Project Milestone 2

Project Title - Predicting Rain with Machine Learning

Group 8 Members:

Dylan D' Andrea - (Team Leader)
Jainam Shah
Leon Silas
Matthew Olajide

Preliminary Project Statement :-

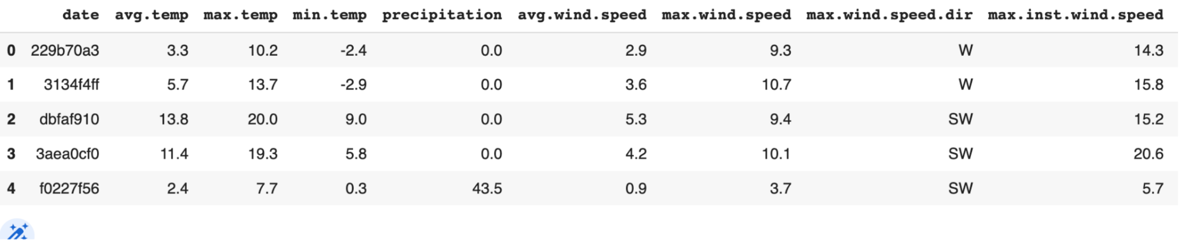
The agriculture industry is heavily reliant on weather conditions and providing forecasts and analysis of them is a surplus for farmers in order to make informed decisions on planning, cost-saving, environmental impact, and maximizing yield.

However, weather prediction involves numerous factors and predicting weather is very much possible but complete accuracy is not guaranteed.The accuracy of prediction depends on various factors such as quality and quantity of historical data and complexity of weather patterns in that specific region. Moreover, localised analysis may be necessary to make more precise predictions for specific farms and fields.

Overall, with proper data and tools, it is possible to forecast upcoming weather conditions which can help agricultural companies optimize their production, reduce costs, and make better decisions on farming activities such as planting and irrigation.

Dataset :-

The dataset has been split into parts i.e. training set and testing set. In each of the sets, there is weather data consisting of anonymized locations names from Region A to Region E.

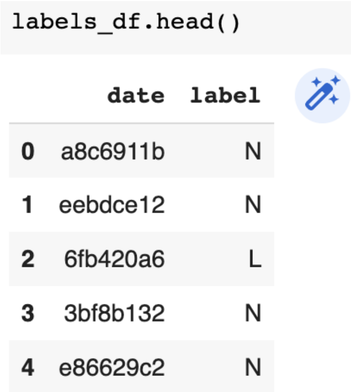


	date	avg.temp	max.temp	min.temp	precipitation	avg.wind.speed	max.wind.speed	max.wind.speed.dir	max.inst.wind.speed
0	229b70a3	3.3	10.2	-2.4	0.0	2.9	9.3	W	14.3
1	3134f4ff	5.7	13.7	-2.9	0.0	3.6	10.7	W	15.8
2	dbfaf910	13.8	20.0	9.0	0.0	5.3	9.4	SW	15.2
3	3aea0cf0	11.4	19.3	5.8	0.0	4.2	10.1	SW	20.6
4	f0227f56	2.4	7.7	0.3	43.5	0.9	3.7	SW	5.7

Figure 1: Glimpse of the dataset

As it is seen from the dataset, the "date" column is anonymized to some random values. There are in total 10 features in the dataset which consist of temperature, wind speed, precipitation, wind speed direction and atmospheric pressure.

The dataset contains 5 csv files in each training and testing set along with a separate csv file named "solution_format.csv" containing target rain predictions for each of the dates, which allows us to use supervised learning when building the model.



```
labels_df.head()
```

	date	label
0	a8c6911b	N
1	eebdce12	N
2	6fb420a6	L
3	3bf8b132	N
4	e86629c2	N

Figure 2: Solution_format.csv

GOAL:-

The goal currently is to predict the weather for the next day based on three labels:

