Logo

Description automatically generated

***Sultan Mahmud***

***Matricola 648496***

**Github Repository:**

**https://github.com/SultanMahmud89/Weather-Forecast-Predictor**

**Weather forecast predictor**

Chapter 1

Introduction

Weather History dataset belongs to various attributes considering for analysis of weather for a particular place. In this project we are using the dataset taken from Kaggle for making analysis with different data mining approaches. At first, we visualize the data set and try to find if there is any specific kind of noise as well as outliers. Then we preprocess the data and after that we have tried to fit our model with different classifier for getting a prediction result so that we can evaluate our model. A small glimpse of time series is also applied at the end of the project for getting or analyzing the attributes like temperature (c), summary, wind speed with apparent temperature. we have used matrix profile for getting the normalized values and after that applied discords and trying to find if there are any anomalies on that specific attributes. we have found anomalies in each attribute we were using during applying the matrix profile. In future we would like to build a model which could predict weather for a specific place based on previous data.

Dataset Description

Weather event is a spatiotemporal entity, where such an entity is associated with location and time. This dataset contains historical daily weather data for Leeds, England from Jan 1st, 2006, to Dec 31st, 2016. Following is the description of available weather event types in this dataset:

Formatted Date: Date year-month-day hours: minutes: seconds, indicates the time of detection in a period between Jan 1st, 2006, to Dec 31st, 2016.

Summary: Weather conditions, including the overall status or description of a specific time of the day (i.e., sunny, partially cloudy, cloudy, etc.)

Daily Summary: Weather conditions, including the overall status or description of that day (i.e., partly sunny throughout the day cloudy, partly cloudy throughout the day cloudy, etc.)

Temperature degrees Celsius, indicates the temperature of the room at the time of detection.

Precipe Type: Short for Precipitation Type meaning any liquid or frozen water that forms in the atmosphere and falls back to the Earth, for instance snow or rain.

Apparent temperature (C): The temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity, and wind speed. The measure is most applied to the perceived outdoor temperature.

Humidity: The amount of water vapor in the air.

Wind Speed: Air moving from high to low pressure, usually due to changes in temperature.

Wind Bearing: Wind direction or the compass heading from which the wind is blowing.

Visibility: Expressed in terms of the horizontal distance at which a person should be able to see and identify.

Pressure: It refers to the weight of the air.

This dataset was taken from Kaggle website. Below is the given link:

<https://www.kaggle.com/datasets/muthuj7/weather-dataset>

Graphical user interface

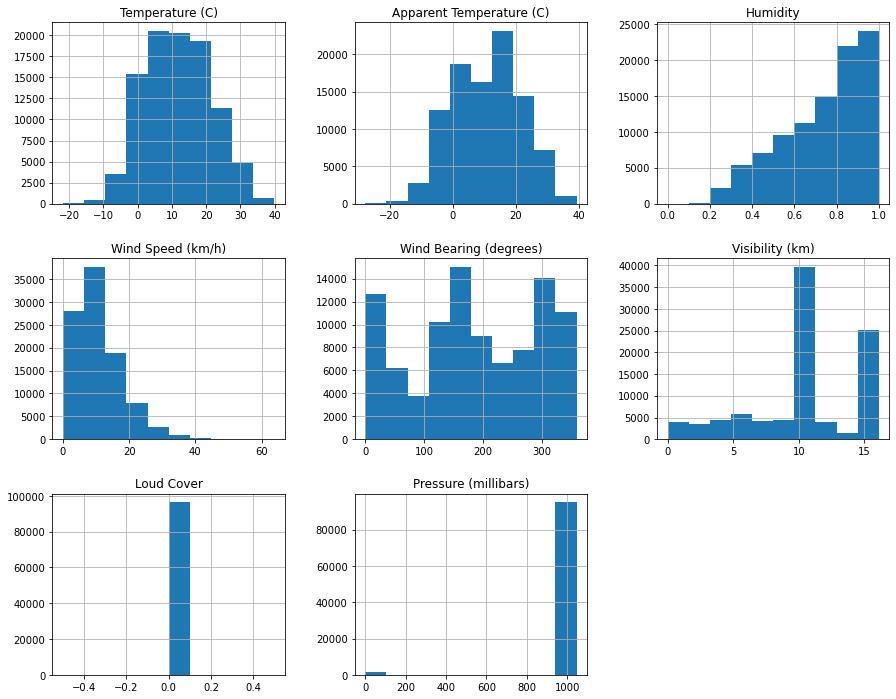
Description automatically generated with medium confidence

