Computer Science Department
Dr. Radi Jarrar
MACHINE LEARNING COMP4388
Section: 1 T, R.

The Weather Dataset
Al-Bireh Municipality

Project two

30. January. 2022

Dear Al-Bireh Municipality representative,

We are pleased to present our analysis of the weather dataset you provided. In this report, we aim to provide conclusions that can help you make informed decisions about your operations concerning the patterns and trends in your weather data. In which you develop strategies plans to ensure public safety and infrastructure planning in different regions of the city.

With respect,

Team:

Sondos Aabed, Manal Makho
1190652, 1191114

Table of Contents

Introduction: Weather dataset analysis

In this document, the supervised machine learning process applied to the Weather dataset is presented. The problem was framed as a binary classification problem. The objective is to build a classification model that predicts whether it will rain tomorrow or not with high accuracy.

The dataset contains 36529 observations and 21 features. The observations are weather conditions days of a specific region including: data, location, minimum temperature, maximum temperature, rain fall, wind direction, wind speed, At 9 am and 3pm: wind direction, wind speed, Humidity, pressure, cloud and temperature and whether it rained that day or not.

The analysis will start by Exploratory Data Analysis in Chapter one. Then will move on to build the classification model. Finally, the conclusion will summarize the findings and insights gained to Al-Bireh Municipality to make informed decisions concerning public safety and infrastructure planning for different regions in Al-Bireh.

Chapter 1: Exploratory Data Analysis (EDA)

Chapter one focuses on Exploratory Data Analysis (EDA) of the Weather dataset. It will uncover patterns, relationships, and distributions of the data. It will also identify problems such as missing data and outliers. After conducting EDA, the features in which will be used for modeling will be selected in the last section.

1.1 Features overview

The features data type varies and they are divided into two categories, Quantitative and Qualitative features. The objective of this section is to have a clear understating of the features using graphs and statistical summaries.

1.1.1 Quantitative Features summary

These are 14 features, include: Min Temp, Max Temp, Rain Fall, Wind Speed, Wind speed 9am, Wind speed 3pm, Humidity 9am, Humidity 3pm, Pressure 9am, Pressure 3pm, Cloud 9am, Cloud 3pm, Temp 9am and Temp 3 pm. Table (1) below, shows useful statics for each quantitative attribute that includes the following: mean, standard deviation, minimum, quartiles and maximum.

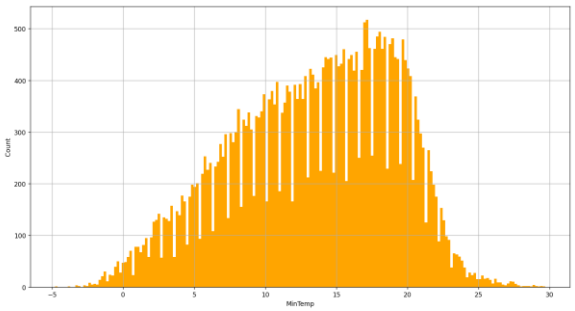
	Count	***Mean***	***Standard deviation***	***Minimum***	***25%***	***50%***	***75%***	***Maximum***
Min Temp	36029	13.41	5.72	- 4.8	9.2	14	18	29.7
Max Temp	36160	23.96	5.91	6.8	19.5	23.4	27.5	47.3
Rain Fall	35839	2.70	9.41	0	0	0	0.8	371
Wind Speed	31597	38.36	13.71	7	30	37	46	135
Wind speed 9am	35698	12.66	9.13	0	6	11	19	130
Wind speed 3pm	35054	17.41	9.32	0	11	17	22	83
Humidity 9am	35862	70.03	17.51	3	58	71	83	100
Humidity 3pm	35196	52.51	20.39	1	37	53	67	100
Pressure 9am	29837	1018.26	6.65	980.5	1013.8	1018.3	1022.7	1039.9
Pressure 3pm	29846	1015.70	6.60	979	1011.3	1015.8	1020.2	1037
Cloud 9am	20712	4.31	2.91	0	1	5	7	9
Cloud 3pm	20390	4.40	2.70	0	2	5	7	8
Temp 9am	36092	17.86	5.29	0.3	14.1	18.3	21.8	37.7
Temp 3 pm	35424	22.49	5.75	6.4	18.3	21.8	25.9	46.7

From the table, it is seen that the humidity values at 9am and 3pm are different, 9am values has higher average. This is an indication that the humidity values changes during the day.

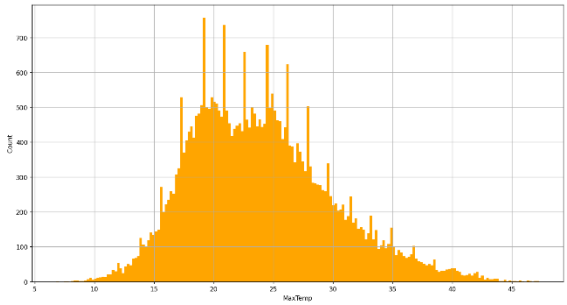
The pressure values at 9am and 3pm are slightly different, with 9am values being slightly higher.

However, there might be potential outliers, the maximum wind speed is 135 when most of the data is below 46. This will be investigated and handled in coming section.