- N - No rain
- L - Light Rain
- H - Heavy Rain

Tools & Technology:-

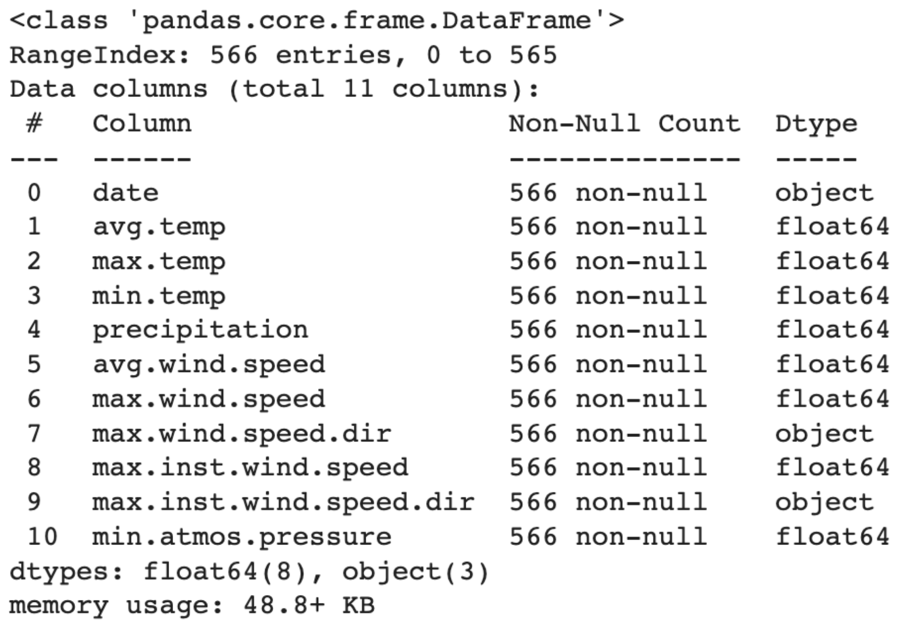
In order to perform weather forecasting for rain occurence, the following tools and technology are going to be utilised:-

1. Python - To implement weather forecasting using machine learning
2. Jupyter- Notebook - To implement Python Code
3. Matplotlib Library in python - To perform EDA

Currently, the above mentioned tools and technology are being employed to predict weather conditions , if time permits and if we can add one or more functionality to make weather prediction more easier for any user, more tools and technologies can be employed.

Preliminary Exploratory Data Analysis (EDA)

