TABLE OF CONTENTS

ABSTRACT

Air pollution affects our daily activities and nature of life, it poses a great threat to humans and the ecosystem that we live in. Owing to the exponential increase in industrial activities over the past few decades, the percent of pollutants in the air we breathe has increased drastically. There is a dire need for the general public to know the extent to which their activities affect air quality. This project proposes to create a publicly accessible dashboard through which anyone with a mobile phone and internet can view various pollutants data over the past 3 years. We have implemented several Machine learning models to forecast and predict for the next 1 month. We have decided to use LSTM (Long Short Term Memory) Machine learning algorithm for this use case as LSTM takes into account the previous data points. We have used RMSE (Root Mean Square Error) for measuring the performance of the LSTM model and the results were quite impressive given the fact that the pollutants data was varying drastically. We have used Tableau Software to create an interactive Visual Dashboard which can display the graphs and analysis for over 15 parameters for the past 3 years and also implemented a forecasting model in the dashboard. This Dashboard is live and can be publicly accessed. The project can be implemented on a national scale and can benefit in providing awareness to the general public.

CHAPTER 01

1.1 INTRODUCTION

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. There are many pollutants that have major role in disease in humans. Among them, Particulate Matter (PM), particles of variable with very small structural diameter, penetrate the respiratory system via inhalation, causing severe respiratory and cardiovascular diseases, reproductive and central nervous system dysfunctions, and cancer. 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 respiratory and cardiovascular system. Furthermore, nitrogen oxide, sulphur dioxide, Volatile Organic Compounds (VOCs), dioxins, and polycyclic aromatic hydrocarbons (PAHs) are all considered air pollutants that are detrimental to humans. Carbon monoxide is a hazard causing direct poisoning when breathed in at high levels. Heavy metals such as lead, when absorbed into the human body, can lead to chronic intoxication, depending on exposure. Diseases occurring from the aforementioned substances include principally respiratory problems such as Chronic Obstructive Pulmonary Disease (COPD), asthma, bronchiolitis, and also lung cancer, cardiovascular events, central nervous system dysfunctions, and cutaneous diseases. 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 is through public awareness coupled with a multidisciplinary approach by scientific experts; national and international organizations must address the emergence of this threat and propose sustainable solutions.

Air lets our living planet breathe—it's the mixture of gases that fills the atmosphere, giving life to the plants and animals that make Earth such a vibrant place. Air is almost entirely made up of gases (78 percent nitrogen and 21 percent oxygen), with a few other gases (such as carbon dioxide and argon) present in much smaller quantities. Air is a core element for the sustenance of life. The cleaner the air, the better your health and well-being. However, various sources, especially anthropogenic, are posing a significant threat to air quality. In this article, we take a closer look at the situation in India, a country whose rapid expansion has seriously diminished air quality. Central Pollution Control Board, a statutory organization under the Ministry of Environment and Forests, Government of India, provides air quality data and AQI at an hourly and daily basis of various stations across cities in India. By analysing the daily AQI data for different cities of India.

Air pollution in India is a serious health issue. Of the 30 most polluted cities in the world, 21 were in India in 2019. ... The 51% of pollution is caused by the industrial pollution, 27 % by vehicles, 17% by crop burning and 5% by fireworks.
The alarming statistics have raised concerns about the poor air quality in the country and the associated health risks. The prevailing air quality issue must be addressed immediately. Air pollution policies and regulation must be revamped, sustainable development must be promoted, and the general public must be enlightened about the health risks associated with poor air quality and how they can mitigate the problem by embracing small yet significant changes in their day to day life.

1.2Air Pollutants:

The World Health Organization (WHO) reports on six major air pollutants, namely particle pollution, ground-level ozone, carbon monoxide, sulphur oxides, nitrogen oxides, and lead. Air pollution can have a disastrous effect on all components of the environment, including groundwater, soil, and air. Additionally, it poses a serious threat to living organisms. In this vein, our interest is mainly to focus on these pollutants, as they are related to more extensive and severe problems in human health and environmental impact. Acid rain, global warming, the greenhouse effect, and climate changes have an important ecological impact on air pollution.

1.3 Sources of exposure:

It is known that the majority of environmental pollutants are emitted through large-scale human activities such as the use of industrial machinery, power-producing stations, combustion engines, and cars. Because these activities are performed at such a large scale, they are by far the major contributors to air pollution, with cars estimated to be responsible for approximately 80% of today's pollution.

1.3.1 Major sources include the emission of pollutants from power stations, refineries, and petrochemicals, the chemical and fertilizer industries, metallurgical and other industrial plants, and, finally, municipal incineration.

1.3.2 Indoor area sources include domestic cleaning activities, dry cleaners, printing shops, and petrol stations.

1.3.3 Mobile sources include automobiles, cars, railways, airways, and other types of vehicles.

1.4 Particulate matter (PM) and Health:

Studies have shown a relationship between particulate matter (PM) and adverse health effects, focusing on either short-term (acute) or long-term (chronic) PM exposure. Particulate matter (PM) is usually formed in the atmosphere as a result of chemical reactions between the different pollutants. The penetration of particles is closely dependent on their size. Particulate matter (PM) pollution includes particles with diameters of 10 micrometers (µm) or smaller, called PM_{10}, and extremely fine particles with diameters that are generally 2.5 micrometers (µm) and smaller. Particulate matter contains tiny liquid or solid droplets that can be inhaled and cause serious health effects.

1.5 Penetrability according to particle size

>11 µm	Passage into nostrils and upper respiratory tract
7–11 µm	Passage into nasal cavity
4.7–7 µm	Passage into larynx
3.3–4.7 µm	Passage into trachea-bronchial area
2.1–3.3 µm	Secondary bronchial area passage
1.1–2.1 µm	Terminal bronchial area passage
0.65–1.1 µm	Bronchioles penetrability
0.43–0.65 µm	Alveolar penetrability

1.6 Consequences of Air Pollution:

There is concrete evidence that air pollution leads to low birth-weight, tuberculosis, ischemic heart disease, cataracts, asthma and nasopharyngeal and laryngeal cancers. Air pollution is linked to diseases and infections that kill around 600,000 children under five years of age per year about 2.2 million school children in Delhi are growing up with irreversible lung damage which they will never recover. The number of premature deaths due to outdoor air pollution is projected to increase from three million people globally in 2010 to a global total of six to nine million people in 2060. The number of cases of bronchitis is projected to increase substantially, going from 12 to 36 million new cases per year for children aged six to twelve and from 3.5 to 10 million cases for adults.

Total welfare losses due to air pollution in India amounted to more than $500 billion (8.5% of country's GDP) in the year 2013 (381% increase from 1990).

Total welfare costs of air pollution in the world is expected to increase from $3,160 billion in 2015 to $18,300 – $25,330 billion in 2060.

