Outdoor Air Temperature Prediction
Phase Two

CSCE 3602 Fundamentals of Machine Learning

Amr Abdelbaky
Computer Science
The American University In Cairo
Cairo, Egypt
amrkhaled122@aucegypt.edu

Ali Eissa
Computer Science
The American University In Cairo
Cairo, Egypt
alieissa@aucegypt.

Dataset of Choice

The chosen Dataset is " Climate Weather Surface of Brazil - Hourly ", only using the North Region of Brazil file , it contains **8.39 Million** samples that were collected starting from May ninth 2000 with a one hour window between each sample till April 30th 2021. The dataset contains about **26** meteorological features that are as follows :

date, hour, Amount of precipitation in millimeters (last hour), Atmospheric pressure at station level (mb), Maximum air pressure for the last hour in hPa to tenths, Minimum air pressure for the last hour in hPa to tenths, Solar radiation KJ/m2, Air temperature (instant) in celsius degrees, 'Dew point temperature (instant) in celsius degrees,Maximum temperature for the last hour in celsius degrees, Minimum temperature for the last hour in celsius degrees, Maximum dew point temperature for the last hour in celsius degrees,Minimum dew point temperature for the last hour in celsius degrees, Maximum relative humidity temperature for the last hour in %, Minimum relative humidity temperature for the last hour in %, 'Relative humidity in % (instant), Wind direction in radius degrees (0-360), 'Wind gust in meters per second, 'Wind speed in meters per second, region, state, station, station_code, latitude, longitude and height. The size of this dataset is about **1.4 GB** and it is available at kaggle[1].

The Rationale Behind The Dataset

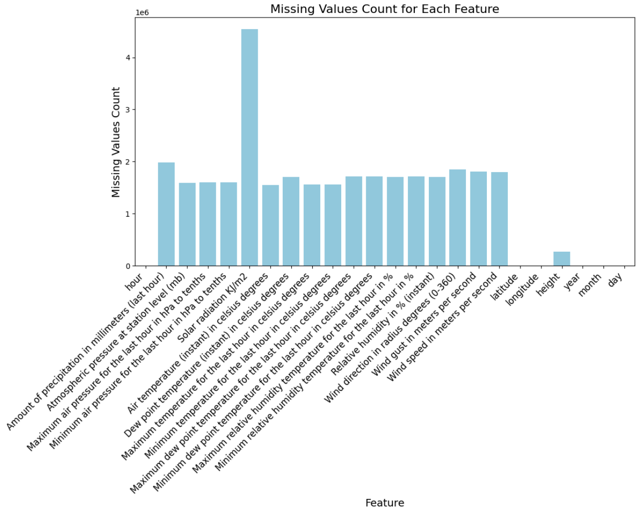
The Dataset was built purely for the purpose of weather forecasting, it contains about 8.39 Million samples , which is a very high amount of a sample size for training , validation and testing . Furthermore, it comes with 26 meteorological features that are correlated well with the label we want to predict (Air temperature (instant) in celsius degrees) . Also having a one hour window between each sample allows for detailed analysis of diurnal and seasonal patterns in air temperature.

A Full Analysis For Features

a. Missing values

The first step that was taken was to determine the amount of missing values and fill them in a meaningful way before proceeding with getting different distributions for different features or doing any data preprocessing .

The following graph was obtained with regard to the missing values from the data set :

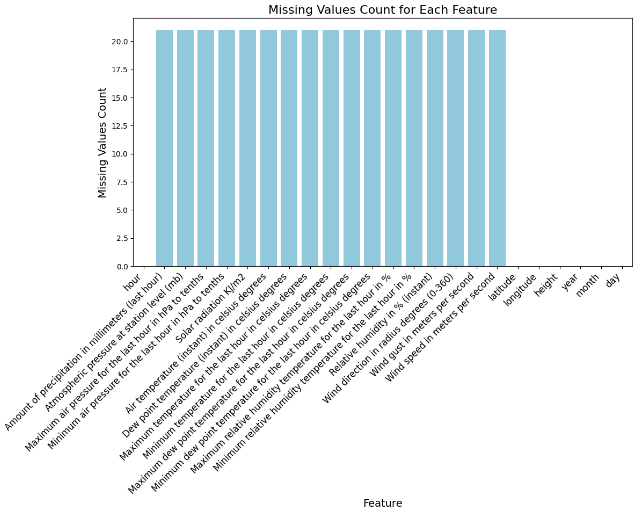


Accordingly , There were about two million missing values for our dataset of choice for 16 features , and the Solar Radiation feature had about four million samples missing , so we had to perform linear interpolation to fill in the missing values . Additionally , we decided to remove the index column (as it is unique for every input and not relevant at all to the value of the temperature).

We also decided to drop region, state, station and station code features because they weren't not highly correlated at all with the Temperature.

The date column was split into day, month and year columns respectively for easier processing.

After using Linear Interpolation to fill in the missing values as we are dealing with time series data, This graph was obtained.



We had now only 20 samples out of 8.39 million whose linear interpolation did not fill in , so we decided to drop them as they are insignificant relative to the size of our dataset and now we have no missing values at all.