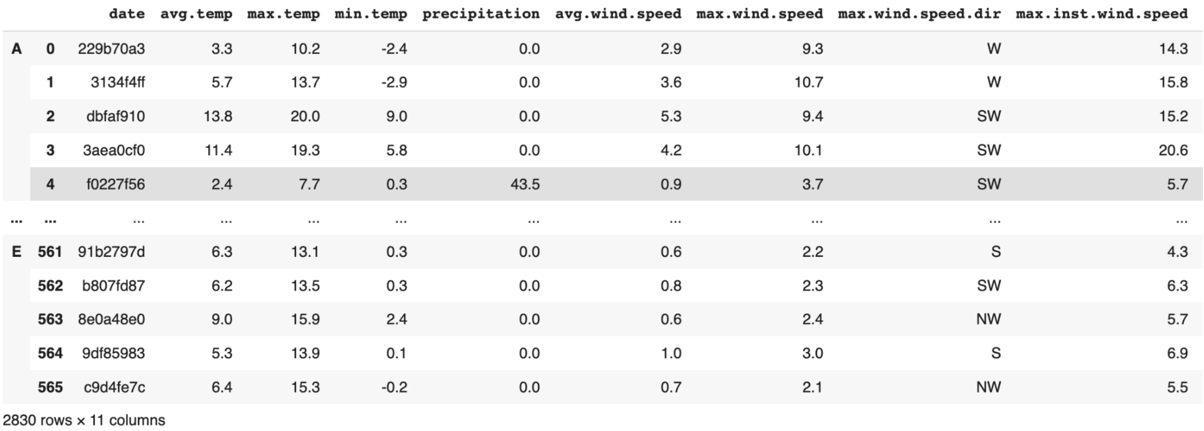
Taking overview of training set of data.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 566 entries, 0 to 565
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   date                     566 non-null    object 
 1   avg.temp                 566 non-null    float64
 2   max.temp                 566 non-null    float64
 3   min.temp                 566 non-null    float64
 4   precipitation            566 non-null    float64
 5   avg.wind.speed           566 non-null    float64
 6   max.wind.speed           566 non-null    float64
 7   max.wind.speed.dir       566 non-null    object 
 8   max.inst.wind.speed      566 non-null    float64
 9   max.inst.wind.speed.dir  566 non-null    object 
 10  min.atmos.pressure       566 non-null    float64
dtypes: float64(8), object(3)
memory usage: 48.8+ KB
```

Figure 3: Overview of data

Now joining all the regions together as they share a primary key "date" with concat() function.



		date	avg.temp	max.temp	min.temp	precipitation	avg.wind.speed	max.wind.speed	max.wind.speed.dir	max.inst.wind.speed