Chapter 2

Visualization

Graphical user interface, text, application

Description automatically generated



Chart, bar chart

Description automatically generated

A picture containing rectangle

Description automatically generated

Chart, scatter chart

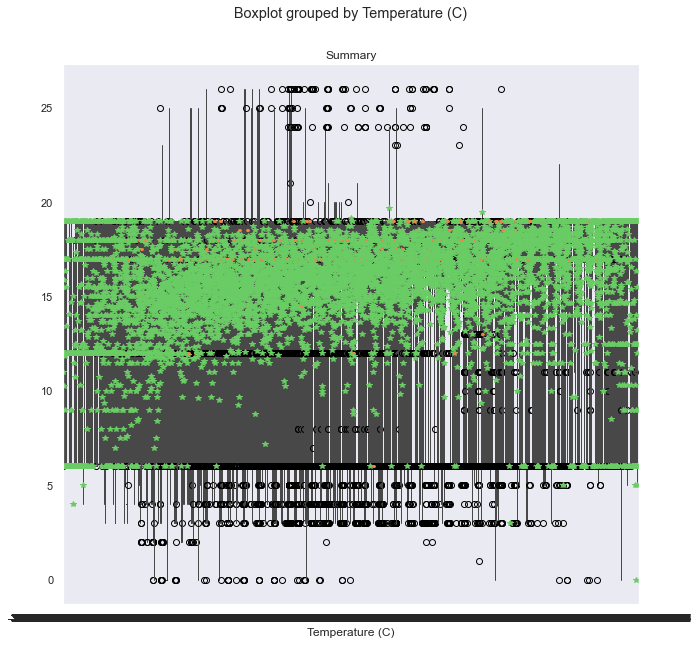
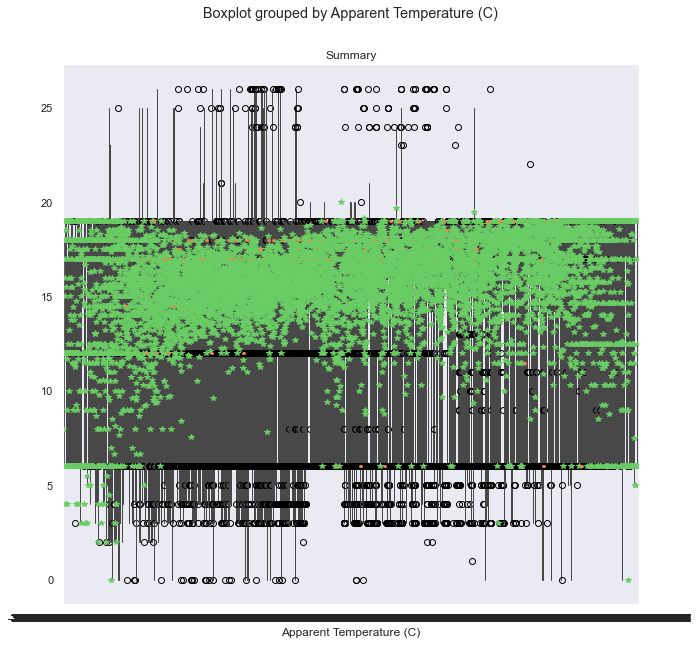
Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, bar chart

Description automatically generated



A picture containing graphical user interface

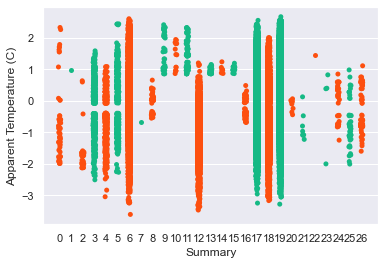
Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Graphical user interface, chart

Description automatically generated Graphical user interface, chart

Description automatically generated



Chart

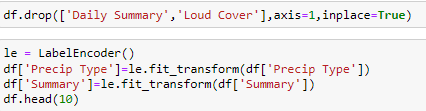
Description automatically generated

Chapter 3

Data Preprocessing

Text

Description automatically generated



Chart

Description automatically generated with medium confidence

Text

Description automatically generated

Table

Description automatically generated

A picture containing treemap chart

Description automatically generated

Chapter 4

Classification

|  |  |
| --- | --- |
| Classifier | Accuracy |
| Random Forest | 0.62798 |
| KNN | 0.53226 |
| Logistic Regression | 0.49798 |
| Decision Tree | 0.69636 |

Chapter 5

Time Series

Calendar

Description automatically generated with low confidence

In this exploratory phase, the matrix was first calculated profile, for each variable of the dataset considered as univariate time series, to discover any motifs or anomalies in the series. It was chosen to use the STOMP algorithm which is one of the most significantly fast for calculation. By fixing the size of the time window w to 1440 (number of observations for each full day) and a maximum number of motifs 5, it was possible to distinguish, as can be seen in Figure

Chart, bar chart

Description automatically generated

Matrix Profiling of Temperature C

A picture containing text, pencil, tool

Description automatically generated

Anomaly Detection of Temperature C

Chart

Description automatically generated

Matrix Profiling of Summary

Chart

Description automatically generated

Anomaly Detection of Summary

Graphical user interface, text, application

Description automatically generated