We now have 24 Features in our dataset , one of them is going to be our label for the purpose of this project.

b. Unique Values for Nominal Features:

Format : Feature Name, Range[MinValue-MaxValue]

hour, Range: [00:00 - 23:00]

Amount of precipitation in millimeters (last hour), Range: [0.0 - 97.2]

Atmospheric pressure at station level (mb), Range: [852.1 - 1050.0]

Maximum air pressure for the last hour in hPa to tenths, Range: [832.0 - 1049.8]

Minimum air pressure for the last hour in hPa to tenths, Range: [830.1 - 1050.0]

Solar radiation KJ/m2, Range: [0.0 - 45305.0]

Air temperature (instant) in celsius degrees, Range: [-9.0 - 42.2]

Dew point temperature (instant) in celsius degrees, Range: [-10.0 - 43.5]

Maximum temperature for the last hour in celsius degrees, Range: [0.0 - 45.0]

Minimum temperature for the last hour in celsius degrees, Range: [-5.6 - 45.0]

Maximum dew point temperature for the last hour in celsius degrees, Range: [-10.0 - 44.4]

Minimum dew point temperature for the last hour in celsius degrees, Range: [-10.0 - 39.8]

Maximum relative humidity temperature for the last hour in %, Range: [7.0 - 100.0]

Minimum relative humidity temperature for the last hour in %, Range: [3.0 - 100.0]

Relative humidity in % (instant), Range: [7.0 - 100.0]

Wind direction in radius degrees (0-360), Range: [0.0 - 360.0]

Wind gust in meters per second, Range: [0.0 - 99.7]

Wind speed in meters per second, Range: [0.0 - 19.9]

latitude, Range: [-12.75055555 - 4.47749999]

longitude, Range: [-72.78666666 - -45.91999999]

height, Range: [9.92 - 798.0]

year, Range: [2000 - 2021]

month, Range: [1 - 12]

day, Range: [1 - 31]

c. Correlation with Label :

hour: 0.45390111

Amount of precipitation in millimeters (last hour): -0.08540796

Atmospheric pressure at station level (mb): 0.14338106

Maximum air pressure for the last hour in hPa to tenths: 0.15316780

Minimum air pressure for the last hour in hPa to tenths: 0.14930714

Solar radiation KJ/m2: 0.45139430

Dew point temperature (instant) in celsius degrees: -0.02131768

Maximum temperature for the last hour in celsius degrees: 0.95887973

Minimum temperature for the last hour in celsius degrees: 0.95842098

Maximum dew point temperature for the last hour in celsius degrees: 0.04912295

Minimum dew point temperature for the last hour in celsius degrees: -0.07612423

Maximum relative humidity temperature for the last hour in %: -0.71418078

Minimum relative humidity temperature for the last hour in %: -0.74521555

Relative humidity in % (instant): -0.76198101

Wind direction in radius degrees (0-360): -0.12089730

Wind gust in meters per second: 0.40028540

Wind speed in meters per second: 0.34701539

latitude: 0.07890954

longitude: 0.02198280

height: -0.14841944

year: 0.00033389

month: 0.11082083

day: -0.00251961

Comment : We noticed that the two features Maximum temperature for the last hour in celsius degrees and Minimum temperature for the last hour in celsius degrees: have a $\simeq\ 0.95$ Correlation with the label , so we decided to drop these two Features.

d. Semantic Importance/Relevance :

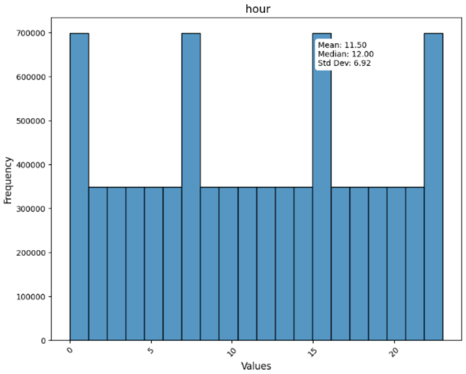
The features encompass diverse meteorological parameters, including temporal indicators such as hour, year, month, and day, as well as spatial characteristics like latitude, longitude, and height. Additionally, meteorological measurements like atmospheric pressure, air pressure variations, solar radiation, and dew point temperature are considered as general factors/variables of weather and temperature prediction.

We found varying degrees of correlation between these features and the target label. Notably, the time-related features, such as hour, month, and year, demonstrated discernible positive correlations, indicating their potential significance in predicting the target variable. Conversely, certain meteorological factors, such as maximum relative humidity, minimum relative humidity, and instant relative humidity, exhibited negative correlations, suggesting their importance in understanding the inverse relationship with the target label.

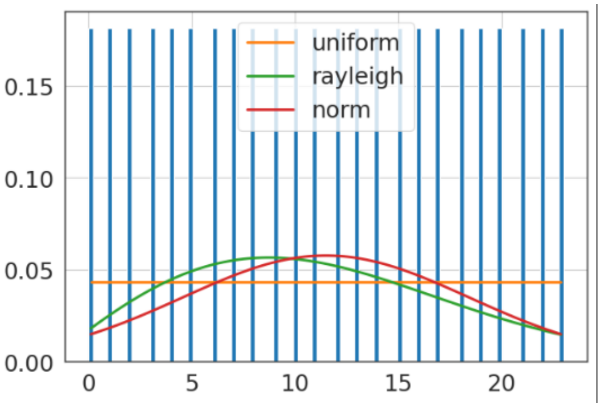
e. Statistical Distributions :

Using the Fitter Library from Python, We fitted the different distributions that we could to the columns' data. Most of which we found fit most closely with the normal distribution. We are gonna break down each feature one by one in this section.

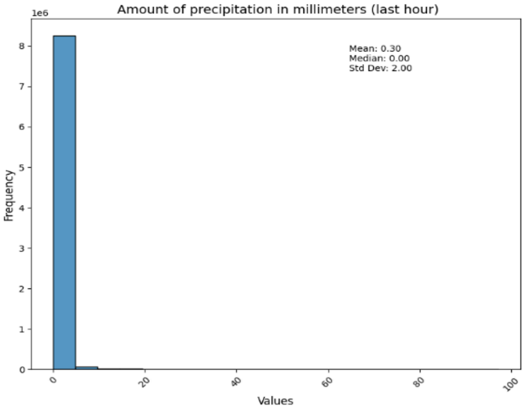
hour Feature :



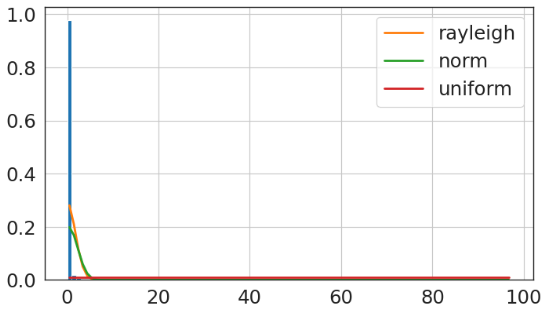
Using the Fitter library to detect the best distributions that fit the data were as shown next :