In developing countries, the problem is more serious due to overpopulation and uncontrolled urbanization along with the development of industrialization. This leads to poor air quality, especially in countries with social disparities and a lack of information on sustainable management of the environment. The use of fuels such as wood fuel or solid fuel for domestic needs due to low incomes exposes people to bad-quality, polluted air at home. It is of note that three billion people around the world are using the above sources of energy for their daily heating and cooking needs. There is spatial heterogeneity in India, as areas with diverse climatological conditions and population and education levels generate different indoor air qualities, with higher $PM_{2.5}$ observed in North Indian states (557–601 $\mu g/m^3$) compared to the Southern States (183–214 $\mu g/m^3$). The cold climate of the North Indian areas may be the main reason for this, as longer periods at home and more heating are necessary compared to in the tropical climate of Southern India. Household air pollution in India is associated with major health effects, especially in women and young children, who stay indoors for longer periods. Chronic obstructive respiratory disease (CORD) and lung cancer are mostly observed in women, while acute lower respiratory disease is seen in young children under 5 years of age.

Technological innovation can only be successful if it is able to meet the needs of society. In this sense, technology must reflect the decision-making practices and procedures of those involved in risk assessment and evaluation and act as a facilitator in providing information and assessments to enable decision makers to make the best decisions possible. Summarizing the aforementioned in order to design an effective air quality control strategy, several aspects must be considered: environmental factors and ambient air quality conditions, engineering factors and air pollutant characteristics, and finally, economic operating costs for technological improvement and administrative and legal costs. Considering the economic factor, competitiveness through neoliberal concepts is offering a solution to environmental problems.

Since the 1990s, the term "digital activism" has been used increasingly and in many different disciplines. Nowadays, multiple digital technologies can be used to produce a digital activism outcome on environmental issues. More specifically, devices with online capabilities such as computers or mobile phones are being used as a way to pursue change in political and social affairs.

1.7 Air pollution scenario in India

As India is fastest developing country median age rises from 27 in 2019 to 32 in 2030, vulnerability to air pollution will increase illnesses, and mortality says clean air fund report. Air pollution costs Indian businesses about $95 billion (about ₹7-lakh crore) every fiscal year, around 3 per cent of India's total GDP, according to a major research report. India has grown to become the world's fifth most polluted country in the last decade and has 21 of the world's 30 most polluted cities. Air pollution affects the overall health of businesses and the economy. While the government has taken aggressive measures to address the issue, the emphasis on air pollution across the globe has continued to be on its public health implications. It has now become important for Indian business to include air emissions in their profit and loss statements."

"Clean air is a precondition for businesses to thrive — and for India to realise its vision of becoming a $5-trillion economy by 2025. Achieving this goal would require industry leaders to take more ownership and become advocates in the movement for cleaner air."

1.8 Need for monitoring

Clean air is vital need for every human being. Polluted air causes many health problems and several damages. Therefore to make any step ahead of controlling the pollution rate it is necessary to monitor the air quality which may help us to make a right decision at right time. There are various causes of increasing the pollution such as smoke automobile exhaust, chemical discharge from industries, radioactive substance etc. These are main reason of decreasing the quality of air. The main gases which directly affect the human health are carbon monoxide (CO), hydrogen sulphide, sulphur dioxide ($SO2$), Nitrogen dioxide ($NO2$) and the main contribution of these gases are traffic related pollutant emission. Huge efforts are required to improve the quality of air in both outdoor and indoor environment. Monitoring of environment has been controlled from manual to the automatic control step by step. There are various improvement in the instrument of environment monitoring but still cannot meet the harsh environment.

1.8.1 Technique used for pollution monitoring

Previously the air pollution monitoring is done via computerized tomography technique which generate a two dimensional map of pollutant concentration. It provides a many advantages over the differential absorption method. In this system there is a single laser source located at the centre of the area. This laser beam is rotated and directed towards the circumference of the circle. There is a cylindrical mirror so that incident laser beam is reflected in a fan beam over angle across the circle. The beam from the mirrors is the circular region and strikes a set of detectors lie in same plane parallel to the ground. This technique focus on lower transmitted laser energy increasing the range and ability to monitor the area that contains several pollutant sources [3]. Another way of monitoring the air pollution is via the online GPRS sensors array which has been designed, implemented and tested. This system unit that consists at a single chip of microcontroller and a pollution server which is a high end personal application server with an internet connectivity where the mobile data acquisition unit that collect the pollution level & pack it into a frame with GPS location, date and time. This frame is uploaded to the GPRS modem and transmitted to the pollution server via the public mobile network. A data base server which is attached to the pollution level which is used by the various client. Pollution server for storing the pollution level which is used by the various clients. Pollution server having an interfaced with the Google map to provide a real time pollutants level as well as the location in large metropolitan area

1.8.2 Wireless sensor network for real time monitoring

A distributed infrastructure consists of a wireless sensor network and grid computing technology for air pollution monitoring as well as mining. However, the two layer network architecture and peer to peer e-science grid architecture and distributed data mining algorithms are used in order to collect the data and tiny operating system is used to examine the operation and performance of the wireless sensor network [5].Wireless sensor network is the great achievement in this field. An effective solution for the pollution monitoring using a wireless sensor network to provide a real time pollution data. The various gases like CO2, NO2 are calibrated by using an appropriate calibration technologies and these precalibrated sensors are integrated with the wireless sensor using a multi hop data aggregation algorithm. A light weight middleware and web interface in order to view the one pollution data in the form of charts and number. It is also available on the internet. The other parameters like temperature and humidity are also sensed along with the gas concentrations which enable the data analysis through the data fusion techniques this system provide accurate pollutant data [6]. The air quality monitoring system combines with the virtual instrument technology & frequency hopping communication technology to achieve the wireless data transmission. By using a spectrum hole detection specimens that adjust a carrier frequency according to the result & made a full use of available radio spectrum with this specimen there is no signal interference during the wireless transmission process & the system can receive the real time information effectively and the gas concentration can show clearly and easy to read by the nonprofessional staff also [7]. The air quality monitoring station are used to monitor the quality of air but most of this method are expensive and provide a low resolution sensing data and these stations are less densely deployed therefore the system consist of sensor mode gateway and back end platform controlled by the lab view program through which the data can be stored in the database the system deployed to the main road in the city to monitor the carbon monoxide concentration caused by the vehicle emission the advantages of these wireless sensor network is that it is easy to set up, inexpensive and also provide a real time data[8]. The system in which several monitoring station communicate wirelessly with the backend server using machine to machine communication & each station equipped with the metro logical sensor and gaseous sensor for data logging and wireless communication capabilities. The backend server collects the real time data from station and converts it in to the information which is used by the user through the web portals and mobile application

1.8.3 Pollution level monitor over the Google map

The main objective of monitoring is to display the collected information in user friendly format. The mobile application and websites are developing in order to display the real time data that contains previous history and recent measurement of pollution level. Only the authorized user can access the website which is easily available to the public when the permission is granted. Website allows displaying the different level of pollution in different area over the Google map with the help of internet connectivity it is possible to display the different level of pollution at different area on the Google map

