Project Title: *Pollution Predictor*

Group ID: *DMP_27*

Group Members:

Bhagvatsinh Jadeja – 19BCP016

Pathik Viramgama – 19BCP093

Vatsal Sevalia – 19BCP137

Subject Name: *Data Mining Project*

Subject Code: *20CP306T*

Abstract

The Machine we made in this project basically predicts what the pollution will be next year when it is provided with previous year values for some location. For this project we have taken the data set of 240 countries and pollution in them for 8 years i.e. from 2010 to 2017. The data set was taken from the kaggle website whose link you can find in the references (along with the link of a Google Drive which is there just in case the kaggle one gets deleted). We have chosen Linear Regression as our model as it gave the best accuracy among others we implemented.

The results were very promising. The key features are listed below:

- Our model successfully predicted every country's pollution for 2017.
- After exploring different Models and Algorithm, we used Simple Linear Regression as our Model which gave us the best results
- This Model Predicted the values of 2017 Pollution with an accuracy of 99%
- Total Mean Square Error occurred was 1.71 units.
- Data was already pre-cleaned. We have checked for the major points of data cleaning and data check out correctly.

Pollution is a major problem for humans currently. Resources are limited and wasting and corrupting them can be fatal to humans later. Air pollution is a major problem in that category. We tried to make an accurate model to predict air pollution and hence tried to tackle the future prediction for the air pollution of different countries of the world so that we can take necessary steps starting from the present to secure the future.

Index

1. Introduction:

Pollution has been increasing on a daily basis in every country on earth. Air pollution is the major problem humanity is facing right now. In developing countries like India, the rapid increase in population and economic upswing in cities have led to environmental problems such as air pollution, water pollution, noise pollution and many more. Air pollution has a direct impact on human health. There has been increased public awareness about the same in our country. Global warming, acid rains, and an increase in the number of asthma patients are some of the long-term consequences of air pollution. Precise air quality forecasting can reduce the effect of maximal pollution on the humans and biosphere as well. Hence, enhancing air quality forecasting is one of the prime targets for the society. Chemicals are one of the major pollutants present in air. They are colorless and have a nasty, sharp smell. They affect human health when breathed in. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest. The concentration of pollution in the atmosphere can influence the habitat suitability for plant communities, as well as animal life. The proposed system is capable of predicting the concentration of pollution in the Air for the year 2017 from the data of past years 2010 to 2016. So if our Project is able to do it successfully we would already have the “intel” we need and a time of more than one year or so to change the future for better, more or less like going in a time machine and viewing the future then coming back to present and changing the harsh future.

Societal Challenges we want to solve through data mining project:

One of our era's greatest scourges is air pollution, on account not only of its impact on climate change but also its impact on public and individual health due to increasing morbidity and mortality. There are many pollutants that are major factors in disease in humans. Among them, Particulate Matter (PM), particles of variable but very small diameter. Despite the fact that ozone in the stratosphere plays a protective role against ultraviolet irradiation, it is harmful when in high concentration at ground level, also affecting the system. Diseases occurring from the aforementioned substances include principally respiratory problems such as COPD, asthma, bronchiolitis and also lung cancer. Last but not least, climate change resulting from environmental pollution affects the geographical distribution of many infectious diseases, as do natural disasters. The only way to tackle this problem through data mining is to educate the public, coupled with a multidisciplinary approach by scientific experts; national and international organizations must address the emergence of this threat and propose sustainable solutions. We tried to tackle this challenge in this project. We studied the world’s current air toxicity level and what expert’s opinions and other model’s predictions for this are and chose a dataset.

2. Dataset:

Dataset Source: Kaggle Datasets Collection

Name: PM2.5 Global Air Pollution 2010-2017

Publisher: Karl Weinmeister

Structured/Unstructured data: Structured Data in CSV format.