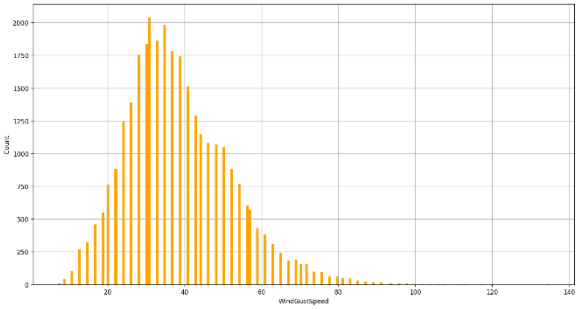
In order to visualize the distribution of the quantitative attributes, univariate histograms were used as follow:



Max Temp



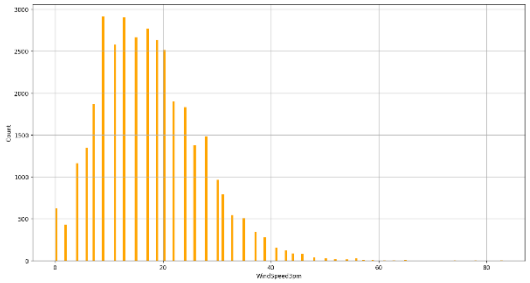
Min Temp



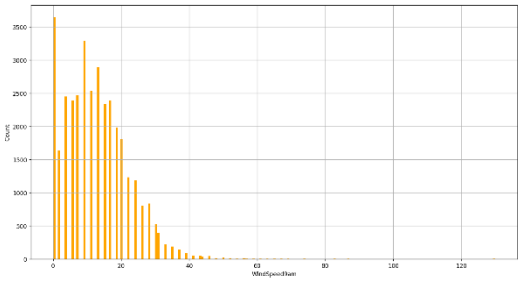
Wind Gust speed

Figure (1)

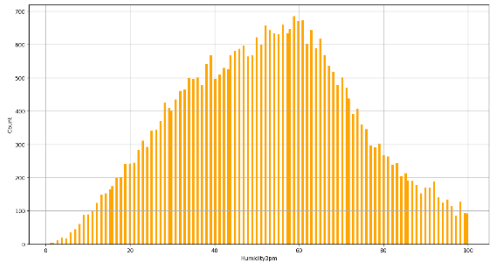
In figure (1), it is shown that the wind gust speed is right skewed distribution with the observation of outliers at the maximum values. As for the min temp distribution, it is very close to bimodal. And the Max Temp it seems to be right skewed.



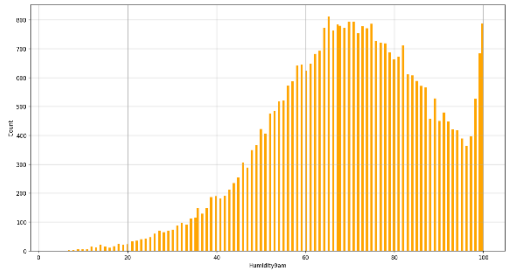
Wind speed 3 pm



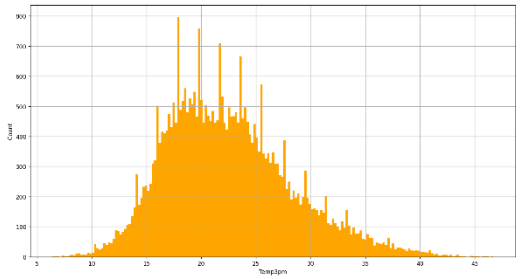
Wind speed 9 am



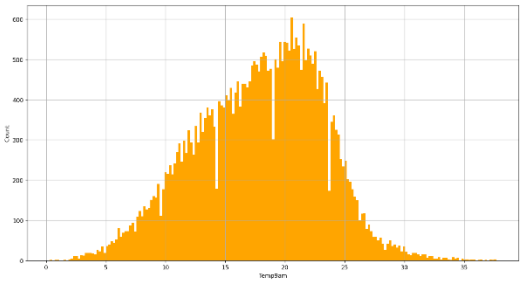
Humidity 3 pm



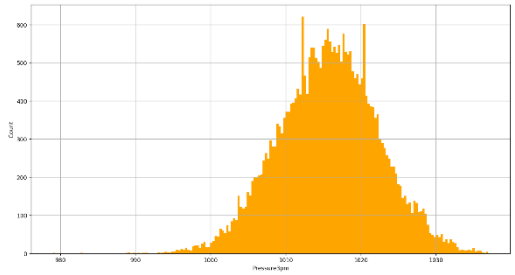
Humidity 9 am



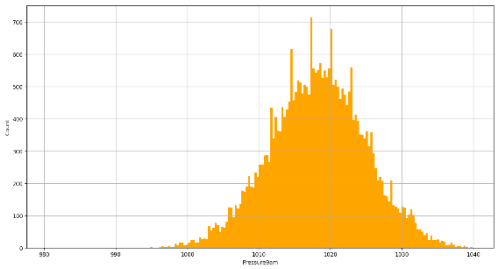
Temp 3 pm



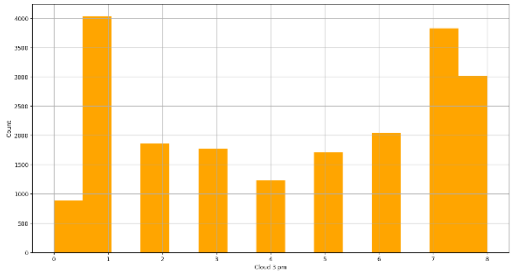
Temp 9 am



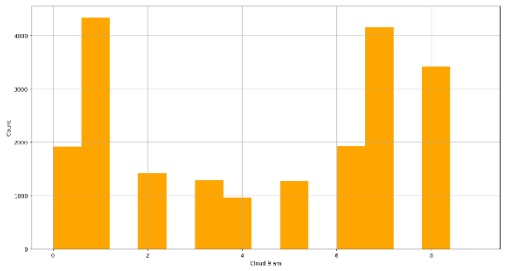
Pressure 3 pm



Pressure 9 am



Cloud 3 pm



Cloud 9 am

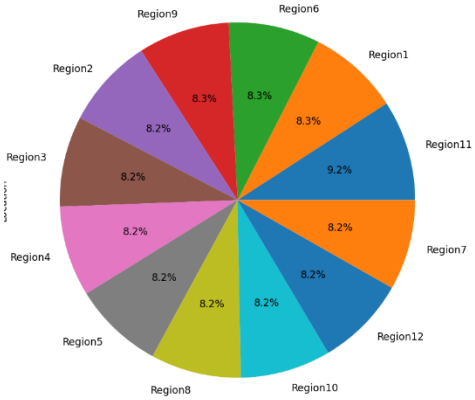
Figure (2)