1.8.4 Centralized monitoring

Different sensors are deploying to the different region and each sensor must send their collected information to server so that the end user can easily see the pollution information in the different area. Centralize monitoring ensure the quality, improve the ability and integrity of data. Collected data are uploaded to the cloud dataset so that it can be analyse or viewed for future use. All these uploaded data are managed in database management system over the centralize database with this available information the user can search the record as per their requirement

1.8.5 Types of sensor

There are different types of sensors are available for collecting the atmospherically data. Such as Temperature sensor, Humidity sensor, Rain sensor, Gas Sensor etc.., different types of gas sensors are available to collect the different gases from the road traffic emission such as CO2 sensor, NO_2 sensor, SO_2 sensor etc. Wireless sensor network built a node where each node is connected to one sensor. With the additional sensors may help to enhance the network and monitoring the additional pollutants. With the help of sensors it may possible to collect the environment related information. It is deployed in several cities to monitor the concentration of dangerous gases for citizen. Air quality measurement can process and presented in a real time to the end user in a friendly format to spread environmental awareness among the population and allow taking appropriate precaution when it is needed.

Keywords: Parts per million, Air pollution monitoring, Machine Learning, Forecasting. CPCB Data Monitoring.

1.9 Objective:

1. To study the present air quality status of air pollution in Visakhapatnam city.
2. To compare various software techniques for monitoring and forecasting of air pollution
3. To create a prediction model using LSTM Machine learning
4. To create a dashboard for visualization entire prediction graphs for public use and awareness.

1.10 Study area:

Visakhapatnam is a city located in the south-eastern region of India, on the coastline facing onto the Bay of Bengal. It has been given names such as the jewel of the east coast, or the city of destiny. With such fanciful names given, duly it is also a prominent tourist destination famous for its beaches, as well as acting as a major cargo port for the whole of India. As one would expect, with activities like tourism and transportation being of chief importance in Visakhapatnam there would also be a large amount of air pollution related to these industries.

As it stands, over the year of 2019 Visakhapatnam came in with an average reading of 44 $\mu g/m^3$ in regards to its levels of PM2.5 in the air, placing it into 'unhealthy for sensitive groups' bracket, which requires a PM2.5 reading of 35.5 to 55.4 $\mu g/m^3$ to be classed as such. This group rating would show that the pollution levels are somewhat high, and would have detrimental effects on those who are vulnerable to air pollution, such as pregnant mothers, young children and those with respiratory related conditions.

This reading of 44 μg/m³ placed Visakhapatnam into 180th place in regards to most polluted cities worldwide in 2019, and 58th place out of all cities ranked in India over the same year. It saw a large amount of fluctuating pollution levels, with some extremely high readings followed by months of significantly improved air quality

The main causes of pollution in Visakhapatnam, also known locally as Vizag, have a number of differing sources, some of which are more prominent than others. As a place that experiences large volumes of both international and local tourism, Vizag would be subject to large amounts of vehicle usage, with a plethora of automobiles such as cars, buses and motorbikes moving in and out of the city, along with the large amount of industrial and recreational ships coming in to dock at the city's ports.

Due to this, there would be a large amount of pollution coming from vehicle and ship fumes, many of which would still be relying on fossil fuels such as diesel, which can lead to a considerably larger output of pollute haze than their cleaner counterparts. There is also the industrial sector to consider, with a large number of factories and other production plants dotted around the city's limits, with ones such as chemical production and garment production factories being prominent.

In short, the main causes of pollution in Vizag would be vehicular and ship emissions, smoke and haze from factories, as well as the large number of industrial materials being moved through the port area, often giving off large amounts of particulate matter. With heavy loads of items such as coal being transported in large quantities, they would give off huge clouds of soot and carbon particles through their movement alone, polluting the air long before they make their way to a furnace for combustion.

Chapter 02

Review of Literature:

Smart air pollution monitoring consists of wireless sensor nodes, server and a database to store the monitored data. Huge amounts of data are generated by gas sensors in air pollution monitoring system. Traditional methods are too complex to process and analyse the voluminous data. The heterogeneous data are converted into meaningful information by using data mining approaches for decision making. The K-Means algorithm is one of the frequently used clustering methods in data mining for clustering massive data sets enhanced K-Means clustering algorithm is proposed to analyse the air pollution data. The correlation coefficient is calculated from the real time monitored pollutant datasets. The Air Quality Index (AQI) value is calculated from the correlation co-efficient to determine the air pollution level in a particular place. The proposed enhanced K-Means clustering algorithm is compared with Possibility Fuzzy C-Means (PFCM) clustering algorithm in terms of accuracy and execution time. Experimental results show that the proposed enhanced K-Means clustering algorithm gives AQI value in higher accuracy with less execution time for when compared to existing techniques.

Kingsy Grace. R I , Manimegalai. R 2 , Geetha Devasena. M.SI , Rajathi. Si , Usha. KI , Raabiathul Baseria. N (2016 IEEE Region 10 Conference (TENCON))

The meteorological parameters have a vital role in the distribution, transition as well as elimination of pollutant concentration in the atmosphere. the influence of climate parameters on the intensity of pollutants is analysed. To analyse the influence, multiple linear regression and artificial neural network techniques are applied. The effect of meteorological data on pollutant concentration data such as particulate matter, nitrogen oxides, carbon monoxide of Coimbatore region are used in this work. The parameters that are positively correlated with the atmospheric pollutants are utilised to estimate the pollutant concentration of PM2.5, NOx and CO through MLR and ANN models. Performance evaluation of the models is done by using statistical measures and it is found that ANN has resulted in better estimation.

(A. Sangeetha, T. Amudha)

2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)\

Over the last ten years, Salamanca has been considered among the most polluted cities in México. Nowadays, there is an Automatic Environmental Monitoring Network (AEMN) which measures air pollutants (Sulphur Dioxide (SO2), Particular Matter (PM10), Ozone (O3), etc.), as well as environmental variables (wind speed, wind direction, temperature, and relative humidity), and it takes a sample of the variables every minute. The AEM Network is mainly based on three monitoring stations located at Cruz Roja, DIF, and Nativitas. In this work, we use the PFCM (Possibilistic Fuzzy c Means) clustering algorithm as a mean to get a combined measure, from the three stations, looking to provide a tool for better management of contingencies in the city, such that local or general action can be taken in the city according to the pollution level given by each station and the combined measure. Besides, we also performed an analysis of correlation between pollution and environmental variables. The results show a significative correlation between pollutant concentrations and some environmental variables. So, the combined measure and the correlations can be used for the establishment of general contingency thresholds.
(B. Ojeda-Magaña1, M. G. Cortina-Januchs2, J. M. Barrón-Adame3, J. Quintanilla-Domínguez2, W. Hernandez 4, A. Vega-Corona3, R. Ruelas1 and D. Andina) 2010 IEEE
They present a knowledge based approach applied to air pollution effects analysis in the case of PM2.5 air pollutant which has potential significant negative effects on human health. The use of knowledge derived from various sources (e.g. literature, databases, questionnaires, human experts' experience, and decision tables) via manual, semiautomatic and automatic methods is proposed for a multiparameters analysis of the PM2.5 air pollution episodes effects on vulnerable people such as children and elderly. Some measures to reduce the negative effects on human health are also proposed by our approach. The knowledge under the form of production rules is incorporated in a knowledge base that is used by the ROKIDAIR intelligent decision support system (ROKIDAIR DSS). The knowledge base coherence was verified with the expert systems generator, VP-Expert. The experimental data sets that were used are for some air pollution monitoring sites situated in the Ploiesti city and included in the Romanian National Air Quality Monitoring Network.
Mihaela Oprea, Hai-Ying Liu
They present a novel method of real-time air pollution analysis. First, we acquire real-time air quality data in several methods via Internet. Second, by combining them with population, elevation, and industry information, we analyse the air quality using the classification model of random forest. Last, we generate on-the-fly thematic maps automatically to visualize the classification results. The development of sensors, Internet of things, and Internet technology,