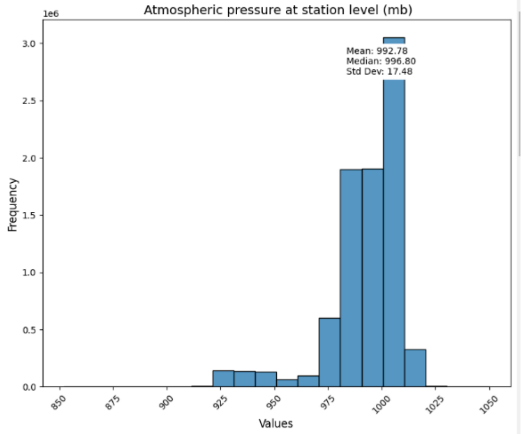
Amount of Precipitation in millimeters (last hour) Feature :



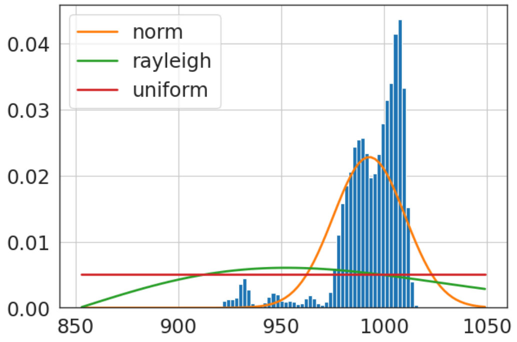
Using the Fitter library to detect the best distributions that fit the data were as shown next:



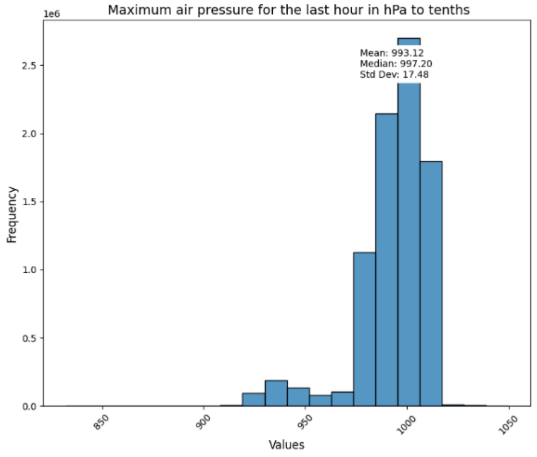
Atmospheric pressure at station level (mb) Feature :



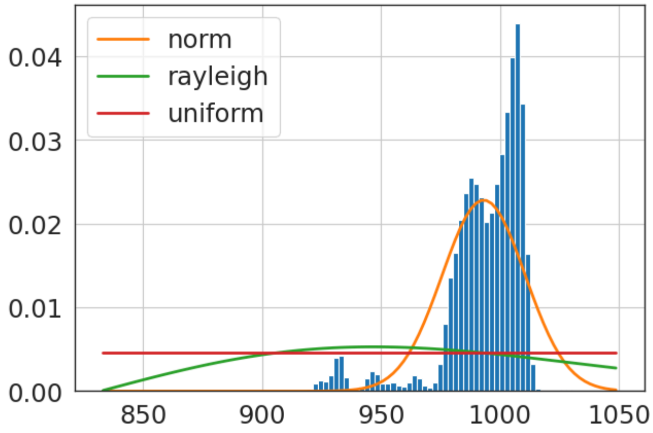
Using the Fitter library to detect the best distributions that fit the data were as shown next:



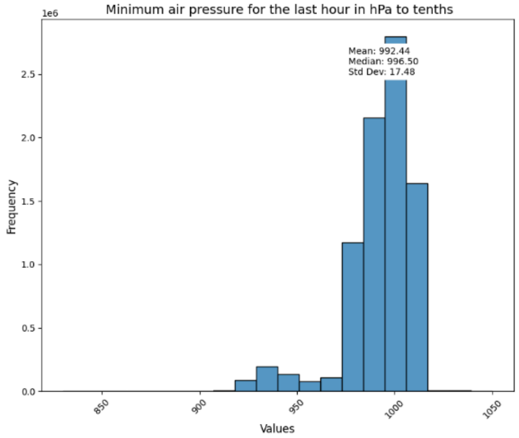
Maximum Air Pressure for the last hour in hPa to tenths Feature :



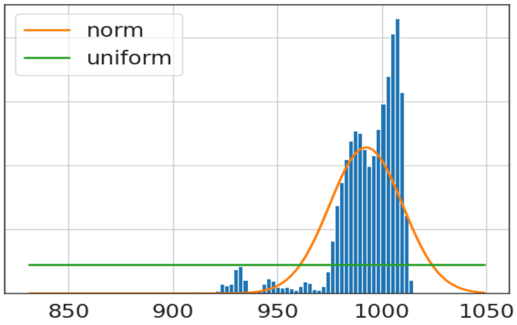
Using the Fitter library to detect the best distributions that fit the data were as shown next :



Minimum Air Pressure for the last hour in hPa to tenths Feature :

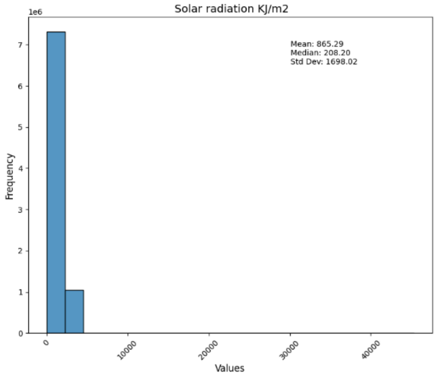


Using the Fitter library to detect the best distributions that fit the data were as shown next:

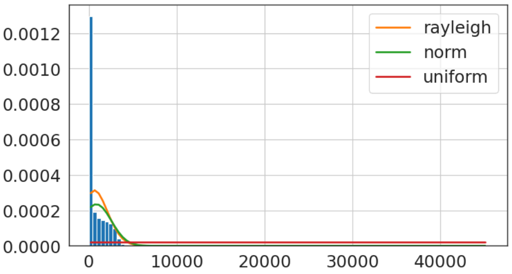


So Minimum Air Pressure for the last hour in hPa to tenths feature follows normal and uniform distributions .