In the above figure (2), it is shown that both pressure at 9 am and 3 pm are symmetrically distributed. As for the Temperature at both 9 am and 3 pm its distribution is skewed to the right. The cloud at both 9 am and 3 pm seems to be bimodal distributed.

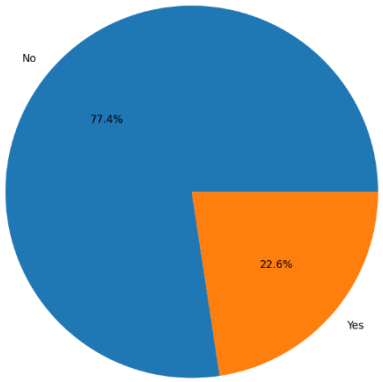
As for Humidity at 9 am it is edge peak distributed, it has a large peak at right tail. Humidity at 3 pm is a little skewed to the right. Both wind speed at 9 am and at 3 am has right-skewed distribution.

1.1.2 Qualitative Features summary

These are 6 features, include: date, location, WindDir9am, WindDir3pm, rain today and rain tomorrow. Figure (3), shows that the observations were collected from 12 different regions, region 1 – region 12. It also shows that 22.6% rained in that specific date and 77.4% it did not rain.

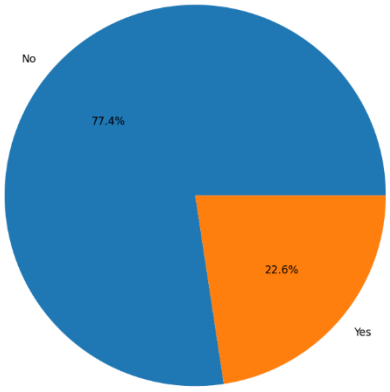


Location



Rain Today

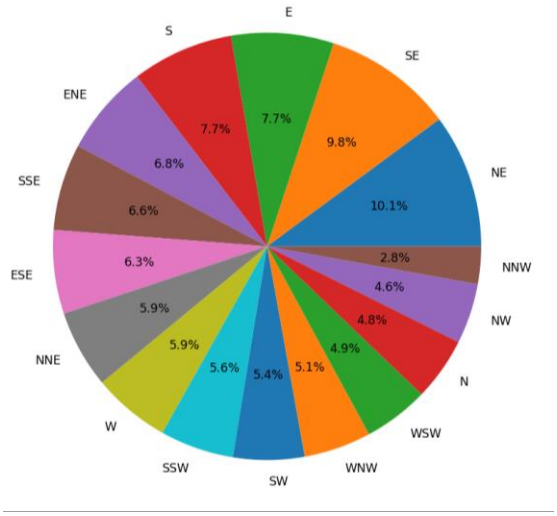
Figure (3)



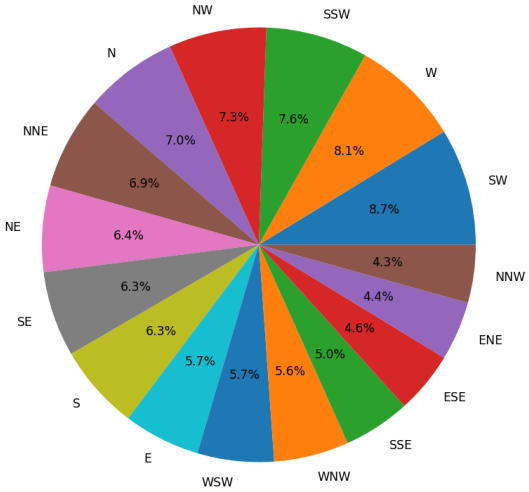
Rain Tomorrow

Figure (4)

As for figure (4), it shows that 22.6% of the observation the day after today was actually raining and 77.4% it did not rain.



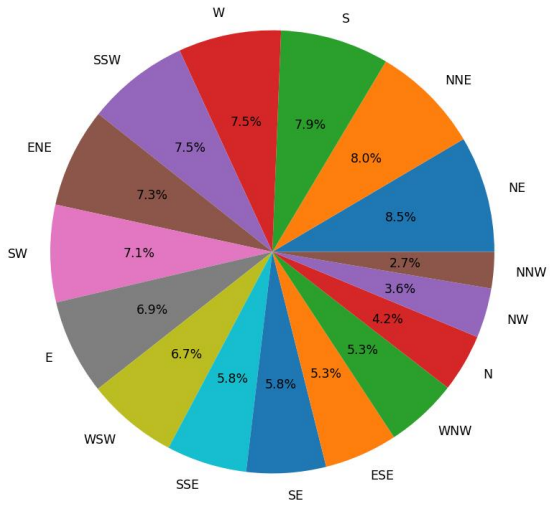
Wind 3 pm



Wind 9 am

Figure (5)

From figure (5) wind at both 3 pm and 9 pm, doesn't seem to be dominated at one direction. It is noticed that NE (North East) at 3 pm has 10.1% wind of the wind at 3 pm feature. For 9 am SW (South West) is 8.7% of the Wind at 9 am feature.