provide a convenient way for real-time access to air quality data, but also lay a good foundation for further development of relevant applications. This paper presents a novel method of real-time air pollution analysis. First, we acquire real-time air quality data in several methods via Internet. Second, by combining them with population, elevation, and industry information in the region, we analyse the air quality using the classification model of random forest. Last, we generate on-the-fly thematic maps automatically to visualize the classification results.
Chen Chen, Qiang Liu, Pei-Wen Liu, Yao-Sen Huang, Wei-Qing Li, Hao Luo
2017 International Conference on Network and Information Systems for Computers (ICNISC)
This study is concerned with computation methods for environmental data analysis in order to enable facilitate effective decision making when addressing air pollution problems. A number of environmental air pollution studies often simplify the problem but fail to consider the fact that air pollution is a spatio-temporal problem. This research addresses the air pollution problem as spatio-temporal problem by proposing a new decentralized computational technique named Online Scalable SVM Ensemble Learning Method (OSSELM). Special consideration is given to distributed ensemble in order to resolve the spatio-temporal data collection problem i.e., the data collected from multiple monitoring stations dispersed over a geographical location. Moreover, the air pollution problem is address systematically including computational detection, examination of possible causes, and air-quality prediction. The proposed OSSELM uses spatio-temporal decentralised analysis to facilitate effective decision making when addressing air pollution problems. It follows that decentralised analysis is vital, as air pollution datasets are obtained from multiple monitoring stations. Obtaining these distributed datasets and then combining them through a centralised computing leads to few concerns, time consumption and privacy are few to names. In this regard, we believe that OSSELM, an SVM aggregation distributed computation technique, is able to address this problem more appropriately than previously reported methods. This work offers significance in the area of big data for computational environmental applications. Moreover, OSSELM presents a novel approach to air pollution prediction via knowledge fusion across spatio-temporal dimensions.
Shahid Ali, Sreenivas Sremath Tirumala, Abdolhossein Sarrafzadeh (2014 IEEE)
The advent of the Internet of Things (IoT) has led to the generation of tremendous amounts of data from various sources. Cloud based systems are effective in storage and application of machine learning algorithms on such datasets. However, in some cases it is important to enable real time processing for making immediate decisions. There are many applications which require instantaneous analysis of the generated data for remedies in event of an anomaly. Data

associated with such use cases remains significant only for a short duration of time. Various electronic sensors, e.g. Temperature, Moisture, Air Quality, Pressure, Wind Velocity etc. present in a Wireless Sensor Network generate streams of values. It can be processed using pipelines which provide prompt and quick analysis for decision making. Stream Processing Systems can be helpful in such cases as they analyse data streams within a few milliseconds to a few seconds. In this paper, we discuss an event based processing of streaming data from air pollution sensors to create a real time anomaly detection system. To reduce the delays associated with the generation of alarms in our pipeline, Apache Foundation's Stream Processing Tools, Kafka and Flink were used for operations on our streams. To further accelerate the process, all the analysis is conducted on an embedded edge computing gateway device rather than sending data to the cloud for batch processing. The results are obtained in the form of a geographical map visualization using ELK stack. The map highlights the coordinates of the location with an unhealthy air quality index in real time.

Utkarsh Kulshreshta, Surya Durbha

IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium

Air pollution is one of the largest environmental and public health challenges in the world today. It leads to adverse effects on human health, climate and the ecosystem. Many techniques for monitoring air pollution have been applied over the years. This paper discusses a survey of air pollution monitoring and what has been done to alleviate its effects. Various air polluting methods are discussed based on recent literature. A further discussion is included to expose the ills of air pollution and how various studies have sought to solve the problem. A special emphasis on South Africa is considered whereby rising levels of pollution is prevalent due to industrial and traffic radiation. The studies show that South Africa is still faced with a crisis in infrastructure. Furthermore, this research focuses on intelligent air pollution monitoring and deliberates on its applications and benefits.

Ntombikayise Koyana, Freestate, South Africa ; Elisha Didam Markus; Adnan M. Abu-Mahfouz

2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)

Environmental parameter monitoring has become major concern in modern megalopolis due to revolution and advancement. The designed system uses Internet of Things (IOT) which provides an economical and an effective system to monitor air pollution level in particular area. IOT empowers vast scope of entities and physical world to be communicated and monitored in fine details. For bestow fascinating services, to exchange and communicate information, IOT embeds connectivity with decision making capability among devices can be used. The strategy of system defines a customized design of IOT based monitoring devices which determine the levels of toxicity in gases in the atmosphere. Using different sensors and GPS connected to ESP8266 air pollution can be determined. The ESP module bridged to cloud, helps to determine the location and pollution level. This integrated system will help in studying testimony of smart city in order to justify the steps taken to control the pollution level.

G Spandana, R Shanmughasundram

2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)

Air quality, especially particulate matter, has recently attracted a lot of attention from governments, industry, and academia, motivating the use of denser air quality monitoring networks based on low-cost sensing strategies. However, low-cost sensors are frequently sensitive to aging, environmental conditions, and pollutant cross-sensitivities. These issues have been only partially addressed, limiting their usage. In this study, we develop a low-cost particulate matter monitoring system based on special-purpose acquisition boards, deployed for monitoring air quality on both stationary and mobile sensor platforms. We explore the influence of all model variables, the quality of different calibration strategies, the accuracy across different concentration ranges, and the usefulness of redundant sensors placed in each station. The collected sensor data amounts to about 50GB of data, gathered in six months during the winter season. Tests of statically immovable stations include an analysis of accuracy and sensors' reliability made by comparing our results with more accurate and expensive standard β-radiation sensors. Tests on mobile stations have been designed to analyze the reactivity of our system to unexpected and abrupt events. These experiments embrace traffic analysis, pollution investigation using different means of transport and pollution analysis during peculiar events. With respect to other approaches, our methodology has been proved to be extremely easy to calibrate, to offer a very high sample rate (one sample per second), and to be based on an open-source software architecture.