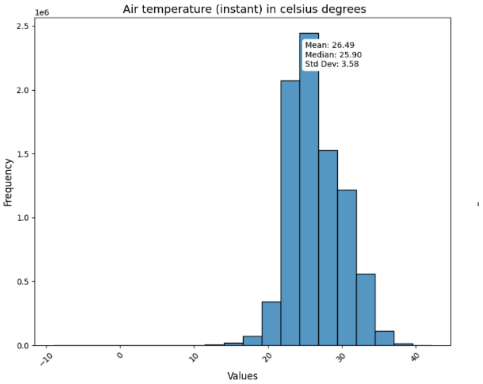
Solar Radiation Kj/m2 Feature:



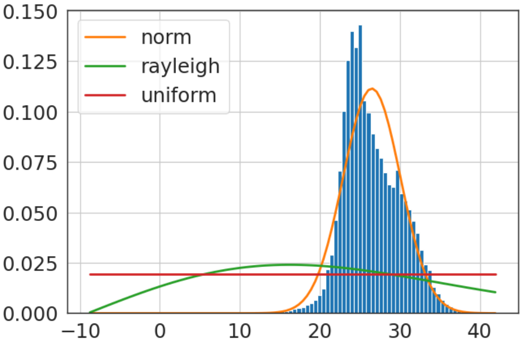
Using the Fitter library to detect the best distributions that fit the data were as shown next:



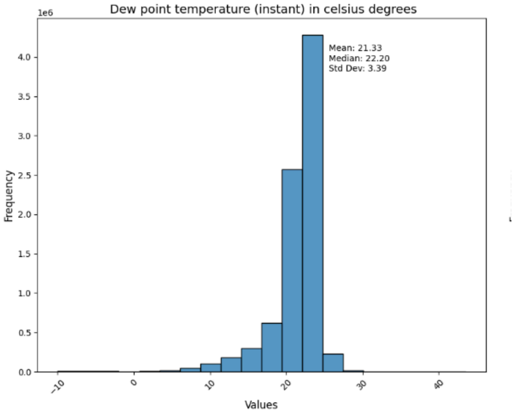
Air Temperature (instant) in celsius degrees Feature :



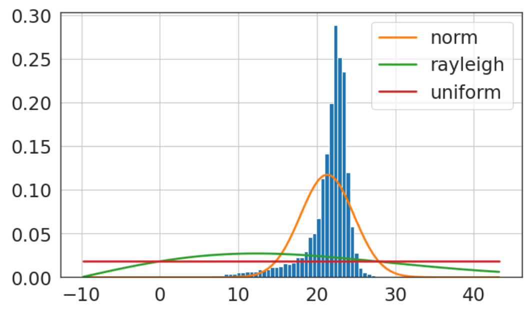
Using the Fitter library to detect the best distributions that fit the data were as shown next:



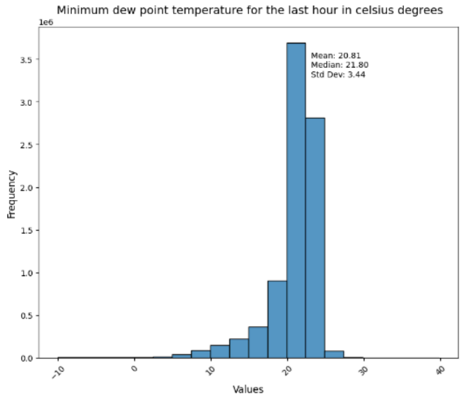
Dew Point temperature (instant) in celsius degrees Feature :



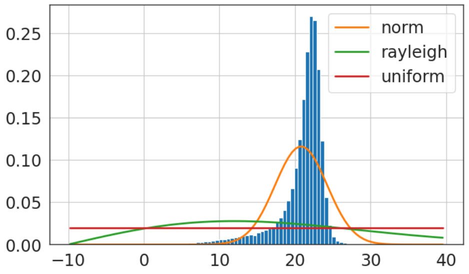
Using the Fitter library to detect the best distributions that fit the data were as shown next:



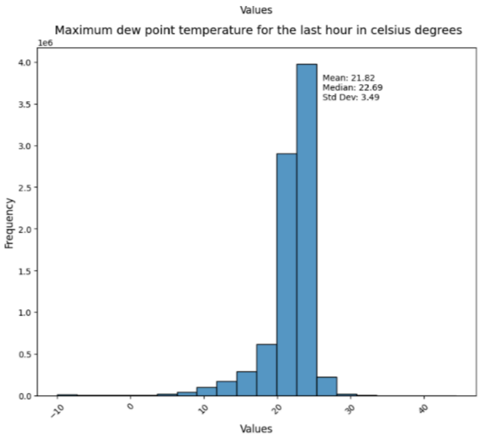
Minimum dew point temperature for the last hour in celsius degrees Feature :



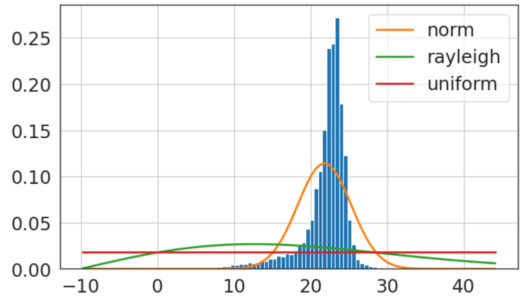
Using the Fitter library to detect the best distributions that fit the data were as shown next:



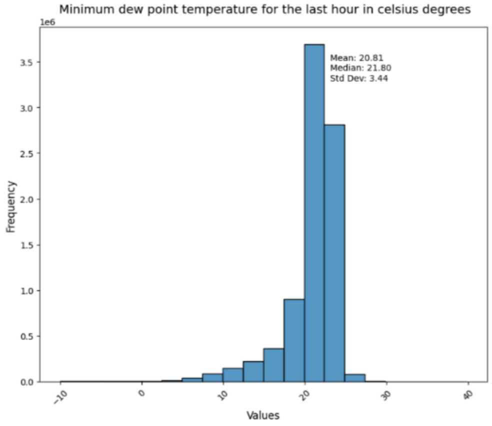
Maximum dew point temperature for the last hour in celsius degrees Feature :



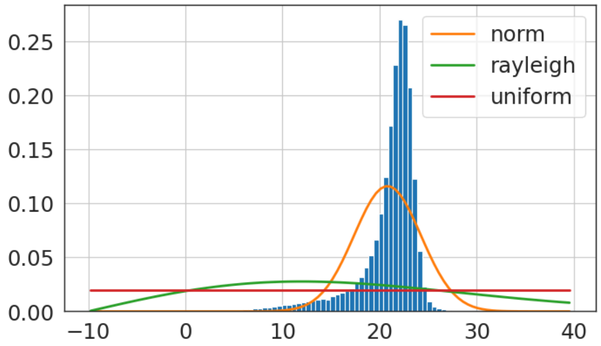
Using the Fitter library to detect the best distributions that fit the data were as shown next:



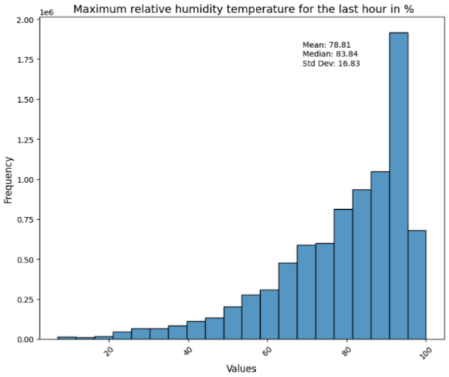
Minimum dew point temperature for the last hour in celsius degree Feature :



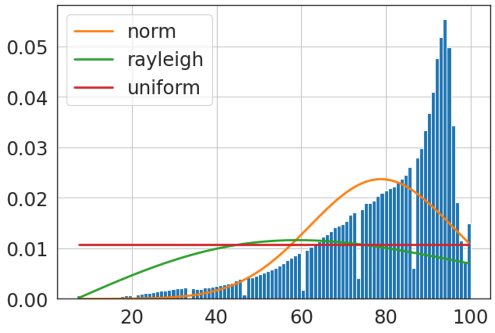
Using the Fitter library to detect the best distributions that fit the data were as shown next:



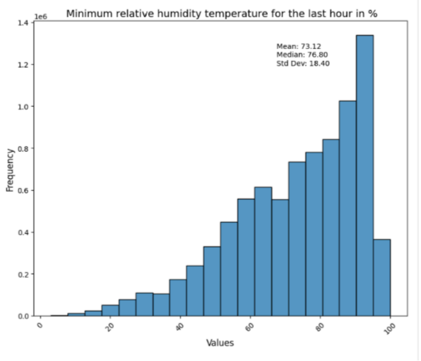
Maximum relative humidity temperature for the last hour in percentage Feature :



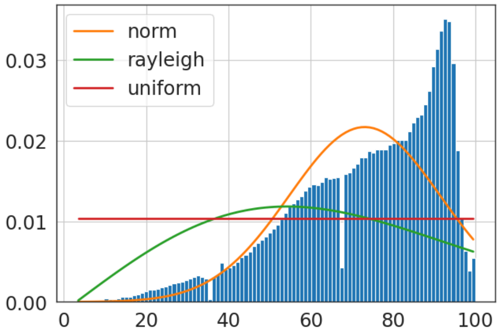
Using the Fitter library to detect the best distributions that fit the data were as shown next:



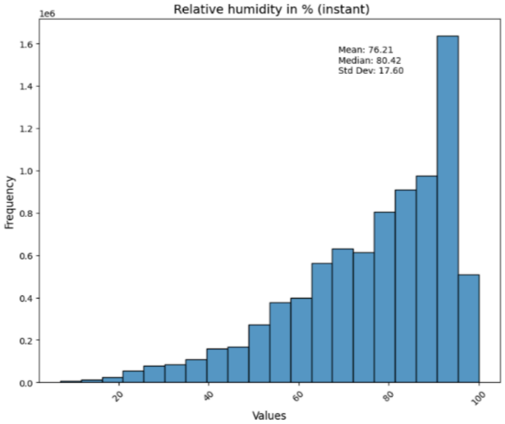
Minimum relative humidity temperature for the last hour in percentage Feature :



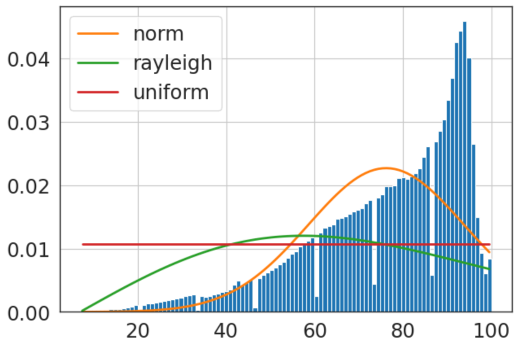
Using the Fitter library to detect the best distributions that fit the data were as shown next:



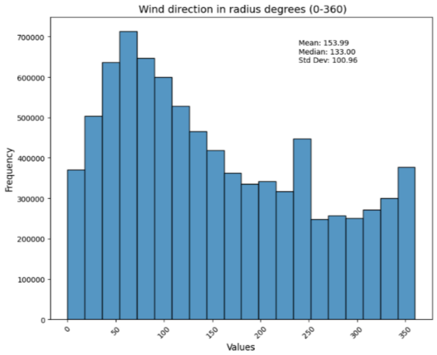
Relative humidity in percentage (instant) Feature :