Wind Gust Direction

Figure (6)

Figure (6) shows wind gust direction, it shows that 8.5% of the days that feature is NE (North East) direction.

1.2 Data Cleansing

In this section of chapter one, data cleaning process is performed to ensure the data used in the modeling has high quality and has no errors. The objective is to handle missing data, detect and handle outliers. The duplicates rows were removed.

1.2.1 Missing Data detection and Handling

The below table (2), shows the count of the missing data for each feature. The table indicates that we have a lot of missing data detected for the features out of 36529 observations.

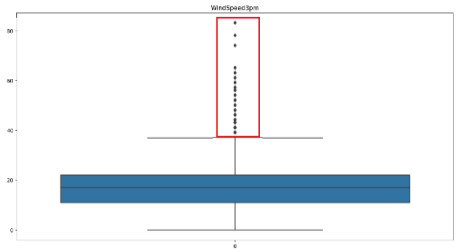
Feature	*Count*	*Missing Data Count*
Min Temp	36029	500
Max Temp	36160	366
Rain Fall	35839	687
Wind Speed	31597	4929
Wind speed 9am	35698	828
Wind speed 3pm	35054	1472
Humidity 9am	35862	664
Humidity 3pm	35196	1330
Pressure 9am	29837	6689
Pressure 3pm	29846	6680
Cloud 9am	20712	15814
Cloud 3pm	20390	16136
Temp 9am	36092	434
Temp 3 pm	35424	1102
Wind Gust Dir	31593	5936
Wind Dir 9 am	32052	4477
Wind Dir 3 pm	34426	2103
Rain Today	35839	690
Rain Tomorrow	35839	690

Table (2)

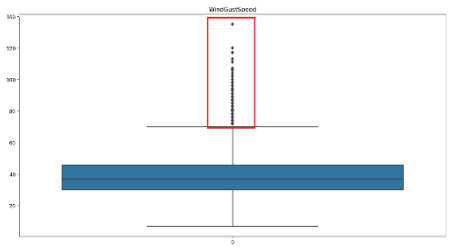
Those missing data were handled by using KNN model to impute the predicted missing value for numerical features. As for the non-numerical features they were imputed by the mode of each feature.

1.2.2 Outliers detection and Handling

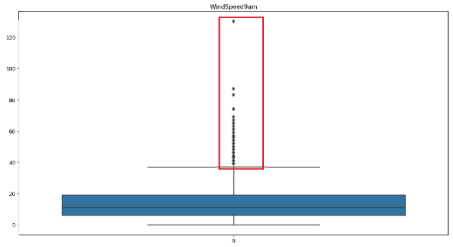
In the features overview section, outliers were indicated in the statistical summary initially by taking a look at the min, max, standard deviation and mean, this is why in this section the detection of outliers is performed using plot boxes shown below and then handled.



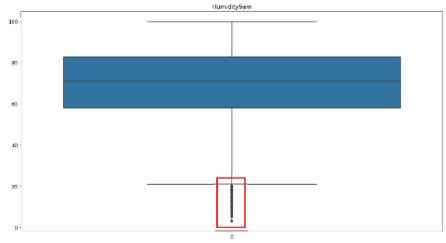
Wind speed 3 pm



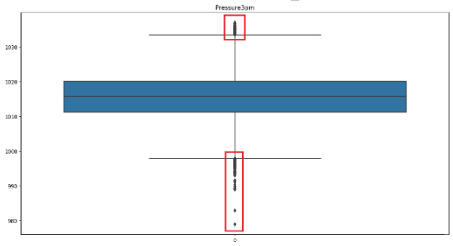
Wind Gust Speed



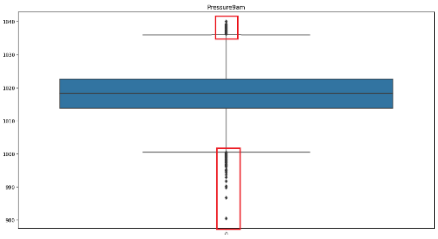
Wind speed 9 am



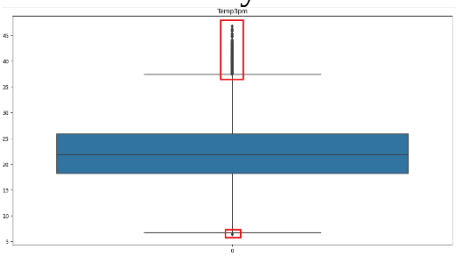
Humidity 9 am



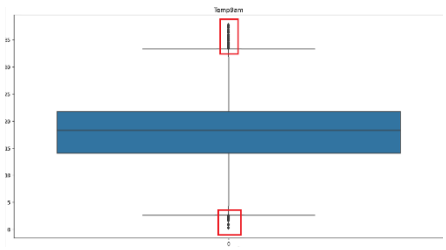
Pressure 3 pm



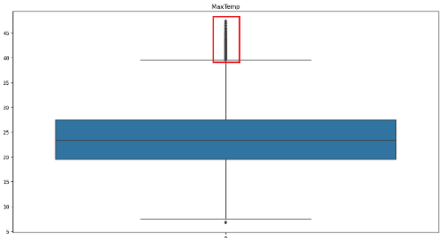
Pressure 9 am



Temp 3 pm



Temp 9 am



Max Temp

Figure (7)

These 9 features were detected to have outliers using plot box. The outliers are the dots surrounded by the red rectangle in each of the plot boxes.