Bartolomeo Montrucchio; Edoardo Giusto; Mohammad Ghazi Vakili; Stefano Quer

IEEE Transactions on Vehicular Technology

Fine ambient aerosols (PM2.5) levels in the atmosphere are continuously worsening over Delhi and National Capital Region (NCR) of India. Complete source profiles are required to be assessed for implementation of proper mitigation measures over the NCR. In this study, emission sources of PM2.5 are reported for the NCR of India for samples collected during December 2016 to December 2017 at three sampling sites in Delhi, Uttar Pradesh and Haryana.

Organic constituents (n-alkanes, isoprenoid hydrocarbons, polycyclic aromatic hydrocarbons, phthalates, levoglucosan and n-alkanoic acids) in PM2.5 were measured to apportion the sources over the study area. Source apportionment of PM2.5 was performed using organic constituents by Positive Matrix Factorization (PMF) and Principal Component Analysis (PCA). Health risk associated with organic pollutants [PAHs and carcinogen BEHP bis(2-ethylhexyl) phthalate] demonstrated the threat of PM2.5 exposure via inhalation. Transport pathways of air masses were evaluated using 3-day backward trajectories and observed that some air masses originated from local sources along with long-range transport which influenced the PAHs concentration during most of the study period over the NCR. PMF and PCA resulted in the five major emission sources [vehicular emissions (32.2%), biomass burning (30%), cooking emissions (16.8%), plastic burning (13.4%), mixed sources (7.6%) including biogenic and industrial emissions] for PM2.5 over the sampling sites. The present study reveals that transport sector is a major source to be targeted to reduce the vehicular emissions and consequent health risks associated with organic pollutants especially PAHs.

Ranu Gadia, Shivania Sudhir, Kumar S, harmab Tuhin, Kumar Mandalb

Chemosphere Volume 221, April 2019, Pages 583-596

Chapter 03

Proposed Methodology:

3.1 What is Machine Learning?

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data.[1] It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.[2] Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

3.2 What is LSTM?

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems). A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

3.3 How did we measure the performance of the model?

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These deviations are called residuals when the calculations are performed over the data sample that was used for estimation and are called errors (or prediction errors) when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various data points into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.[1] RMSD is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower RMSD is better than a higher one. However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used.

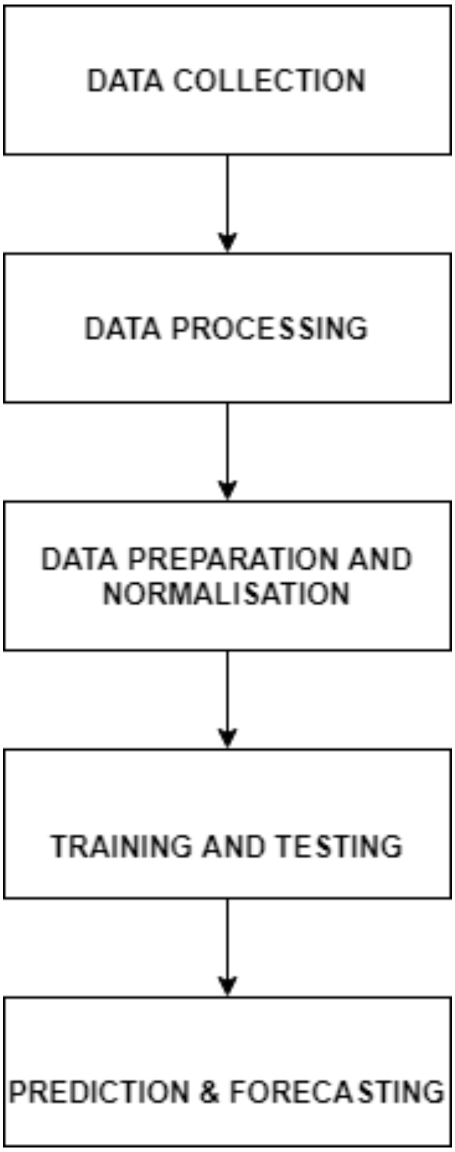
$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}.$$

RMSD = Root-Mean-Square-Deviation.

y_t = Observed Values

$\hat{y}_t$ = Expected Values

T = Total values



3.4 Data Collection:

We collected the required pollutant and meteorological data from CPCB website have collected the following parameters

1. PM 2.5
2. PM 10
3. NO
4. NO_2
5. NOx
6. NH_3
7. SO_2
8. CO
9. OZONE

10. BENZENE
11. TOLUENE
12. TEMPERATURE
13. RELATIVE HUMIDITY
14. WIND SPEED
15. XYLENE

collected the data from its oldest date digitally recorded in Visakhapatnam. From July 2017 to April 2021. I have made sure to include both pollutant and meteorological data for in-depth analysis.

3.5 Data Processing:

The collected data was in a very dis-organised and sophisticated manner. This led us to process the data to enable meaningful inference making later. There were many missing values in the data, we have assigned the mean value of the data to missing data points. we could also have removed the rows of data with even one missing data point, but chose putting mean values because the number of missing values was large enough to diminish the dataset. we also employed Data Normalisation to make the data sensitive for model training and testing. Data Normalisation is the process of bringing down the actual values to a scale of 0-1.

a) Dealing with missing data
b) Normalising data

3.6 Data Preparation:

To train and test the model, we had to prepare the dataset into two partitions namely Train Data and Test Data. we have used 90% of the data in spite of the industry standard being 80-20 or 75-25, because due to covid the previous several months of data follows a different trend due to the change in pollution levels and lockdown effects, having a large training data set allows the model to train on covid affected months too.

3.7 Prediction & Forecasting:

we have trained the model to use the previous 100 days of data. This concept is called LSTM Training, which used the previous data to predict the future variables. We have forecasted each parameter for the next 30 days after 18th April 2021. The forecast for each of the parameter is present in the later part of this document

