# Weather Forecast for Indian Climate Using Bayesian Approach

## 1. Introduction

Nowadays, people tend to pay more attention on environment development and wildlife sustainment for our generations. The additional heat in the earth's atmosphere that has produced the rise in global temperature is referred to as global warming. Climate change is caused by, and will continue to be caused by, global warming. Climate change can result in increasing sea levels, community devastation, and harsh weather events. Here are ten global warming sources that are contributing to the climate disaster, for example, overfishing, industrialization, farming, consumerism, transport and vehicles, oil drilling, power plants, waste, deforestation and gas. In this project, we applied Bayesian approach to predict whether the weather in Delhi, India with multiple features is expected to be extreme hot or not. The dataset was obtained from the Kaggle website which had collected from Weather Undergroud API.

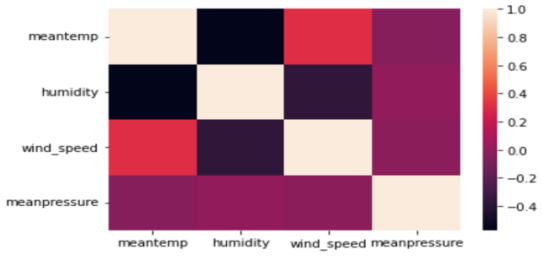
## 2. Exploratory Data Analysis

Exploratory data analysis (EDA) is the process of studying and visualizing data in order to gain a better knowledge of it and gain insight. In this project, we used some tools for EDA including importing data, descriptive statistics, cleaning data, preprocessing data and visualizing data. In our project, this dataset contains information from the 1st of January 2013 to the 24th of April 2017 in the city of Delhi, India. The five columns in this case are date, mean temperature, humidity, wind speed, and mean pressure. Mean temperature averaged out from multiple 3 hours intervals in a day. Humidity value was measured per day in grams of water vapor per cubic meter volume of air. Wind speed was measured in kmph and mean pressure was measured in atm.

Except date column, all other columns are numeric values. For this dataset, we used descriptive statistics to get the minimum, median, mean, standard deviation, and maximum values for each numeric variable, which is shown below. The correlation matrix allows us to determine whether or not there are strong relationships between the variables. Humidity is negatively connected with mean temperature, with a value of -0.57, according to the generated correlation matrix for our dataset. The correlation plot depicts the same conclusion as the correlation matrix.

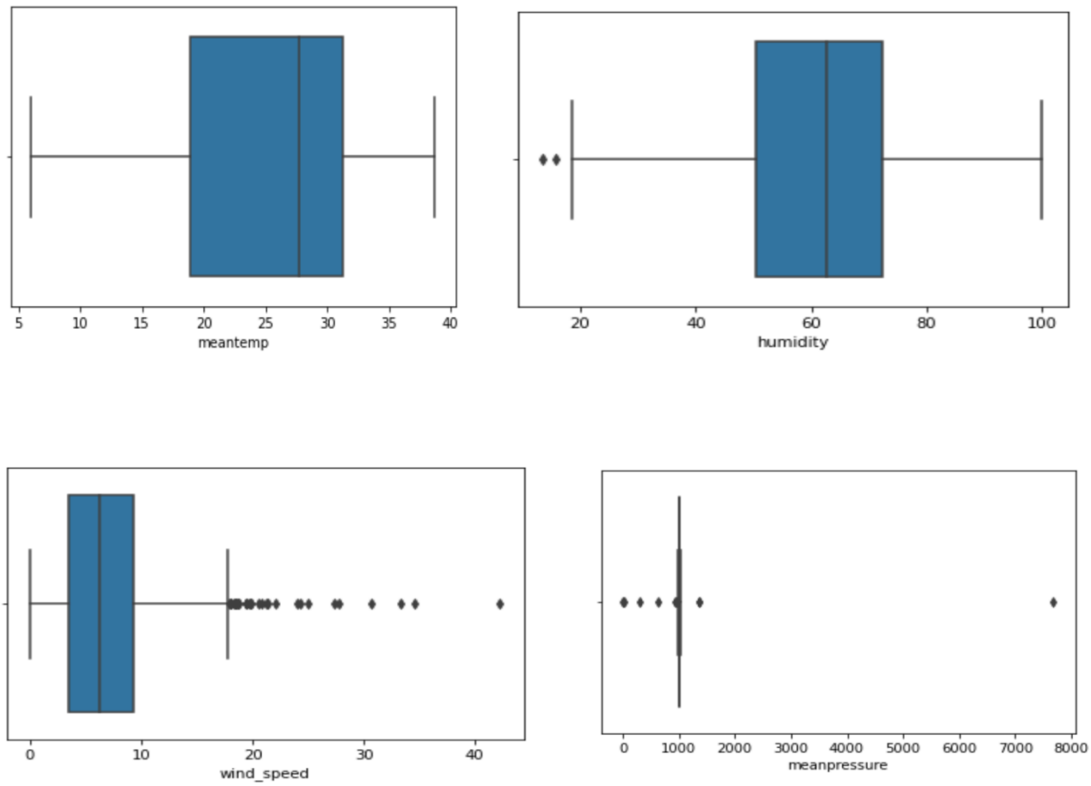
| | meantemp | humidity | wind_speed | meanpressure |
|---|---|---|---|---|
| **count** | 1462.000000 | 1462.000000 | 1462.000000 | 1462.000000 |
| **mean** | 25.495521 | 60.771702 | 6.802209 | 1011.104548 |
| **std** | 7.348103 | 16.769652 | 4.561602 | 180.231668 |
| **min** | 6.000000 | 13.428571 | 0.000000 | -3.041667 |
| **25%** | 18.857143 | 50.375000 | 3.475000 | 1001.580357 |
| **50%** | 27.714286 | 62.625000 | 6.221667 | 1008.563492 |
| **75%** | 31.305804 | 72.218750 | 9.238235 | 1014.944901 |
| **max** | 38.714286 | 100.000000 | 42.220000 | 7679.333333 |

