-dimensional data.

1. **SVMs**

An SVM is a classification tool aimed at finding the best hyperplane that increases the margin between two categories of data points, as illustrated in Figure 1 below.

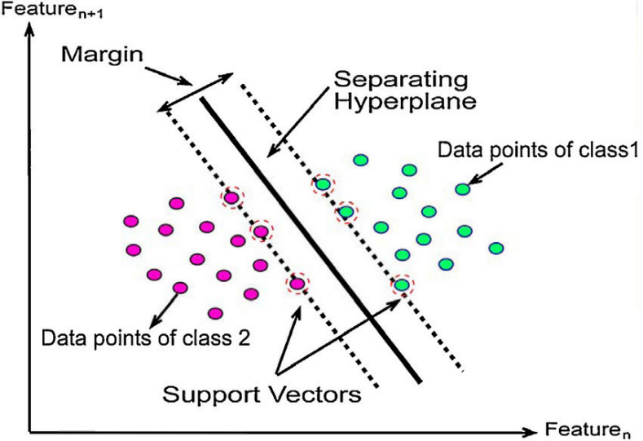


Figure 1

The margin refers to the gap between the hyperplane and the closest data point from both classes. SVMs aim to increase this margin while reducing classification mistakes. SVMs can be divided into linear and non-linear classifications.

**Advantages & Disadvantages**

Random Forest

Advantages:

* Robust and Accurate
* Handles Missing Data

Disadvantages:

* Less Interpretable
* Needs Tuning

KNN

Advantages:

* Simple and adaptable
* No training phase

Disadvantages:

* Slow on large datasets
* Sensitive to noise

SVMs

Advantages:

* Works well in high dimensions
* Avoid overfitting

Disadvantages:

* Needs careful tuning
* Slow on large data sets

**Implementation of Classification Models**

Step 1: Data Preprocessing

Before training any model, we must:

* Handle missing values (e.g., precipitation = 0)
* Split the data into training and testing sets.

Step 2: Train and Evaluate Each Model

* Split data into training and testing sets.
* Use cross-validation to optimize hyperparameters.

Step 3: Compare Model Performance

After running all models, check the accuracy, precision, recall, and F1-score to determine the best one.

If GBM has the highest accuracy, but training time is a concern, Random Forest may be the best option.

**Performance Metrics**

The effectiveness of RF, KNN, SVMs can be measured using three performance metrics:

* Precision
* Recall
* F1-Score

These metrics are utilized to assess the predictive performance of the models.

Precision, determined by Equation (1), is an important performance measure for assessing the correctness of a model’s positive prediction. It specifically assesses the ratio of accurately recognized positive cases among all cases that are labeled as positive. In this measure, True Positives indicates the total of data points accurately recognized as positive by the model, whereas False Positives refers to the count of cases incorrectly labeled as positive when they are negative.

|  |  |
| --- | --- |
|  | (1) |

Recall assesses the model’s ability to correctly recognize positive instances, and it is determined using Equation (2). It assesses the ratio of true positive cases identified out of all the genuine positive instances in the dataset. Recall primarily emphasizes the model’s responsiveness to positive cases and its efficiency in reducing false negatives.

In Equation (2), False Negatives signify the quantity of instances that are truly positive yet are mistakenly labeled as negative by the model.

|  |  |
| --- | --- |
|  | (2) |

F1 Score, determined by Equation 3, merges precision and recall. A rise in the F1 Score, which varies from 0 to 1, signifies enhanced overall performance. It takes into account both precision and recall attaining equilibrium between two metrics.

|  |  |
| --- | --- |
|  | (3) |

# **Conclusion**

The aim of this study was to better understand the climate of London and to predict the weather. To carry out this study, the London weather dataset was used, and a machine learning approach was implemented to make predictions on the data set, enhancing the accuracy and detail of weather forecasts, particularly when applied to short-term predictions. This report demonstrates how to use a set of Machine Learning based models and large amounts of data to make informed weather or climate-related predictions within hours.

1. There can be an end-to-end data analysis on different temporal in weather, while still providing a foundation for forecasting other weather conditions.
2. Can be used as a tool to tackle weather prediction including, for example, tuning such tools for a seamless prediction framework

However, based on the interpretability of the model, further research will be required to fully integrate machine learning into Climate Insights’ operational weather forecasting system, addressing the need for reliable predictions across diverse weather conditions.

# **References**

<https://www.ecmwf.int/sites/default/files/elibrary/092024/81582-evaluation-of-ecmwf-forecasts_0.pdf>

<https://www.climatechange.ai/blog/2024-02-07-forecast-tutorials>

Climate Science Investigations South Florida - Global Wind Patterns (2023) www.ces.fau.edu. Available at: <https://www.ces.fau.edu/nasa/content/resources/global-wind-patterns.php>.

<https://serc.carleton.edu/eslabs/weather/2a.html>

Tahafaisalkhan. (n.d.). *Tahafaisalkhan/weather-classifier: Weather type classification is a machine learning project that predicts weather types-rainy, sunny, cloudy, or snowy-using a synthetically generated dataset. this project focuses on classification algorithms, data preprocessing, and outlier detection, making it ideal for honing data science and machine learning skills.*

GitHub. <https://github.com/tahafaisalkhan/weather-classifier>

Aldous, K. (2023, May 2). *Comparing the Results: Classification vs. Regression Models in Machine Learning*. Team Acua.

<https://acua.qcri.org/blog/comparing-the-results-classification-vs-regression-models-in-machine-learning/>

Safia, M., Abbas, R., & Aslani, M. (2023). *Classification of weather conditions based on supervised learning for Swedish cities. Atmosphere, 14(7), 1174.* <https://doi.org/10.3390/atmos14071174>

*GeeksforGeeks. (2025, January 29). Classification vs Regression in Machine Learning.* GeeksforGeeks. <https://www.geeksforgeeks.org/ml-classification-vs-regression/>

Perera, J. (2022, January 6). *Applying linear regression on a weather dataset - Analytics Vidhya - medium.*

*Medium*. <https://medium.com/analytics-vidhya/applying-linear-regression-on-a-weather-data-set-e84901120f88>

Ak, A. (2023, October 9). *Regression vs. Classification: Understanding the Differences and Use Cases.*

*Medium*. <https://medium.com/@abelkuriakose/regression-vs-classification-understanding-the-differences-and-use-cases-6cd51c236112>

Dattaradon, G. (2024, February 26). *Advancing Weather Prediction : Comprehensive Classification Models using Python.*

*Medium*. <https://medium.com/@gigi.dattaradon/weather-prediction-with-python-a-comprehensive-analysis-of-classification-models-f7149d39995d>

<https://www.springboard.com/blog/data-science/regression-vs-classification>

<https://shannwang.github.io/Machine_Learning/supervised_learning_classifying_weather_conditions.html>

<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>

<https://www.datacamp.com/blog/classification-machine-learning>

<https://www.researchgate.net/figure/A-typical-example-of-a-KNN-classification-for-a-two-class-problem-ie-the-pink-and_fig2_322358139>