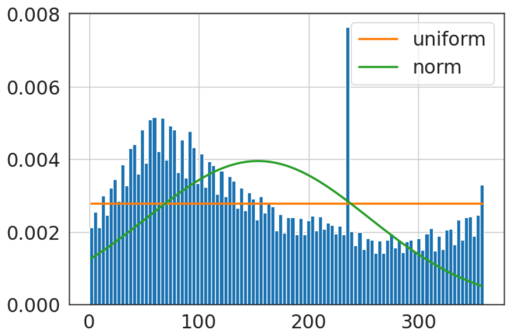
Using the Fitter library to detect the best distributions that fit the data were as shown next:



Wind direction in radius degrees (0-360) Feature :

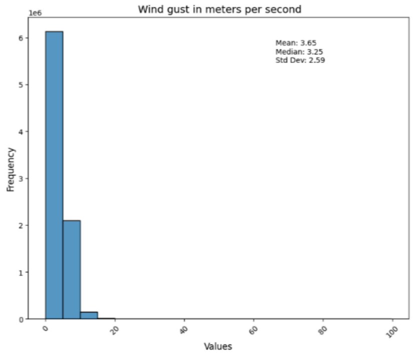


Using the Fitter library to detect the best distributions that fit the data were as shown next:

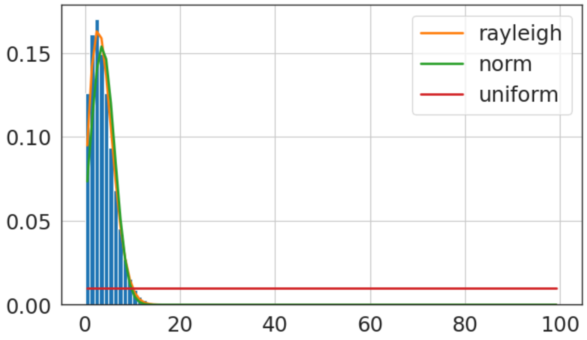


so the Wind direction in radius degrees (0-360) feature follows uniform and normal distributions only .

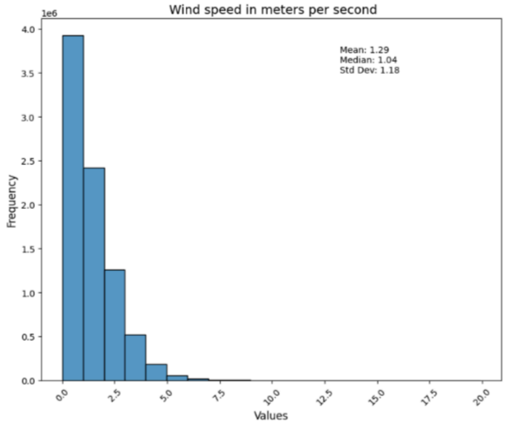
Wind gust in meters per second Feature :



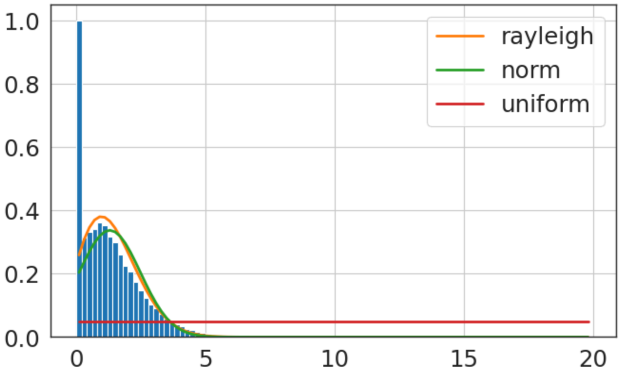
Using the Fitter library to detect the best distributions that fit the data were as shown next:



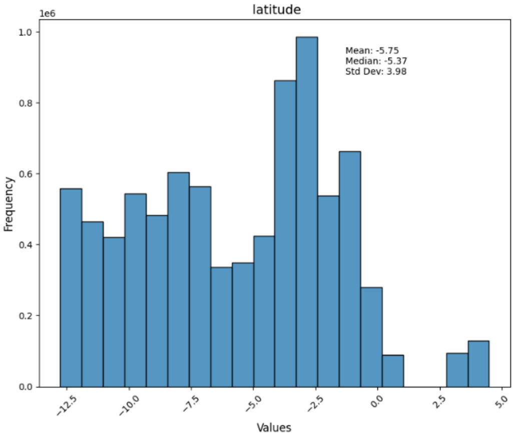
Wind speed in meters per second Feature :



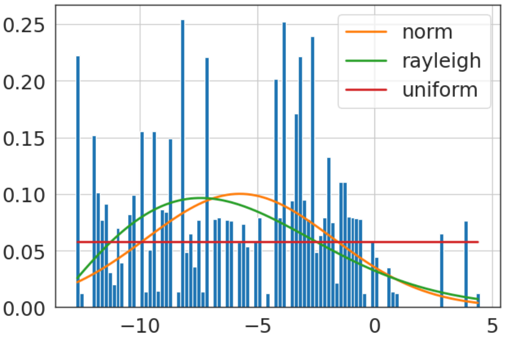
Using the Fitter library to detect the best distributions that fit the data were as shown next:



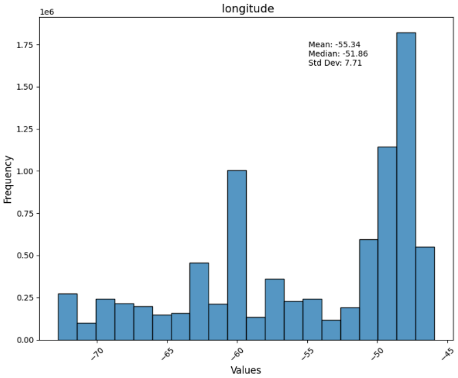
Latitude Feature :