| | meantemp | humidity | wind_speed | meanpressure |
|---|---|---|---|---|
| **meantemp** | 1.000000 | -0.571951 | 0.306468 | -0.038818 |
| **humidity** | -0.571951 | 1.000000 | -0.373972 | 0.001734 |
| **wind_speed** | 0.306468 | -0.373972 | 1.000000 | -0.020670 |
| **meanpressure** | -0.038818 | 0.001734 | -0.020670 | 1.000000 |

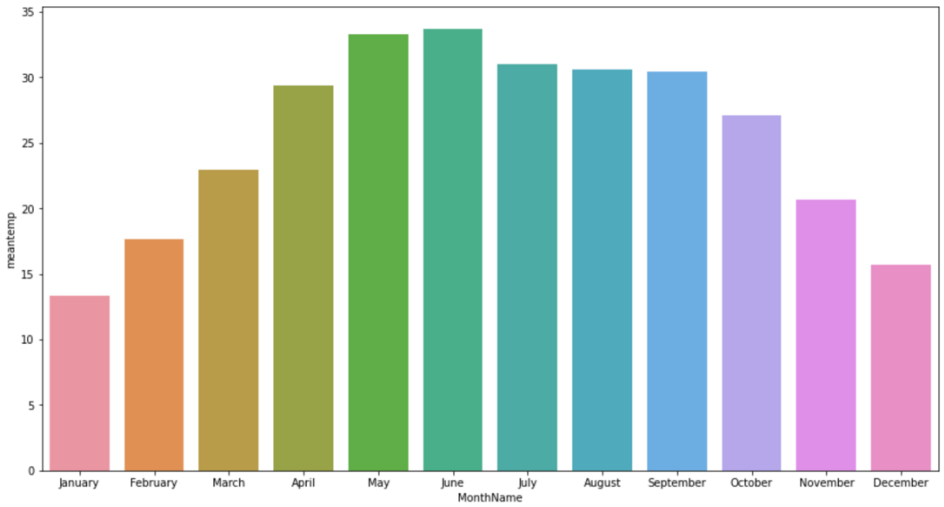


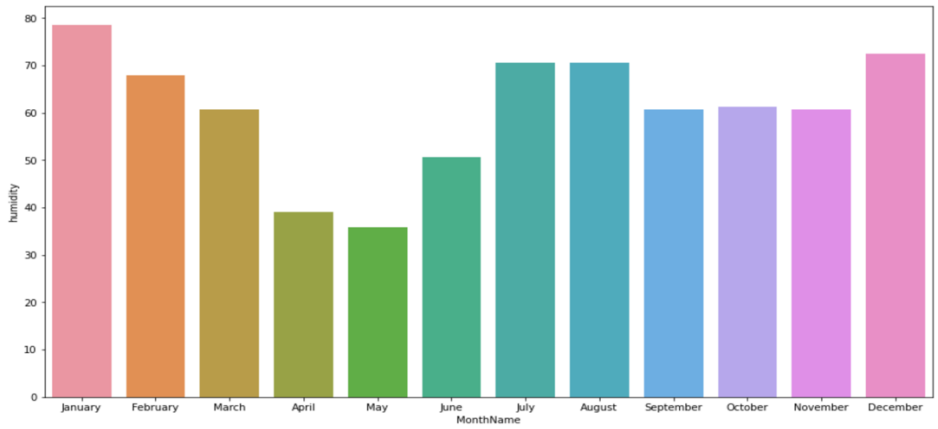

In the data cleaning section, we mostly check for missing, null, NA, NaN, or empty entries in our dataset. Missing values are often handled in two ways: removal and imputation. There are no missing or null values in our dataset. The duplicate rows add nothing to the model's or algorithm's learning process, but they do add storage and processing cost. Our dataset has no duplicate rows. In some circumstances, models, such as linear regression, may be particularly sensitive to outliers, and outliers may impair model performance. In our project, we utilized