Link of the Dataset:

https://www.kaggle.com/kweinmeister/pm25-global-air-pollution-20102017

(Google Drive link for same dataset if original kaggle one doesn't open)

https://drive.google.com/file/d/1B6OiqIdIUAGq7OK0hJ4eiX8MOc8U4QNX/view?usp=sharing

The dataset consists of around 2400 records of air pollution of major countries of the world for 8 years. It has following Attributes:

1. Country Name
2. Country Code
3. Pollution in 2010
4. Pollution in 2011
5. Pollution in 2012
6. Pollution in 2013
7. Pollution in 2014
8. Pollution in 2015
9. Pollution in 2016
10. Pollution in 2017

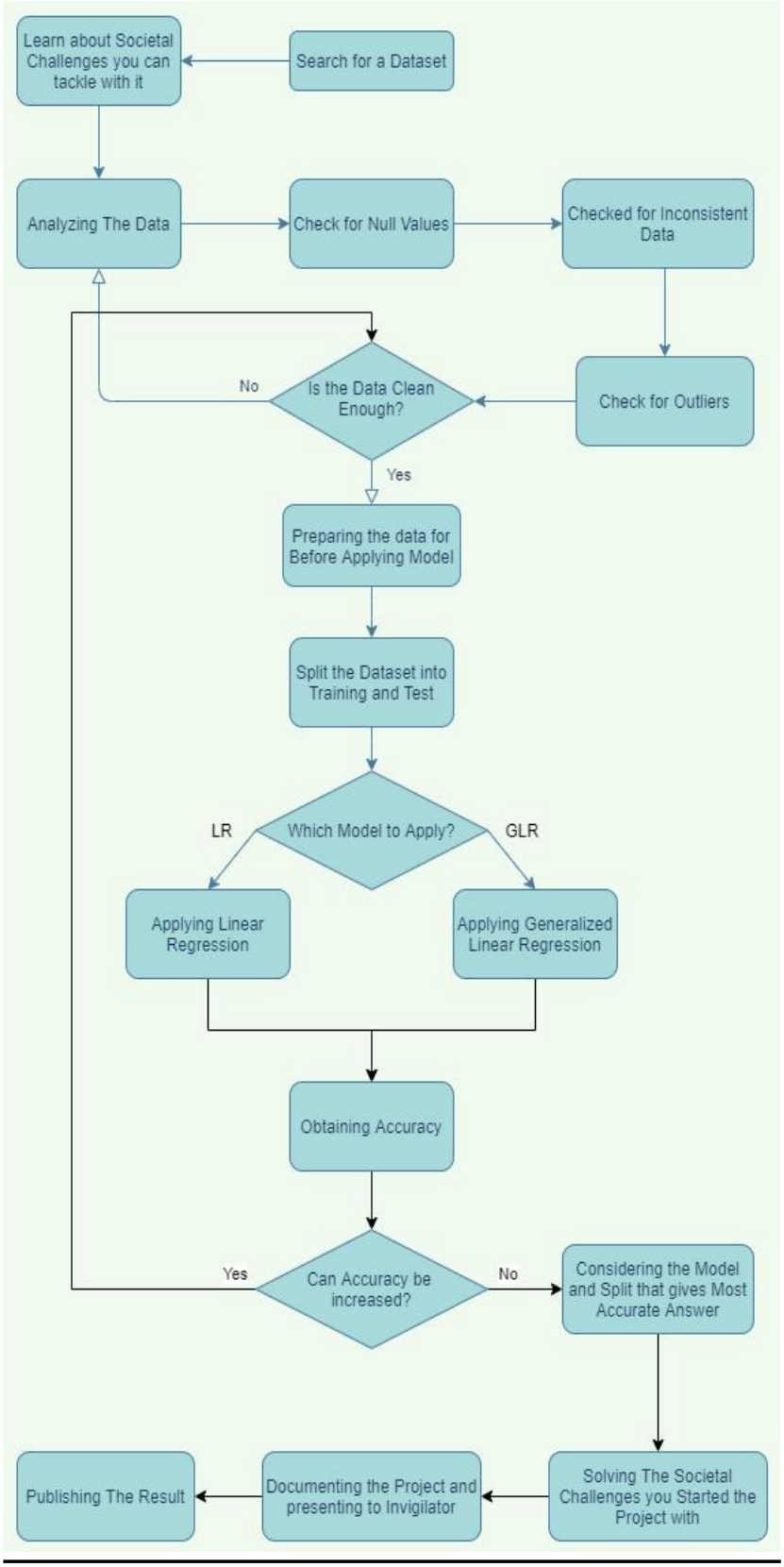
Measurement Unit of Dataset:

PM2.5 refers to atmospheric particulate matter (PM) that has a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair. The unit of Air pollution in the dataset is Mean Annual Exposure (Micrograms per Cubic Meter).