Code:

```
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import MinMaxScaler
import tensorflow as tf
import math
from sklearn.metrics import mean_squared_error
df=pd.read_csv('pollution.csv')

df.head()

df1=df.reset_index()['PM10']
plt.title('Overall Data, values shown in µg/m3')
plt.plot(df1)
plt.savefig('PM/Overall data.png')

#Normalisation
scaler=MinMaxScaler(feature_range=(0,1))
df1=scaler.fit_transform(np.array(df1).reshape(-1,1))
print(df1)

##splitting dataset into train and test split
training_size=int(len(df1)*0.90)
test_size=len(df1)-training_size
train_data,test_data=df1[0:training_size,:],df1[training_size:len(df1),:1]
training_size,test_size

train_data

# convert an array of values into a dataset matrix
def create_dataset(dataset, time_step=1):
        dataX, dataY = [], []
        for i in range(len(dataset)-time_step-1):
                a = dataset[i:(i+time_step), 0]   ###i=0, 0,1,2,3-----99   100
                dataX.append(a)
                dataY.append(dataset[i + time_step, 0])
        return np.array(dataX), np.array(dataY)

# reshape into X=t,t+1,t+2,t+3 and Y=t+4
time_step = 100
X_train, y_train = create_dataset(train_data, time_step)
X_test, ytest = create_dataset(test_data, time_step)

print(X_train.shape), print(y_train.shape)

print(X_test.shape), print(ytest.shape)

# reshape input to be [samples, time steps, features] which is required for LSTM
X_train =X_train.reshape(X_train.shape[0],X_train.shape[1] , 1)
```

```
X_test = X_test.reshape(X_test.shape[0],X_test.shape[1] , 1)

### Create the Stacked LSTM model
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.layers import LSTM

model=Sequential()
model.add(LSTM(50,return_sequences=True,input_shape=(100,1)))
model.add(LSTM(50,return_sequences=True))
model.add(LSTM(50))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')

model.summary()

model.fit(X_train,y_train,validation_data=(X_test,ytest),epochs=100,batch_size
=64,verbose=1)

### Lets Do the prediction and check performance metrics
train_predict=model.predict(X_train)
test_predict=model.predict(X_test)

##Transformback to original form
train_predict=scaler.inverse_transform(train_predict)
test_predict=scaler.inverse_transform(test_predict)

### Calculate RMSE performance metrics

math.sqrt(mean_squared_error(y_train,train_predict))

### Test Data RMSE
math.sqrt(mean_squared_error(ytest,test_predict))

### Plotting
# shift train predictions for plotting
look_back=100
trainPredictPlot = np.empty_like(df1)
trainPredictPlot[:, :] = np.nan
trainPredictPlot[look_back:len(train_predict)+look_back, :] = train_predict
# shift test predictions for plotting
testPredictPlot = np.empty_like(df1)
testPredictPlot[:, :] = np.nan
testPredictPlot[len(train_predict)+(look_back*2)+1:len(df1)-1, :] = test_predict
# plot baseline and predictions
plt.plot(scaler.inverse_transform(df1))
plt.plot(trainPredictPlot)
plt.plot(testPredictPlot)
plt.title('Training + Testing Data, values shown in µg/m3')
plt.savefig('PM2.5/Training+Testing_Data.png')
plt.show()

len(test_data)
```

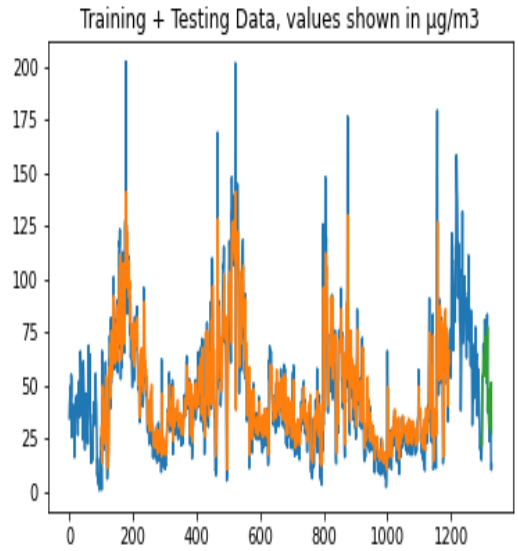
```
x_input=test_data[33:].reshape(1,-1)
x_input.shape

temp_input=list(x_input)
temp_input=temp_input[0].tolist()

temp_input

# demonstrate prediction for next 10 days
from numpy import array

lst_output=[]
n_steps=100
i=0
while(i<30):

    if(len(temp_input)>100):
        #print(temp_input)
        x_input=np.array(temp_input[1:])
        print("{} day input {}".format(i,x_input))
        x_input=x_input.reshape(1,-1)
        x_input = x_input.reshape((1, n_steps, 1))
        #print(x_input)
        yhat = model.predict(x_input, verbose=2)
        print("{} day output {}".format(i,yhat))
        temp_input.extend(yhat[0].tolist())
        temp_input=temp_input[1:]
        #print(temp_input)
        lst_output.extend(yhat.tolist())
        i=i+1
    else:
        x_input = x_input.reshape((1, n_steps,1))
        yhat = model.predict(x_input, verbose=2)
        print(yhat[0])
        temp_input.extend(yhat[0].tolist())
        print(len(temp_input))
        lst_output.extend(yhat.tolist())
        i=i+1


print(lst_output)

day_new=np.arange(1,101)
day_pred=np.arange(101,131)

len(df1)

plt.plot(day_new,scaler.inverse_transform(df1[1229:]))
plt.title('PM 2.5 Forecast along with previous data, values shown in µg/m3')
plt.plot(day_pred,scaler.inverse_transform(lst_output))
plt.savefig('PM2.5/Forecast_along_previous_data.png')

df3=df1.tolist()
df3.extend(lst_output)
```

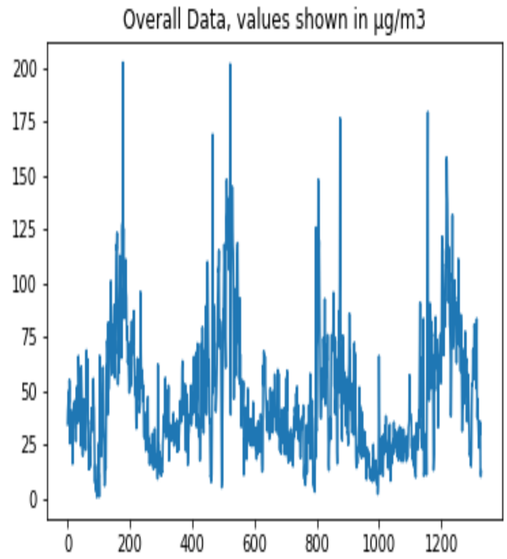
```
plt.plot(df3[1329:])

df3=scaler.inverse_transform(df3).tolist()

#1month forecast
plt.title('PM 2.5 1 Month Forecast, values shown in µg/m3')
plt.plot(df3[1329:])
plt.savefig('PM2.5/1 monthh forecast.png')
```

