machine learning report (classification project weather)

By:

Name	ID
Moustafa Karam	42010428

Under the Supervision of:

Eng. Ahmed Nousir

Contents

1. Introduction

Begin by providing a brief overview of the problem at hand – the challenges and implications associated with weather. Mention the importance of predicting and managing weather, and how the given dataset can contribute to this.

2. The Problem understanding:

The data set is about weather this data contain 21 columns and 25000 rows

2.1 Data Description

Explain the significance of each column in the dataset. Highlight the role of factors like rain today, wind, and others in contributing to rain tomorrow.

2.2 Problem Statement

Define the problem statement clearly. For example, you could state that the goal is to predict based on various factors in the dataset.

3. Data Exploration

This data from Kaggle is about 25000 rows and 21 columns in his part we need to get more information about data and another statistics.

3.1 Descriptive Statistics

Provide summary statistics for each column. This could include mean, median, standard deviation, and other relevant metrics. Identify any trends or patterns.

Out[43]:

	minimum_temp	maximum_temp	rain_fall	Wind_gustspeed	Wind_speed9am	Wind_speed3pm	Humidity_9am	Humidity_3pm	Pressure_9am	Pressu
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.
mean	13.356868	23.956832	2.644620	36.836760	12.396640	16.451520	70.294920	53.113400	1018.390880	1015.
std	5.834302	6.105938	9.669995	12.486118	9.228075	9.029664	18.010043	20.820351	5.838693	5.
min	-3.300000	6.800000	0.000000	7.000000	0.000000	0.000000	3.000000	1.000000	980.500000	979.
25%	9.000000	19.400000	0.000000	30.000000	6.000000	9.000000	58.000000	37.000000	1015.000000	1012.
50%	14.100000	23.300000	0.000000	33.000000	11.000000	15.000000	71.000000	55.000000	1019.300000	1015.
75%	18.000000	27.700000	0.600000	43.000000	19.000000	22.000000	84.000000	67.000000	1021.400000	1018.
max	29.700000	47.300000	371.000000	135.000000	130.000000	83.000000	100.000000	100.000000	1039.900000	1036.

Figure 1 Statistics for data

3.2 Correlation Analysis

Explore the correlation between different factors and the target variable (Rain tomorrow). This will help understand which features are more influential.

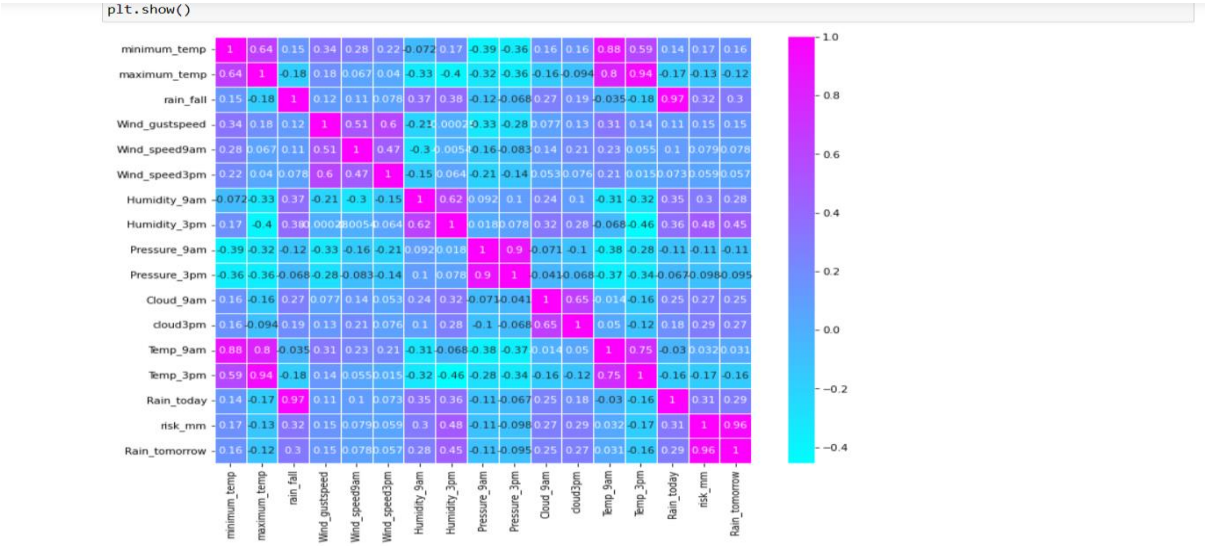
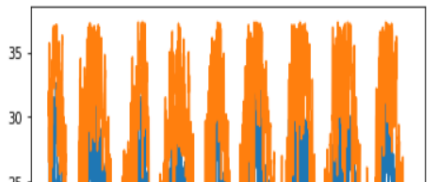


Figure 2 Correlation

Create visualizations such as histograms, scatter plots, or box plots to better understand the distribution of data and relationships between variables.

In [63]:
```
df.groupby('date')[['Temp_9am', 'Temp_3pm']].max().plot()
```

Out[63]: <AxesSubplot:xlabel='date'>



4. Data Preprocessing

This part for clean data from nulls and duplicates and check outliers to make data ready for deploy model.

4.1 Handling Missing Values

Address any missing values in the dataset through imputation or removal.

4.2 Encoding Categorical Variables

If there are categorical variables, encode them into numerical format for model compatibility.

4.3 Feature Scaling

Normalize or standardize numerical features if necessary to ensure fair treatment by the model.

5. Model Building

5.1 Splitting Data

Divide the dataset into training and testing sets.

5.2 Model Selection

Choose a suitable machine learning model for the task (e.g., Random Forest, Logistic Regression).

5.3 Training the Model

Train the chosen model on the training dataset.

5.4 Model Evaluation

Evaluate the model's performance on the testing dataset. Mention metrics such as accuracy, precision, recall, and F1 score.

	Model	Testing Accuracy
1	DecisionTree	1.00
5	GradientBoosting	1.00
0	RandomForest	1.00
6	NaiveBayes	0.98
2	LogisticRegression	0.95
4	KNN	0.82
3	SVM	0.78

Figure 4 Accuracy

6. Conclusion

Summarize the key findings and insights from the analysis. Discuss the limitations and potential areas for improvement. Offer recommendations for further research or model enhancements.