Preprocessing:

We checked for 3 major points of data cleaning i.e. null values, outliers & inconsistent data. Moreover we checked for data duplicity. So after preprocessing our dataset is clean so it still contains 240 rows and 10 columns.

3. Architecture:



The data for Pollution in 2010 to 2016 is taken in as input for each country and the output given by machine is a predicted value of the pollution level in that country for the year 2017. The air pollution is expressed in Mean Annual Exposure (Micrograms per Cubic Meter).

4. Experiment:

We started this project to learn different skill sets. So instead of allocating work according to the skill set of team members, we focused all of our energy on a single problem at a time. From searching for a dataset to creating a model, we have added lots of new experience to our skills.

Google Collab Link:

(Instructions for each line of code is there in the Collab. Go according to sections to understand properly)

https://colab.research.google.com/drive/1vplOK1UgY0yTNDJMyOkRVo92CtFIBKcD?usp=sharing

(Alternatives for Google Collab)

.ipynb File Link:

https://drive.google.com/file/d/1p0ptr6eFV6vOCxrqOixRLlQ0_wAXfQF1/view?usp=sharing

.py File Link:

https://drive.google.com/file/d/1IGj5SVchdLwOeUp-YKchZk-gBFrUHewO/view?usp=sharing

Google Collab File PDF Link:

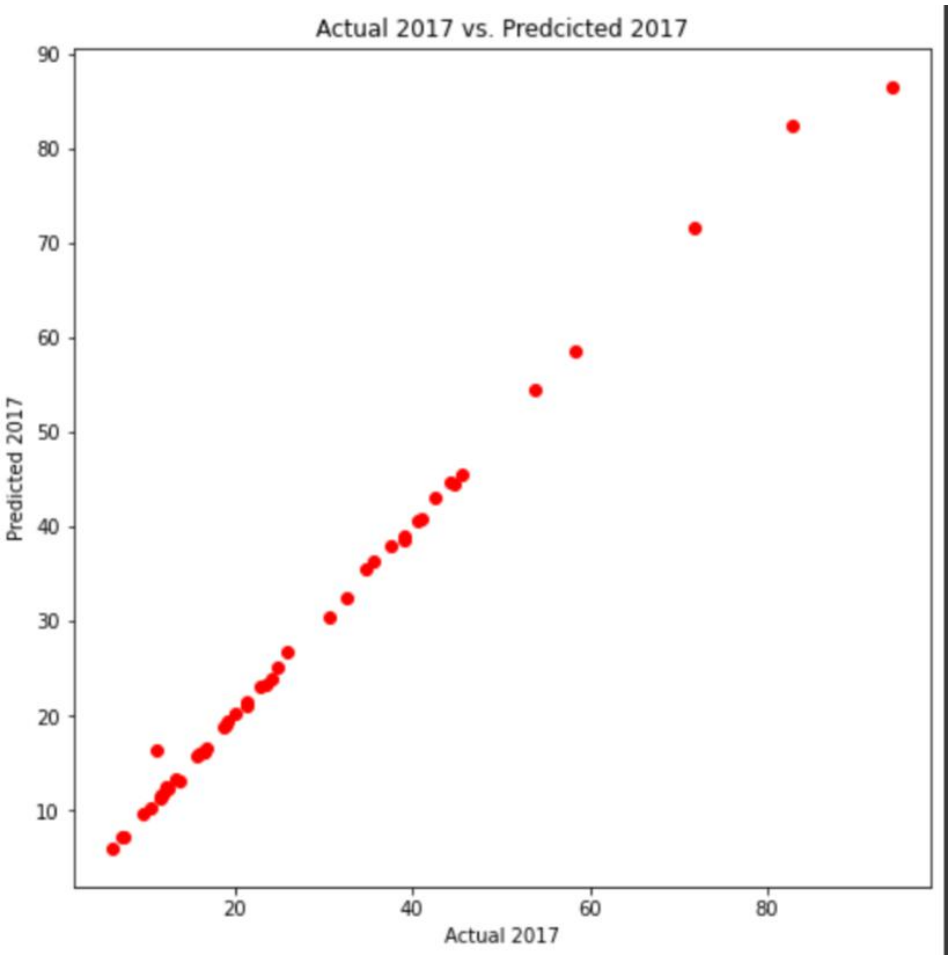
https://drive.google.com/file/d/1dryKcoeQ7VbOaed2nUyse2CKxfIg6aYG/view?usp=sharing