A	**0**	229b70a3	3.3	10.2	-2.4	0.0	2.9	9.3	W	14.3
	1	3134f4ff	5.7	13.7	-2.9	0.0	3.6	10.7	W	15.8
	2	dbfaf910	13.8	20.0	9.0	0.0	5.3	9.4	SW	15.2
	3	3aea0cf0	11.4	19.3	5.8	0.0	4.2	10.1	SW	20.6
	4	f0227f56	2.4	7.7	0.3	43.5	0.9	3.7	SW	5.7
...	**...**	...	...	...	...	...	...	...	...	...
E	**561**	91b2797d	6.3	13.1	0.3	0.0	0.6	2.2	S	4.3
	562	b807fd87	6.2	13.5	0.3	0.0	0.8	2.3	SW	6.3
	563	8e0a48e0	9.0	15.9	2.4	0.0	0.6	2.4	NW	5.7
	564	9df85983	5.3	13.9	0.1	0.0	1.0	3.0	S	6.9
	565	c9d4fe7c	6.4	15.3	-0.2	0.0	0.7	2.1	NW	5.5

2830 rows × 11 columns

Figure 4: Dataset after joining all training datasets

We don't want the regions as the index, so we reset the index and then rename some columns to get the data in the right shape.

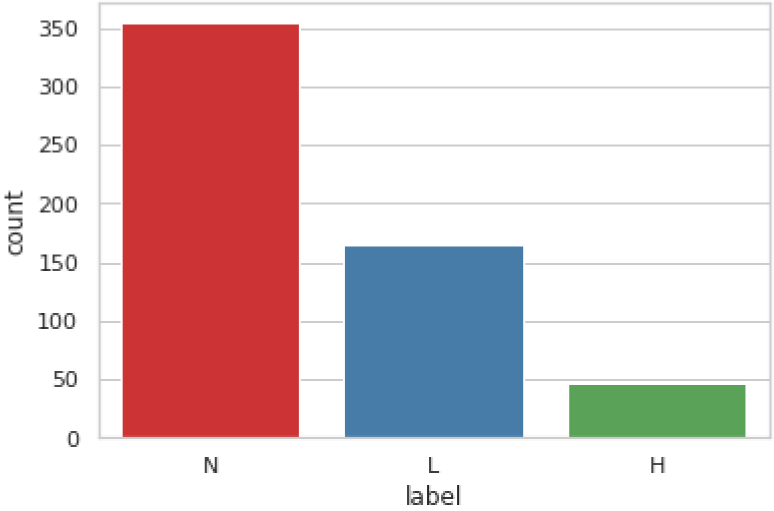


Figure 5: Visualizing Target Class

As it is visible from the data, we have an imbalanced class, as label "N" is the dominating rest of the classes.

This means that model can be biased towards classes with larger samples. This happens because the classifier has more information on classes with more samples, so those classes will be predicted better than smaller classes. It means Label “N” will be predicted more.

