Name: Aman Patel

Subject: Report for CS475 Project

Along with economic growth and development, air pollution has become a severe concern in recent decades. According to the World Health Organization, long-term and short-term exposure to toxic air pollutants is considered as the major risk factor in human diseases including cancer, eye irritation, skin diseases, cardiovascular diseases, and respiratory diseases. Air pollution is now the world’s fourth-leading fatal health risk, causing one in ten deaths. The health and environmental agencies have made air pollution control the top priority. The key to air pollution control lies in timely air quality monitoring and accurate prediction on the change of air pollutants, including particle pollution (P2.5 and P10), CO, sulfur oxides, nitrogen oxides, lead, etc.

Above is the provided problem, what it means is that global warming has caused lot of deaths in recent years and the only way to get ahead of this problem is to carefully and systematically analyze the reason why and how its increasing meaning we must find the cause for it and then work against it to nullify that cause or decrease the cause gradually to eventually nullify the global warming. And the best way to analyze this problem is to create a machine learning algorithm which can analyze and then learn the pattern of the data and then successfully predict the data using the other factors in the dataset. This will be able to help us determine the cause of global warming. Now let’s understand the global warming in depth.

Global warming is not just caused by one thing it is caused by collection of deadly gases, now you would ask what these deadly gases are and who is creating them well we humans are for our own pleasure and happiness. This gas consists of CO2 from car, gases that come from the exhaust of cars, airplanes, helicopters, and other motorized vehicles. Refrigerators also emits gases that are deadly for earth. What these gases do is they are thrown in the atmosphere and these gases react with each other and causes to emit heat and this heat causes other reaction and as a result the entire globe starts heating, it’s a slow process that will heat our planet and eventually we wont be able to do anything about it, that is if we don’t act against it now.

So, the purpose of this project will be to gather the data in a csv form and then get data ready to analyze and study and then create two different models for predicting deadly gas level in the atmosphere and then to compare the results of both models. And later on, we will also compare what other parameters or things causes the gas to increase in concentration. In our csv file we will have following datas.

1. Date
2. Cityname
3. Citycode
4. SO2
5. NO2
6. PM10
7. CO
8. O3\_8
9. PM25
10. Station
11. airPressure
12. sunHours
13. highTemperature
14. lowHumidity
15. year
16. month
17. season
18. longitude
19. latitude
20. day

Now let’s understand how we will get our data ready for the models.

As you could see the data set, we have above will have twenty different columns and some of them are very useful and some of those are redundant for the purpose of this project. Like we have date and then we have three different columns for month year and day. So, we don’t need these redundant columns when we are predicting the data which depends on the other columns. So, what we will do is drop all the redundant data and just keep the data on which our to be predicting data is dependent on. Following is the shortlisted columns that I think are necessary for the purpose of this project.

1. Citycode
2. SO2
3. NO2
4. PM10
5. CO
6. O3\_8
7. PM25
8. airPressure
9. sunHours
10. highTemperature
11. lowHumidity
12. year
13. month
14. season
15. day

Here we removed total of five columns and these columns are first we removed “date” column reason we dropped this column is because we already had day, month, and year columns and we don’t need extra date column. Next, we dropped “cityname” column and that was for two reasons first it was a string, and we would have to convert it anyway of we wanted to use it and we already has “citycode” column which would serve the same purpose as the city name. Next up is the station column and reason we dropped that column is because we already had the “citycode” and each city has one “citycode” and one station, so it was not efficient to have that column for our dataset. Lastly, we removed “longitude” and “latitude” columns and the reason behind those removal is nothing but same as the reason for removing the “station”.

Now we will discuss about the two different tree based models that I have used for this project. Let’s discuss what decision trees are and how they are useful for our problem here. What decision tree does is they create a tree with the root splitting the given parameters in smaller nodes and those nodes maybe leaf or maybe they could be non-leaf if they are non-leaf there are more questions to answer and these trees are known as decision trees. And this trees are created based on the test data that we give our model and then train it on our training data. First tree based I have used in my solution is Random Forest. Beauty of the random forest is that most of the other decision models are over fitting to the current dataset but if the new dataset is introduced it changes the results and is no longer accurate, but what random forest does is that it creates different trees randomly with different root nodes and then takes average of all the answers and the result is not the most accurate but consistently average accurate algorithm.

* **Root node:** The base of the decision tree.
* **Splitting:** The process of dividing a node into multiple sub-nodes.
* **Decision node:** When a sub-node is further split into additional sub-nodes.
* **Leaf node:** When a sub-node does not further split into additional sub-nodes; represents possible outcomes.
* **Pruning:** The process of removing sub-nodes of a decision tree.
* **Branch:** A subsection of the decision tree consisting of multiple nodes.

Evaluation:For the evaluation of my models what I have done is that have used methods from the sklearn package like metrics for getting the accuracy of my model compared to test and training dataset. For MAE(Mean Absolute Error) there is a method which allows us to get the answer directly with out writing any codes. Mean Absolute Error means the addition of the different of the actual data and test data with absolute value divided by the total amount of data. Following is the formula for it.

Text

Description automatically generated with medium confidence

To conclude after carful analysis of the data I am sure to say that concentration of PM2.5 depends on the concentration of PM10.