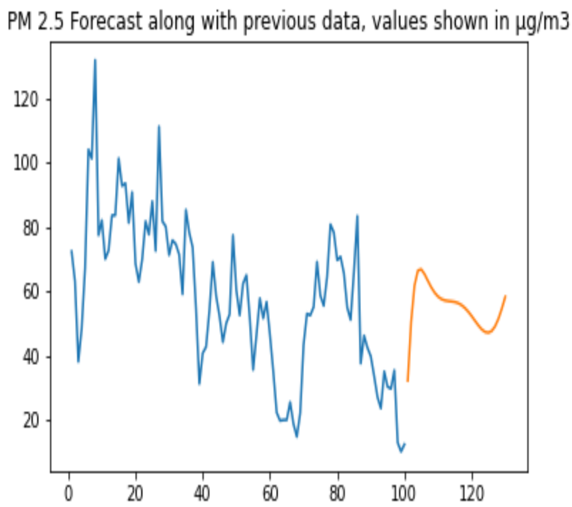
Chapter 04

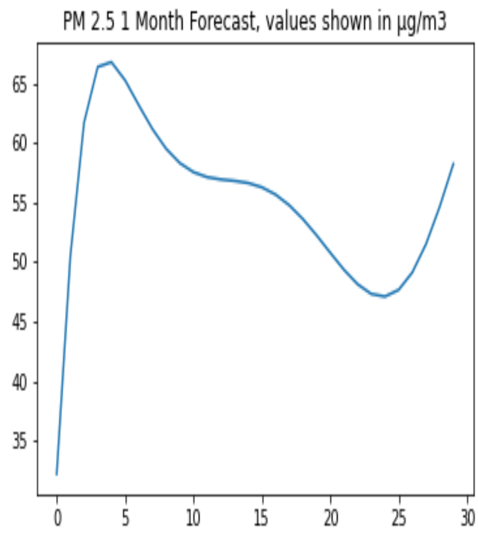
RESULTS AND GRAPHS

PM 2.5 **prediction probability** (April 18th to May 18th 2021)

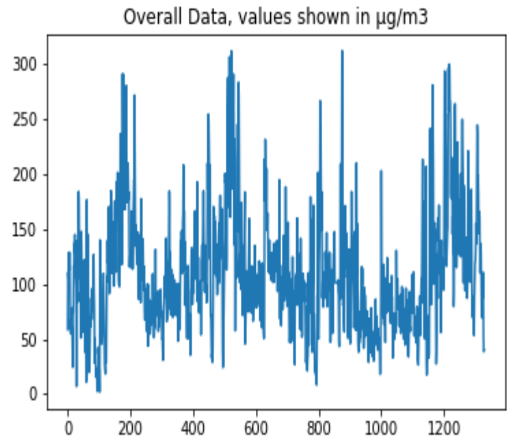


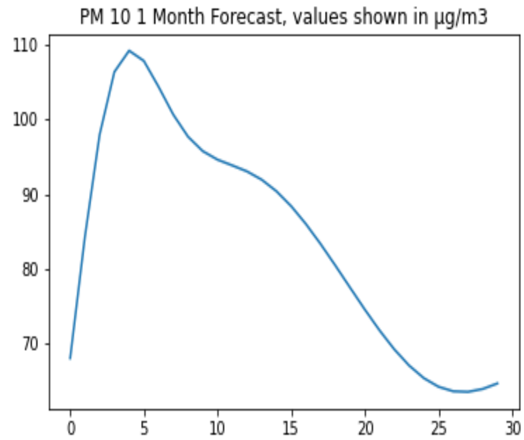


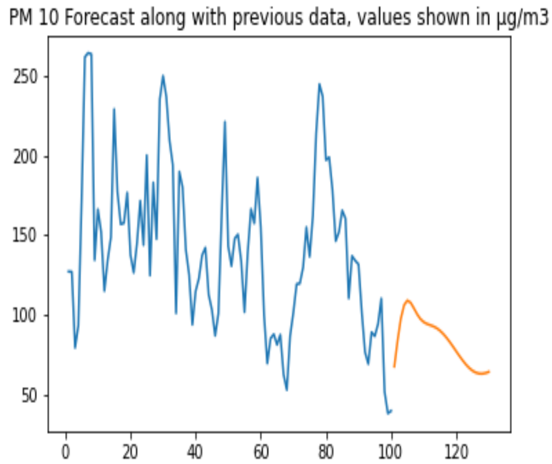


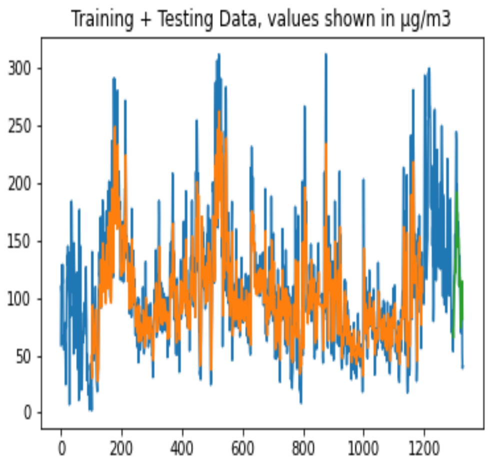


PM 10: prediction probability (April 18th to May 18th 2021)

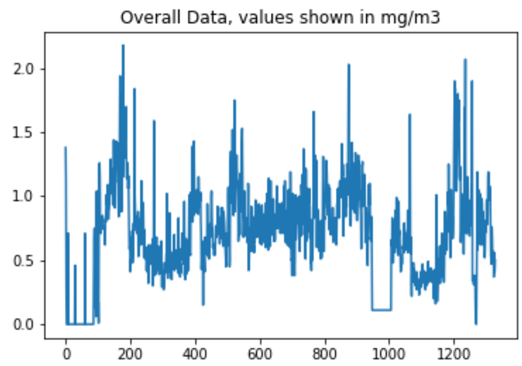


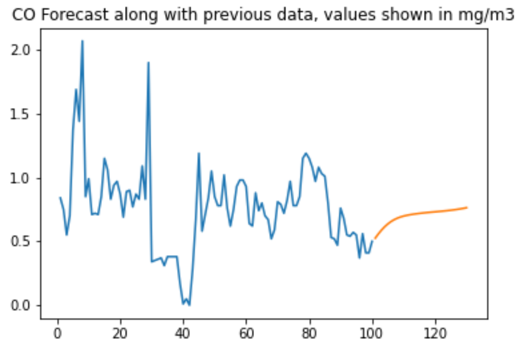


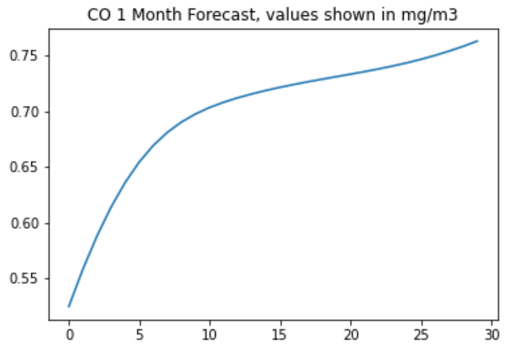


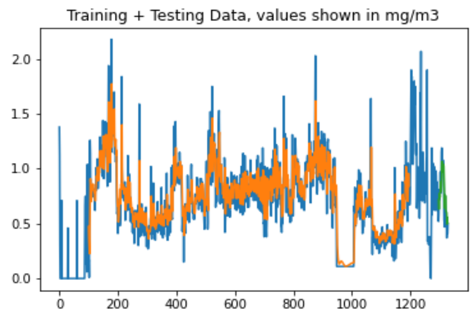


CO: prediction probability (April 18th to May 18th 2021)