As all features are used to predict the next day’s weather. Let’s see whether all regions share similar patterns or whether any outliers or anomalies exist.

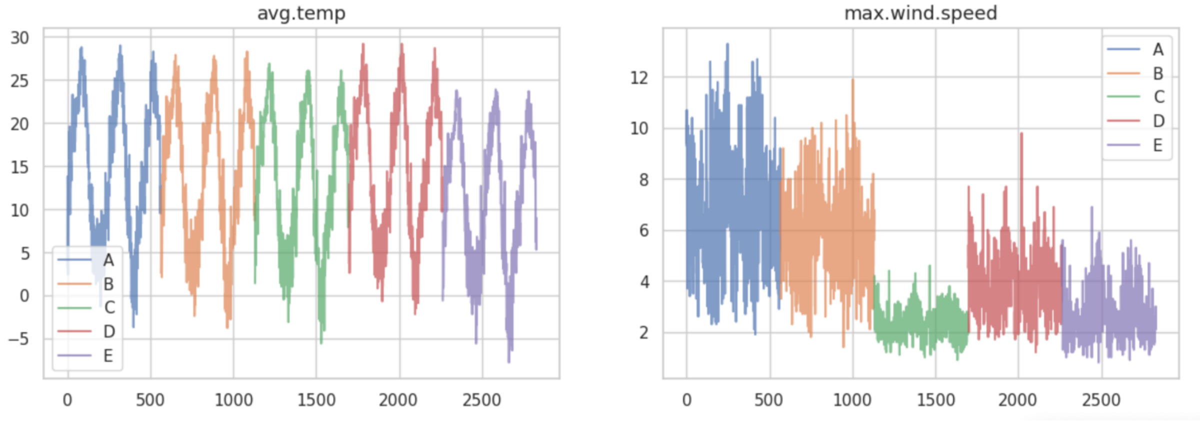


Figure 6: Average temperature and maximum wind speed in all regions

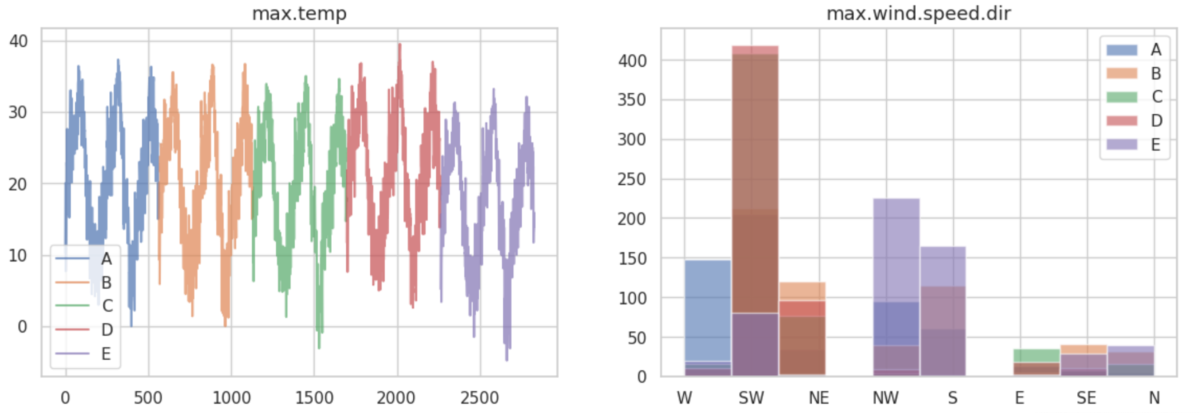


Figure 7: Maximum Temperature and maximum wind direction in all regions

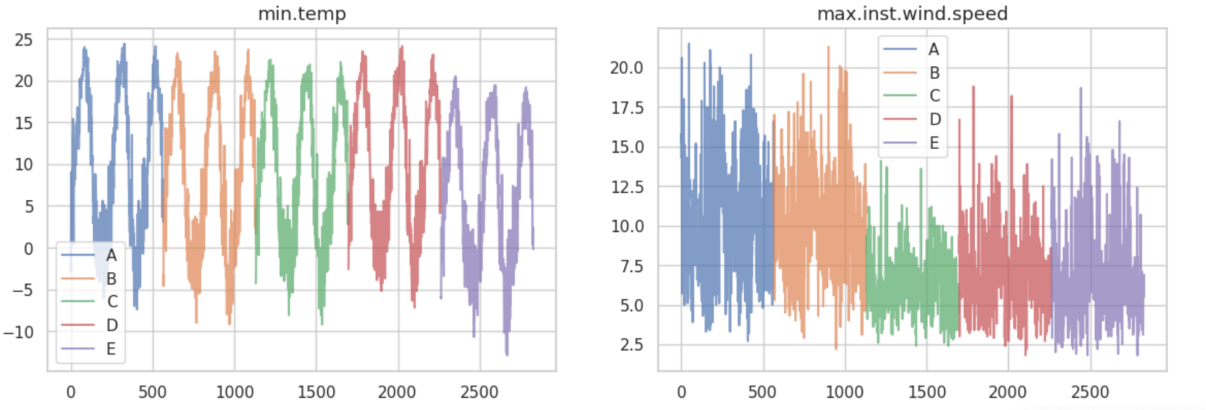


Figure 8: Minimum Temperature and maximum instant wind speed in all regions

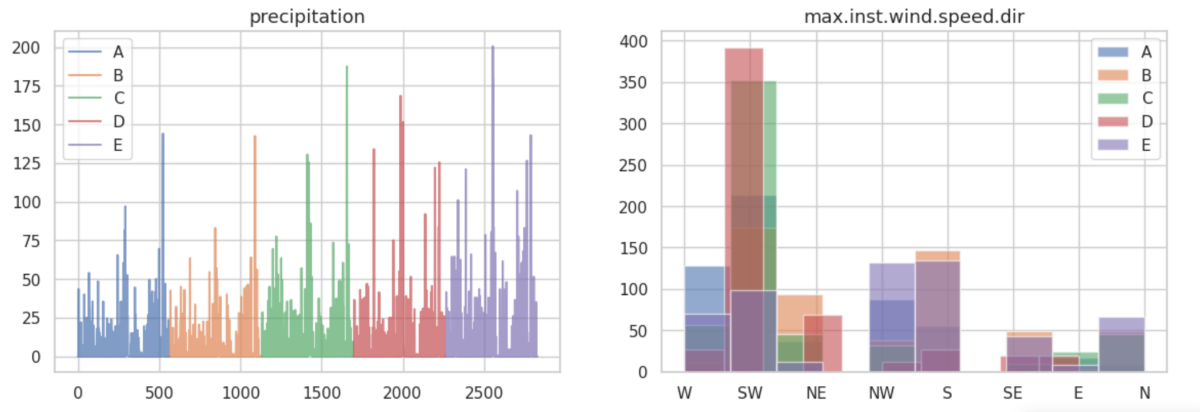


Figure 9: Precipitation and maximum instant wind speed in all regions

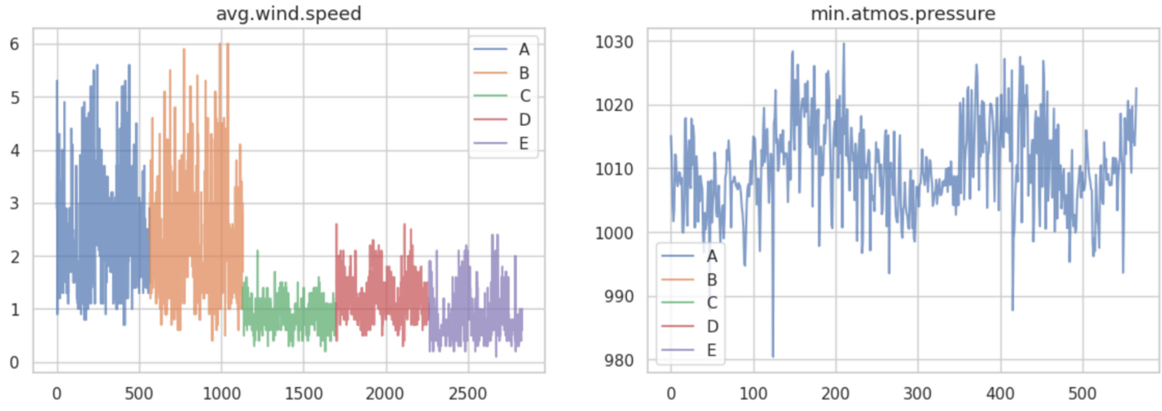


Figure 10: Average wind speed and minimum atmospheric pressure in all regions

From the plots, there are patterns in the data that are very similar except for regions C,D,E where minimum wind speed and average wind speed are on a lower scale.

Now that preliminary EDA is complete, there is further requirement of checking the dataset about missing values and to create the model to predict weather conditions for rain.