5. Results:

After feeding the Clean dataset to multiple Algorithms in the model, Linear Regression gives the most effective result. And Generalized Linear Regression the second most effective. We will Consider only Linear Regression as result.

- The Model predicts the Air Pollution of 239 countries for the current year (2017) based on the previous 6 years (2010-2016) data for the same.
- Model predicts the target attribute with an accuracy of 99%. This alarming accuracy is because the dataset is squeaky clean and data smooth. This was calculated using the r2_score function from sklearn.metrics library.
- The mean Root Square Error is 1.71 units for the 72 splitted test values. This was calculated using mean_squared_error function from numpy Library

For the 70-30 data split when we plot the result of prediction against Actual Test Values we get this graph. The function we get from Plotting of this graph should be the linear line Y = X for the most accurate result.



For the 30% test cases (72), following is the difference found per example. On Squaring and adding the difference and then taking square root, we get The RMS Error of 1.71 units.

	Country Name	Country Code	Actual Value	predicted value	Difference
109	Kazakhstan	KAZ	13.824288	13.100193	0.724095
71	France	FRA	11.814964	11.699489	0.115474
37	Cote d'Ivoire	CIV	25.886266	26.639946	-0.753679
74	United Kingdom	GBR	10.472690	10.279060	0.193631
108	Japan	JPN	11.704778	11.636303	0.068475
...	...	...	...	...	...
218	Sub-Saharan Africa (IDA & IBRD countries)	TSS	44.602096	44.445224	0.156872
129	Late-demographic dividend	LTE	40.000207	40.596130	-0.595923
73	Gabon	GAB	44.385548	44.017877	0.367672
4	Arab World	ARB	58.689259	59.253161	-0.563902
107	Jordan	JOR	33.006081	33.506304	-0.500223

6. Conclusion:

As addressed earlier we have devoted our resources all together to tackle one problem at a time. We have scratched through the internet to find an appropriate dataset. Then we studied the dataset and looked for societal challenges we can tackle while working on it. After analysis of data, we made our model after studying different models and their application on the internet.

Each and every member of this project team has given their best to make the model as accurate as possible. We achieved the experience of a data analyst and a machine learning enthusiast. We have gained knowledge on different algorithms and techniques while learning about models over the internet. We also obtained an insight on the Air pollution levels in the previous years and also learnt above current levels while searching over the internet.

Results we obtained were very promising. So we can say we solved our societal challenge of getting accurate data for Air Pollution Prediction and giving an advance on decreasing that pollution.

We have tried to do everything possible to get the best result in our project. However there is one shortcoming we detected while doing the project. COVID-19 has been a deadly pandemic virus and has forced people to stay at home. This resulted in lockdowns and curfews. The effects of these were seen later in that year when pollution decreased and also the Ozone Hole healed itself. When the situation got better and lockdowns were lifted off, pollution began to rise again. The data for that part if included in this dataset, our results would have been highly inaccurate for the next year as the model would not have information that there has been a major event (those pandemic years). What we need for that is to add a resistance factor for years with some major event affecting the air pollution as in our case that event is COVID 19. We studied further about customized Neural Networks as a solution for the resistance factor. That is something one can work on to add in our model.