The following scatter plot shows the Rain Fall feature. Which detect some outliers to the feature as shown in figure (8):

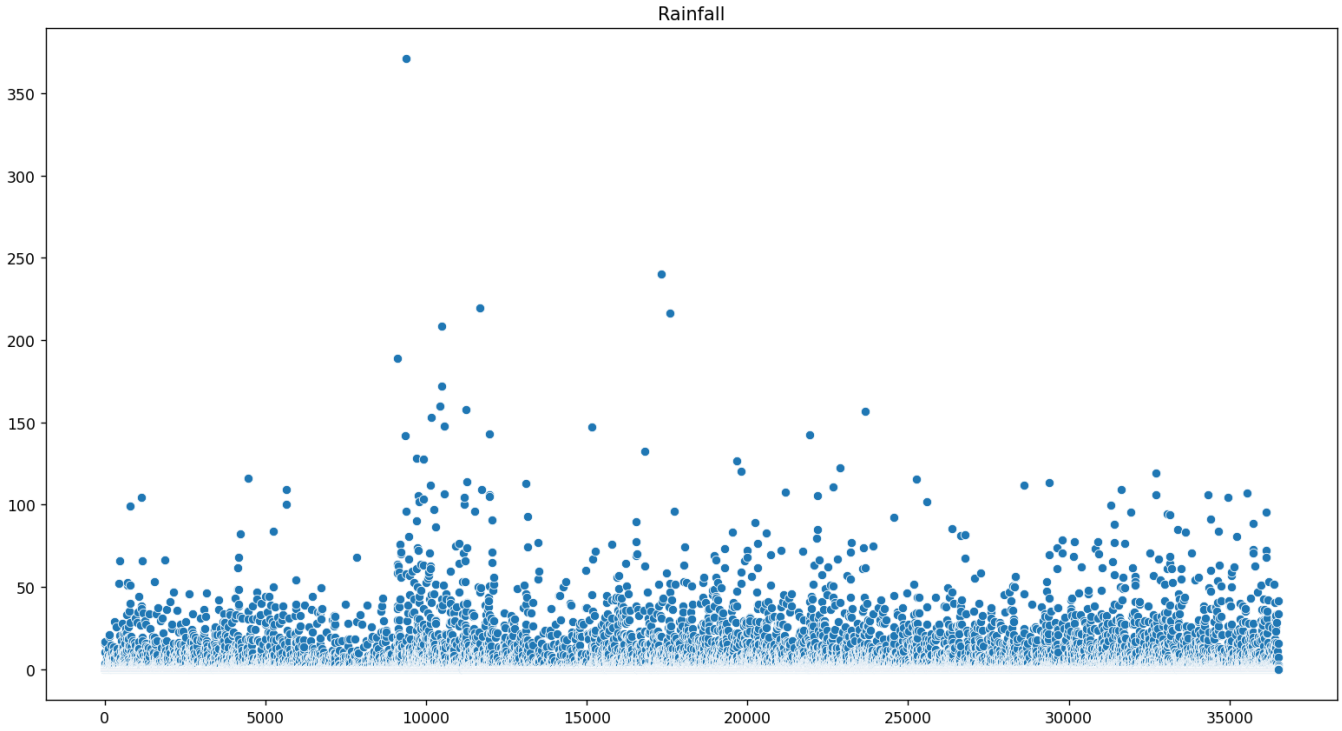


Figure (8)

As for handling the detected outliers, the values that fall outside of the lower and upper bounds of the feature was replaced with the bounds themselves. (Capping and flooring).

After performing data cleansing, the count of observations is 36421 and it was 36529, as for the features they are now 21 feature.

1.3 Data processing

After the data cleaning and the features overview, the decision was made to perform some processing of the data. The objective of this section is to perform features discretization and features scaling.

It is noticed that Date feature has high cardinality (very high number of unique values) on its nature so it needed to be processed. The date feature was divided into Month, day and year features.

1.3.1 Features Scaling

After the features missing values and outliers were handled, by looking at the ranges of different features in the following table(3):

Feature	*Minimum*	*Maximum*
Min Temp	- 4.8	29.7
Max Temp	7.75	39.35
Rain Fall	0	2
Wind Speed	7	70
Wind speed 9am	0	38.5
Wind speed 3pm	0	38.5
Humidity 9am	20.5	100
Humidity 3pm	1	100
Pressure 9am	100.2	1035
Pressure 3pm	999.9	1031.9
Cloud 9am	0	9
Cloud 3pm	0	8
Temp 9am	2.55	33.35
Temp 3 pm	6.9	37.3

Table (3)

It is noticed that the pressure at both 9 am and 3pm has dominated ranges and there are different ranges for other features. It is decided at this point to perform feature scaling and train for different models that performs better for feature scaling.

Because outliers were handled in the previous section, min-max scaler was used for features scaling.

1.4 Data Correlation

In this section, data correlation will be tested and features in which will be used in the classification task will be selected.

1.4.1 Multivariate analysis

In the following page the heat map is shown, were each feature was found to be perfectly correlated with itself, the diagonal values are all equal to 1.

There is high correlation between day and month given the attributes are related, the decision is to keep the month feature and drop the day feature they are positively correlated by 1.

It was also noticed that the year feature is negatively correlated with the target feature so it will be dropped.

Rain fall feature and rain today feature were found to be highly positively correlated with 0.95. The decision is to keep rain fall and drop rain fall because it had higher correlation by 0.01 with the target class rain tomorrow.