boxplots to see if there were any outliers in the columns, and we discovered some in the humidity, wind speed and mean pressure and no outliers in mean temperature column.

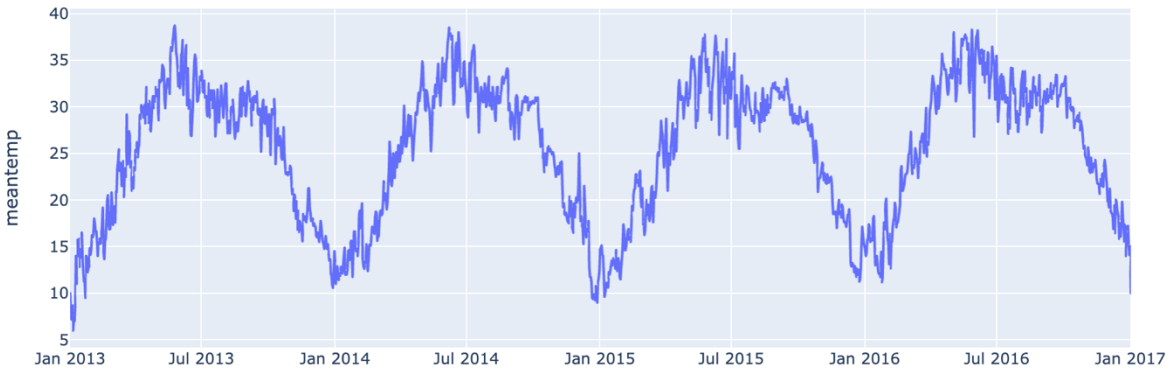


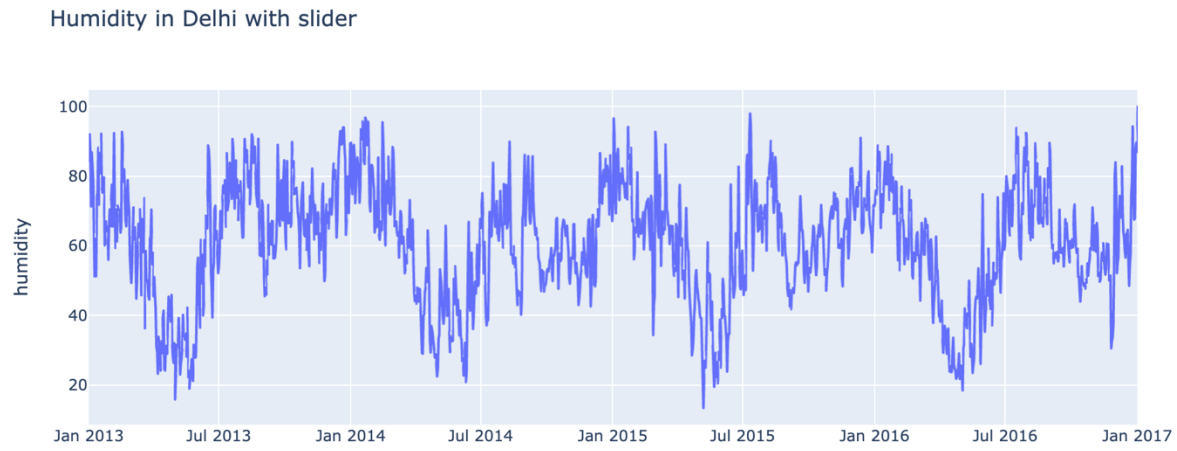

Data visualization is extremely useful for observing data distributions and trends. We extracted the month and day from the Date column and created three bar plots across the month, day, and year columns, revealing that the mean temperature in June was the highest compared that in other months. Starting from the January to June, the mean temperature went up as the humidity decreased. Looking at the slider pictures regarding to mean temperature, humidity and wind speed, we noticed that fluctuations in mean temperature was the opposite for that in humidity across the time, which was reasonable. Because higher temperature, lower chance in raining and less humidity.