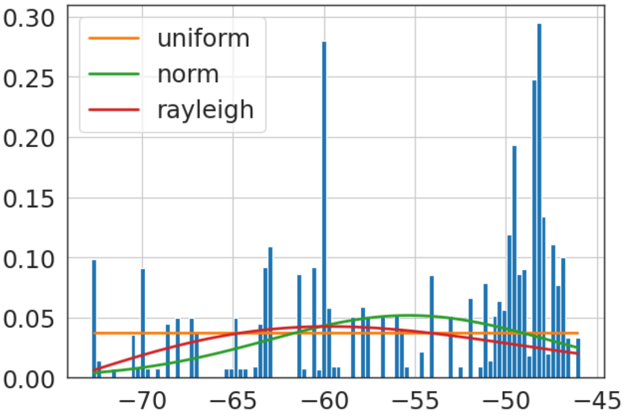
Using the Fitter library to detect the best distributions that fit the data were as shown next :



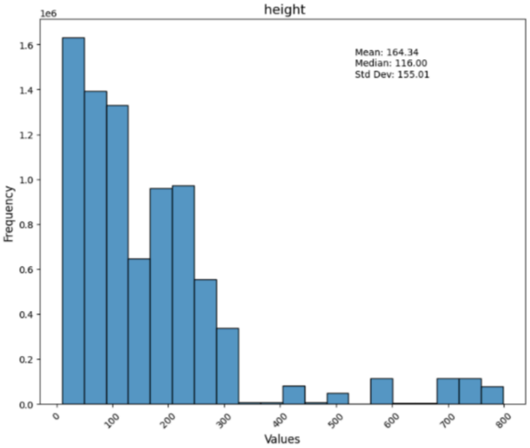
Longitude Feature :



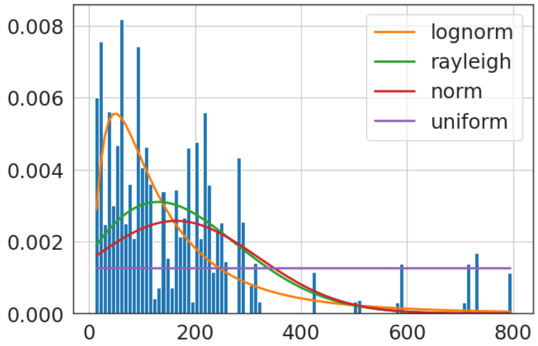
Using the Fitter library to detect the best distributions that fit the data were as shown next :



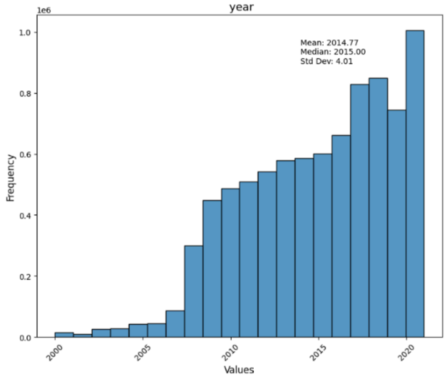
Height feature :



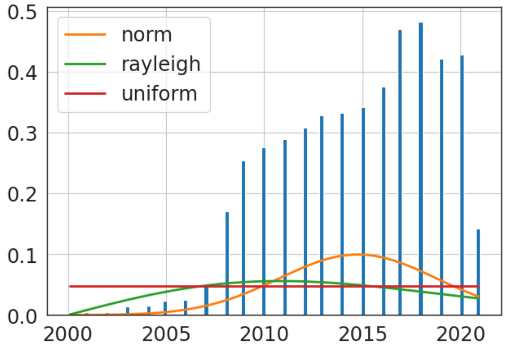
Using the Fitter library to detect the best distributions that fit the data were as shown next :



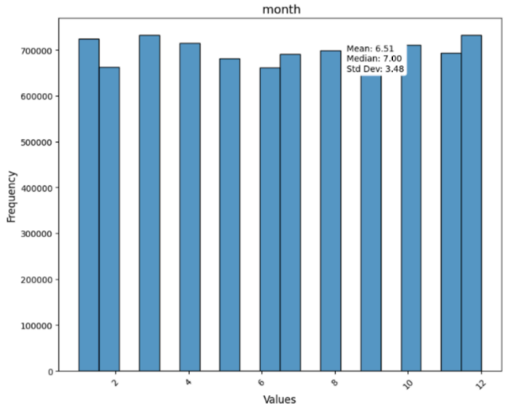
Year Feature :



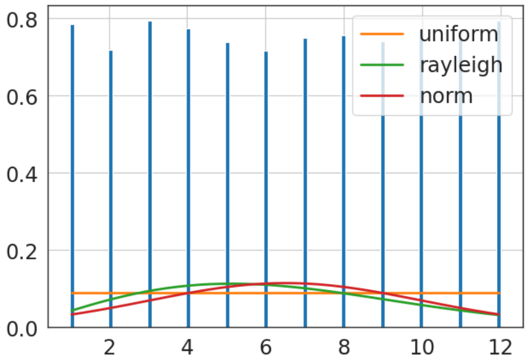
Using the Fitter library to detect the best distributions that fit the data were as shown below :



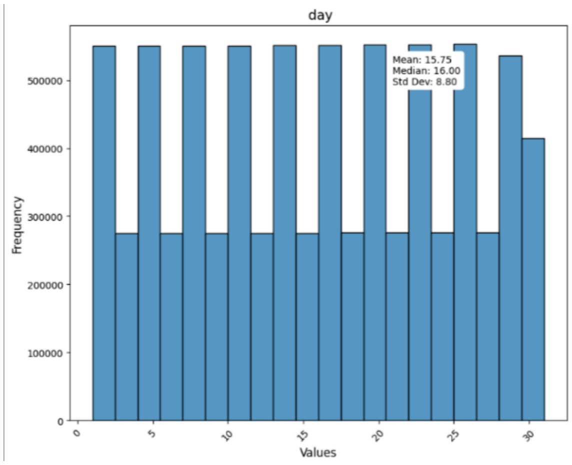
Month Feature :