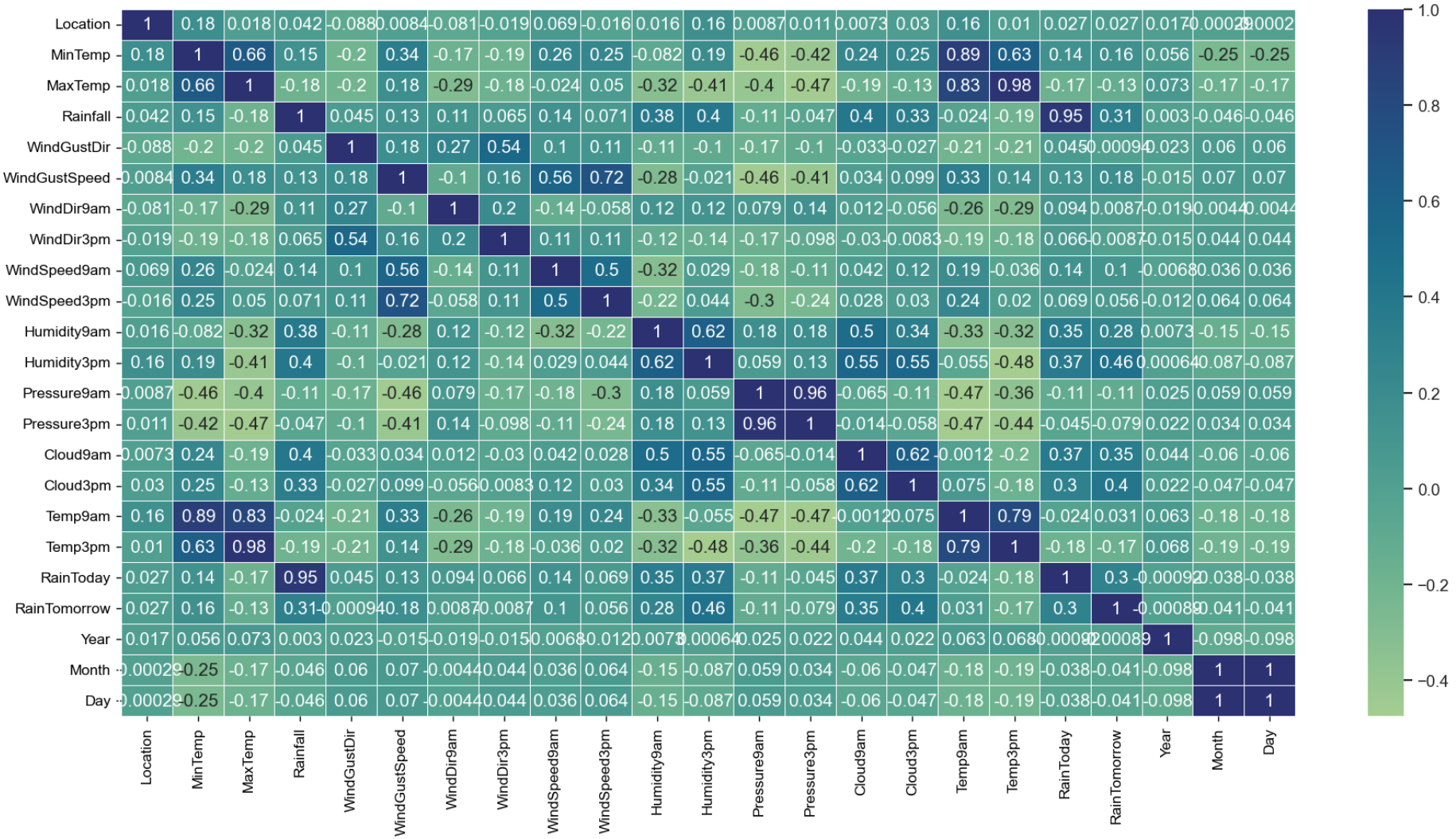
Both pressure at 3 pm and at 9 am was found to be negatively correlated with the target class by -0.11 and -0.079. The decision is to drop both of them.

Maximum temperature and temperature at 3 pm were found to highly correlate by 0.98. So temperature at 3 pm will be dropped. As for Minimum temperature and temperature at 9 am by 0.89 so temperature at 9 am will be dropped.

Wind gust speed has positive correlation with both wind speed at 9 am and wind speed at 3 pm with 0.65 and 0.72. Wind gust speed will be kept and wind speed at both 3 pm and 9 am will be dropped.

Wind gust direction was found to be very negatively correlated with the target feature so it will be dropped too.

Cloud at both 9 am and 3 pm was found to be positively correlated with the target feature they will both be kept.



Heat Map

Figure (9)

1.4.2 Features selection

After multivariate analysis using heat map, these attributes will be **dropped**:

1- Day
2- Year
3- Rain fall
4- Wind gust direction
5- Pressure at 3 pm
6- Pressure at 9 am
7- Temperature at 3 pm
8- Temperature at 9 am
9- Wind speed at 3 pm
10 - Wind speed at 9 am

These attributes will be **kept** for the learning process:

1- Month
2- Location
3- Min Temp
4- Max Temp
5- Humidity at 3 pm
6- Humidity at 9 am
7- Wind Gust Speed
8- Wind direction at 9 am
9- Wind direction at 3 pm
10 - Cloud at 3 pm
11 - Cloud at 9 am
12 - Rain today

Chapter 2: Classification of Tomorrow Rain

Now that the EDA was performed and the attributes was cleansed and selected, it is split into test and training sets, based on 20:80 percentage.

In this section, different Classification Algorithms will be used in the weather dataset to predict the rain tomorrow feature. At the end of this section a comparison of performance for the selected three models will be presented.

2.1 Classification Algorithms

Three models are trained on the weather dataset. These models are: Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Network (ANN).