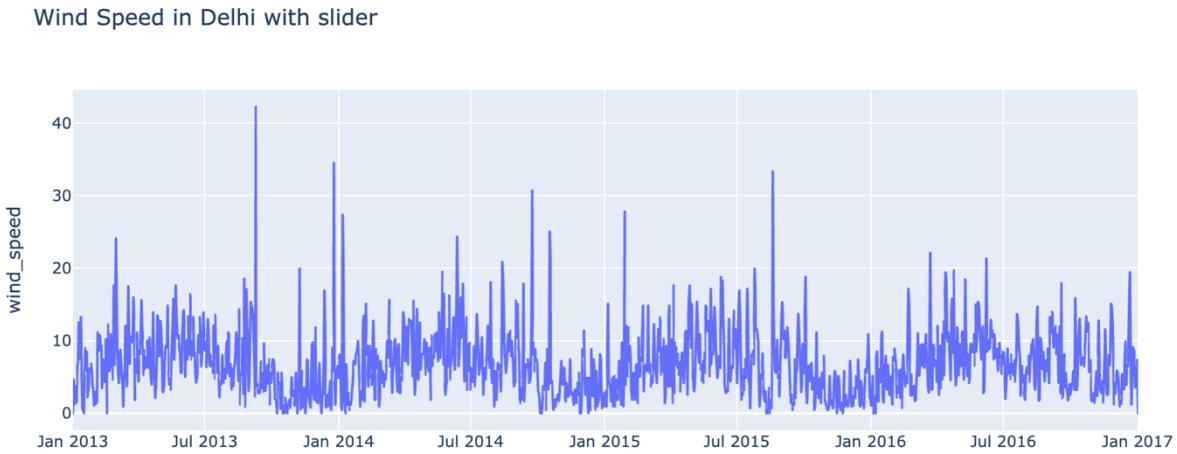



Mean temperature in Delhi with slider






## 3. Method Overview and Results

Naive Bayes techniques are a type of advanced statistical method that use the Bayes theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. The following connection is stated using Bayes' theorem, given a class variable y and a dependent feature vector $x_1$ via $x_n$:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}$$

The naïve conditional independence assumption is presented that:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y), i = 1,2, \dots . n$$

This equation can be simplified to:

$$P(y|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, \dots, x_n)}$$

Therefore, we can use the following classification rule because of the constant inputs:

$$P(y|x_1, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

$$\hat{y} = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

and maximum posterior estimation for $P(y)$ and $P(x_i|y)$ can be calculated. In our dataset, the median of mean temperature is 27.71℃ and we created the target column that considered the mean temperature greater than 27.71℃ is extremely hot noted as 1, otherwise 0. Therefore, Bernoulli naïve Bayes was implemented in our project. The decision rule for Bernoulli naïve Bayes is defined as:

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i)$$

Additionally, we also considered the logistic regression model in this case. Logistic regression is a classification model rather than a regression model. Logistic regression is also known as logit regression, maximum-entropy classification or the log-linear classifier in the literature. A logistic function is used in this model to describe the probability defining the probable outcomes of a single experiment. Next, we split the data into train data and test data. The train data was used for estimation and test data was used for evaluating the model performance. The model prediction accuracy was defined by the accuracy score which is the correct classification rate. We calculated that the accuracy score for Bernoulli naïve Bayes is 0.9 and accuracy score for logistic regression model is 0.76.

## 4. Conclusion

In this project, we mainly demonstrated the Bayesian and logistic regression model performances, respectively. The accuracy score told us the Bayesian approach is better than the logistic regression model on predicting whether the extreme weather is presented or not given some weather characteristics. The reason of logistic regression model not outperforming the naïve Bayes approach could be the existed outliers in predictors. Through this project, we learn there are several advantages using naïve Bayesian approach. First and foremost, the categorization rule is straightforward to grasp. Second, the approach uses only a modest quantity of training data to estimate the classification parameters. Third, the classifier assessment is rapid and simple, and fourth, the approach can be a decent alternative to logistic regression.

However, there are also some limitations of naïve Bayesian method. The assumption of conditional independence of the independent variables is highly unrealistic. The density function must be known or assumed to be normal in the case of continuous independent variables. The probability of categorical independent variables cannot be determined if the count in any conditional category is 0. Moreover, the model interpretability can be worse than other models if there are large number of predictors. In the future study, we can explore more explainable and robust models to predict climate in change such as time series analysis.