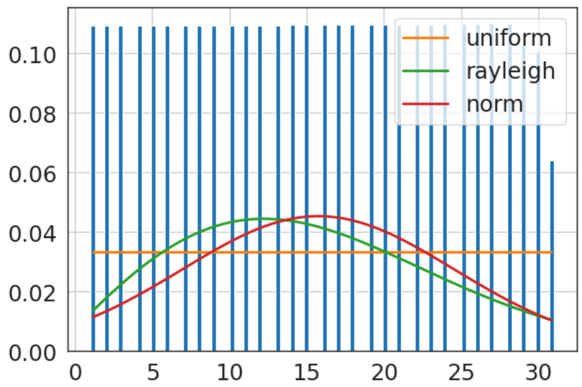
Using the Fitter library to detect the best distributions that fit the data were as shown below :



Day Feature :



Using the Fitter library to detect the best distributions that fit the data were as shown below :



Cleaning :

We dropped a few features that were either highly correlated with our label or irrelevant to our use case.

Many the column's sensor data contained values like -9999 which is by convention considered as a placeholder value for missing values or NaNs in datasets, so we substituted them with NaN, then we followed by linear interpolation on these values to fill the missing gaps, as linear interpolations is typically used with time-series data.

PreProcessing :

The Following Features were scaled using MinMax scaling due to their numeric nature :

- Hour
- Day
- Month
- Year

Due to the nature of the distributions that were found in this dataset , we decided to do Standard Z-scaling to the following features :

- Amount of precipitation in millimeters (last hour)
- Atmospheric pressure at station level (mb)
- Maximum air pressure for the last hour in hPa to tenths
- Minimum air pressure for the last hour in hPa to tenths
- Solar radiation KJ/m2
- Dew point temperature (instant) in celsius degrees
- Maximum dew point temperature for the last hour in celsius degrees
- Minimum dew point temperature for the last hour in celsius degrees
- Maximum relative humidity temperature for the last hour in %
- Minimum relative humidity temperature for the last hour in %
- Relative humidity in % (instant)
- Wind direction in radius degrees (0-360)
- Wind gust in meters per second
- Wind speed in meters per second
- Latitude
- Longitude
- Height

Removed Features :

Due to very low relevance to temperature , the following features were removed :

- Region
- State
- Station
- Station code

Due to a very high correlation(.95) between the label and these two features , they were removed to reduce any bias in the data for a better training :

- Maximum temperature for the last hour in celsius degrees
- Minimum temperature for the last hour in celsius degrees

Chosen Features :

The following 21 Features were left after removing the above features :

- Hour
- Day
- Month
- Year
- Amount of precipitation in millimeters (last hour)
- Atmospheric pressure at station level (mb)
- Maximum air pressure for the last hour in hPa to tenths
- Minimum air pressure for the last hour in hPa to tenths
- Solar radiation KJ/m2
- Dew point temperature (instant) in celsius degrees
- Maximum dew point temperature for the last hour in celsius degrees
- Minimum dew point temperature for the last hour in celsius degrees
- Maximum relative humidity temperature for the last hour in %
- Minimum relative humidity temperature for the last hour in %
- Relative humidity in % (instant)
- Wind direction in radius degrees (0-360)
- Wind gust in meters per second
- Wind speed in meters per second
- Latitude

- Longitude
- Height

And our label is :

- Air temperature (instant) in celsius degrees

Dataset Size and Metrics Statistics post Cleaning :

Number of rows

- 8,392,299 Sample

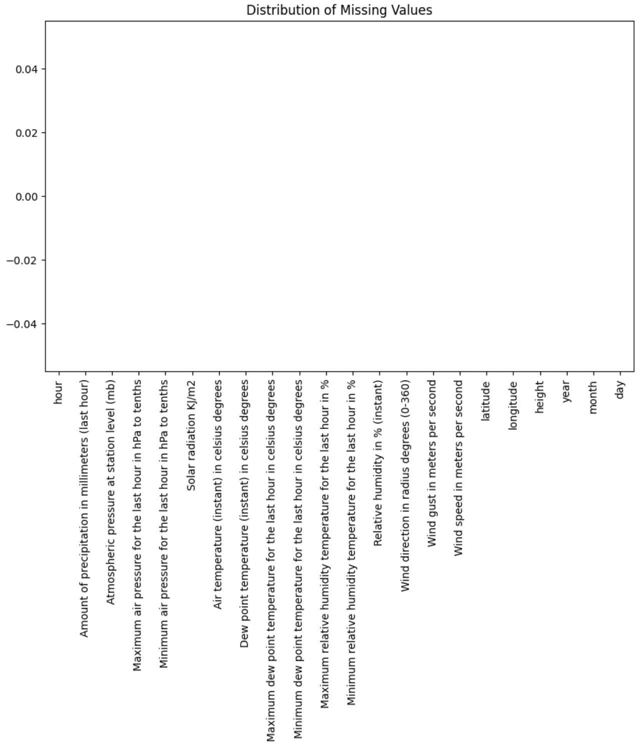
Number of columns

- 22 Column

Size of the dataset after cleaning:

- 3190.85 MegaBytes or 3.116067GigaBytes

There are no current missing values as per this final graph obtained after cleaning the dataset :



A dump of the cleaned dataset :

Due to the size of our dataset after cleaning (3.14 GigaBytes) we are going to share the link to the Final .Csv file in our shared google drive folder instead of uploading it , access the dataset by pressing here

References

[1] INMET (National Meteorological Institute - Brazil). (2000, January). Climate Weather Surface of Brazil - Hourly, Version 9. Retrieved February 18, 2024 from https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region/data?select=north.csv