2.1.1 Explanation

Following a briefly explanation of each algorithm:

Logistic Regression (LR):

Used to model the relationship between the dependent variable (rain tomorrow or not) and independent variables such as temperature, humidity, etc. by estimating the probabilities of the dependent variable using a logistic function.

Support Vector Machine (SVM):

Used to find the best hyperplane or a set of hyperplanes that separates the data into the two classes (rain tomorrow or not). This can be used to make binary predictions of whether it will rain tomorrow or not based on the input features.

Artificial Neural Network (ANN):

Designed to mimic the structure and function of the human brain. ANNs can be used to model complex relationships between the input features and the target variable (whether it will rain tomorrow or not).

2.1.2 Implementation

Following, The details of each models implementation is presented. As for the analysis of the training and test error, Bias- variance tradeoff analysis: the cross validation was performed and the learning graph of the model was plotted

Logistic Regression (LR):

Table (4) below shows the confusion matrix of applying Logistic Regression classifier into the weather dataset to predict rain tomorrow:

	Predicted Positive	***Predicted negative***
Actual positive	*5374*	*334*
Actual Negative	*827*	*750*

Table (4)

The LR classifier achieved a precision of 0.69 and the AUC score was 0.85. The performance of the LR classifier on weather dataset was good, with a high AUC score.

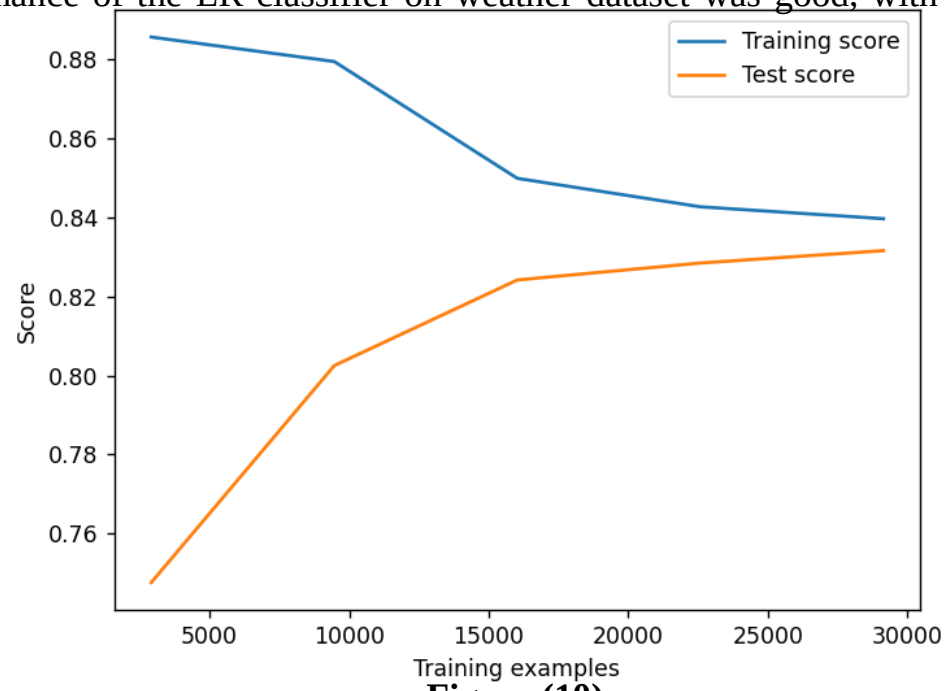


Figure (10)

The training score is very high when using little samples for training. It decreases when the number of samples increases. The test score is low at the beginning and then increases when adding samples. The training and test scores become more realistic when all the samples are used for training. This suggests that the model might be underfitting to the training data.

Support Vector Machine (SVM):

The below **table (5)** shows the confusion matrix of applying Support Vector Machine classifier into the weather dataset to predict rain tomorrow:

	Predicted Positive	***Predicted negative***
Actual positive	*5507*	*199*
Actual Negative	*909*	*670*

Table (5)

The SVM classifier achieved a precision of 0.77 and the ROC/AUC score was 0.82. The performance of the SVM classifier on weather dataset was good, with a higher precision score and a relatively high ROC/AUC score.

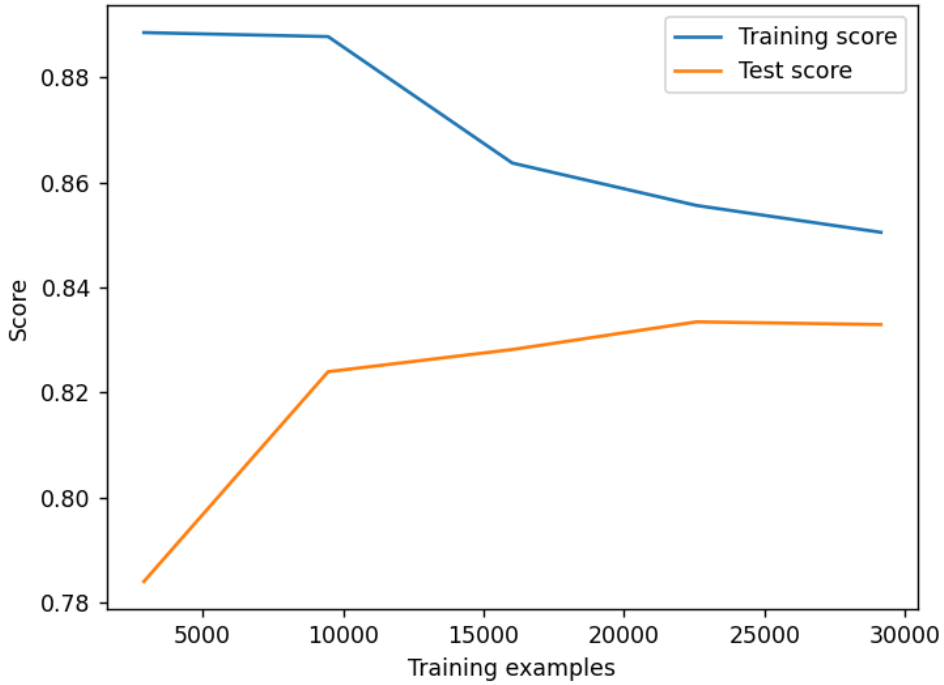


Figure (11)

The test score is lower than the training score. It is not generalizing too well to new data.

Artificial Neural Network (ANN):

Table (6) shows the confusion matrix of applying Artificial Neural Network classifier into the weather dataset to predict rain tomorrow:

	Predicted Positive	***Predicted negative***
Actual positive	*5485*	*195*
Actual Negative	*910*	*695*

Table (6)

The ANN classifier achieved a precision of 0.78 and the ROC/AUC score was 0.86. The performance of the ANN classifier on weather dataset was relatively good, with a high precision score and a relatively high ROC/AUC score.

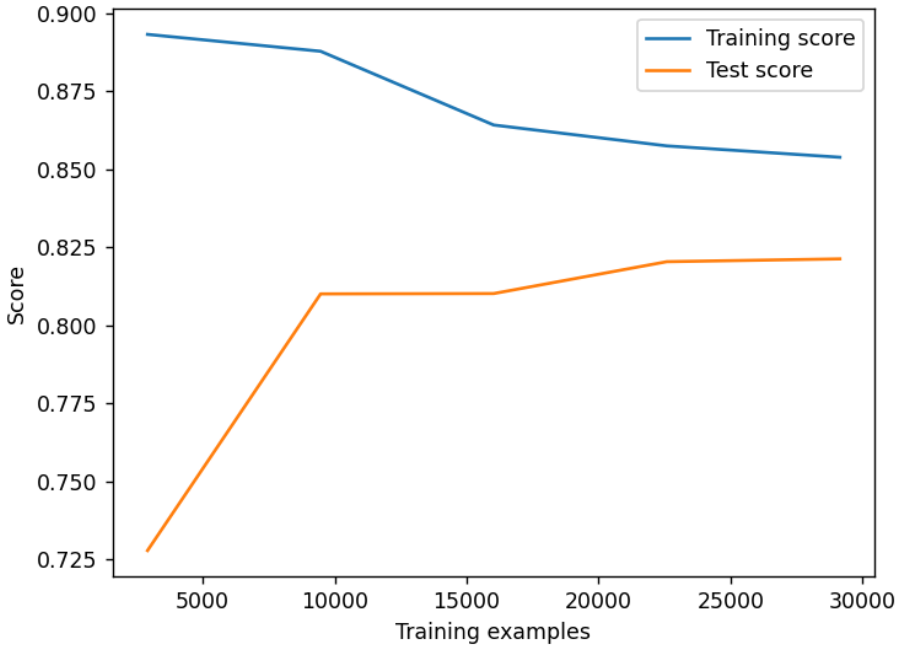


Figure (12)

The training score is very high when using little samples for training. It decreases slowly when the number of samples increases. The test score is low at the beginning and then increases fast at the beginning but becomes steady at some point as seen when adding more samples. The test score is lower than the training score that indicates high variance.

2.3 Performance

Below in **table (7)**, a comparison of the performance of the Logistic Regression, Support vector machine, and Artificial Neural network classifiers on the weather dataset using ROC/AUC and precision:

	LR	***SVM***	***ANN***
ROC/AUC	*0.85*	*0.82*	*0.86*
Accuracy	*0.84*	*0.85*	*0.85*
Precision	*0.69*	*0.77*	*0.78*
Recall	*0.48*	*0.42*	*0.43*
f-score	*1*	*0.54*	*0.56*

Table (7)

The ANN model has the highest ROC-AUC score among the three models (0.86).

The accuracy scores of all three models are close to each other, with LR and SVM having an accuracy score of 0.84 and 0.85 respectively, while ANN has an accuracy score of 0.85.

The precision scores of the models are also quite close to each other, with the highest score of 0.78 belonging to ANN.

When it comes to recall, SVM has the lowest score (0.42), while LR and ANN have scores of 0.48 and 0.43 respectively.

The F1-Score of ANN (0.56) is slightly higher than that of LR (0.54) and much higher than that of SVM (1).

ANN model has the best performance with the highest ROC/AUC score of 0.86.

Conclusion: Weather dataset Analysis

In conclusion, the analysis of the weather dataset revealed the presence of missing values and outliers that were effectively handled using KNN imputer and Capping and Flooring. It also showed that the ranges of some features may dominate the others in respect to their contribution to the classification task so feature scaling had to be performed.

Multivariate analysis was performed to determine the correlation between features and the target, leading to the removal of features with low positive correlation and highly correlated features to avoid redundancy.

After evaluating the performance of three different classification algorithms (LR, ANN and SVM), the ANN classifier was found to be the best performer with the highest ROC/AUC score and precision. For that reason the selected algorithm will be ANN to predict whether it will rain tomorrow or not.

These results can be valuable for Al-Bireh municipality in making informed decisions about weather predictions, such as allocating resources for potential rain-related events or planning outdoor activities based on the predicted weather. By using the best performing classifier, the municipality can have more confidence in its weather predictions and respond more effectively to potential weather-related challenges.