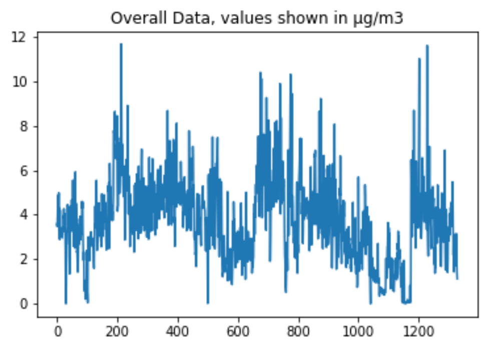


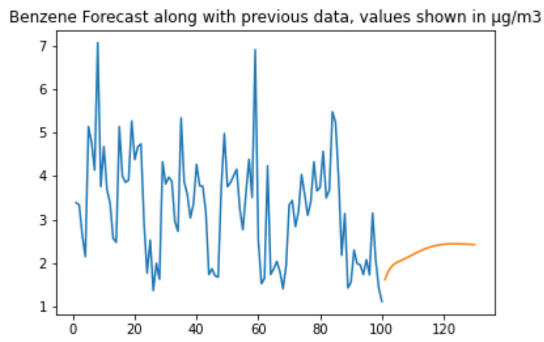


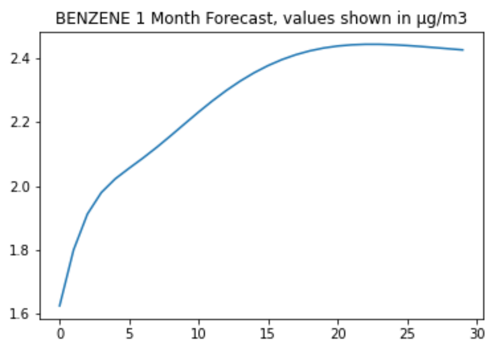


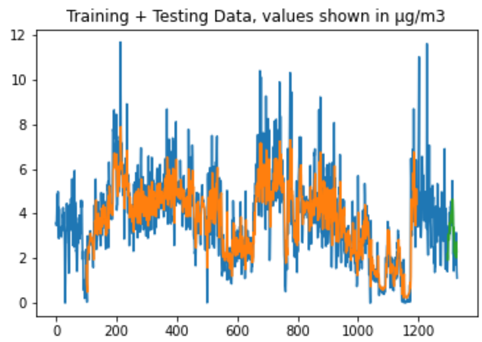


BENZENE: prediction probability (April 18th to May 18th 2021)

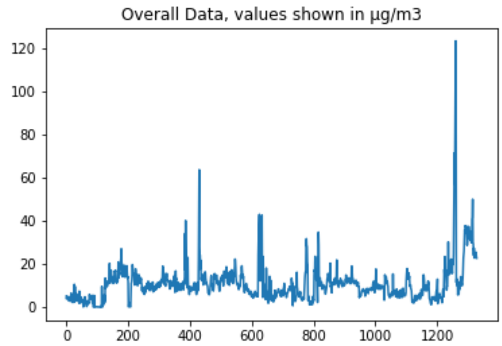


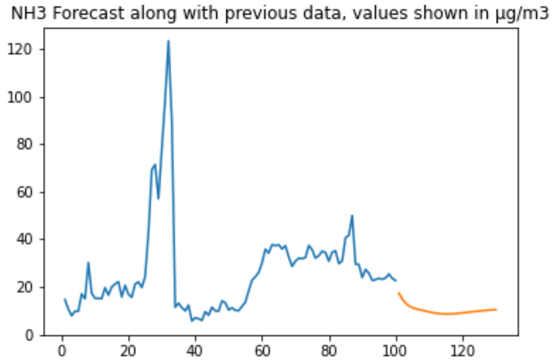


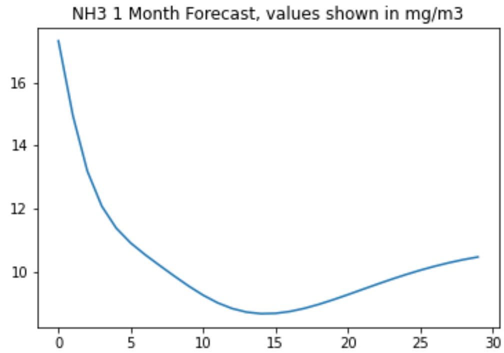


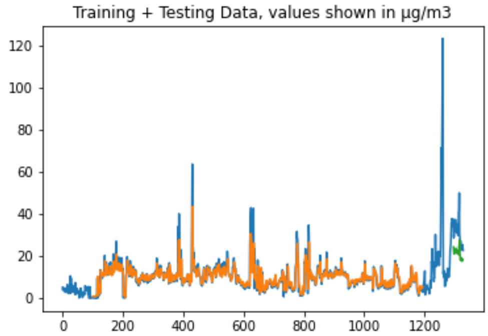


NH3: prediction probability (April 18th to May 18th 2021)

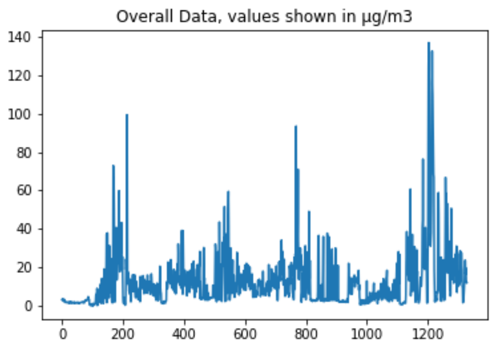


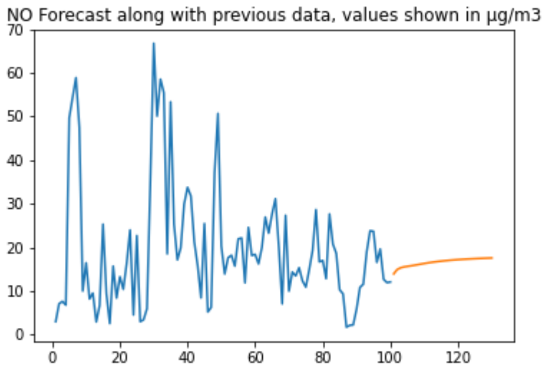


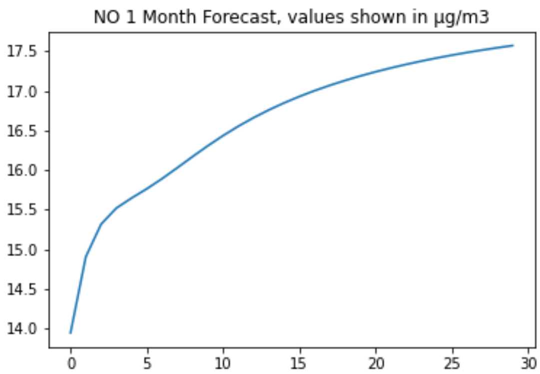


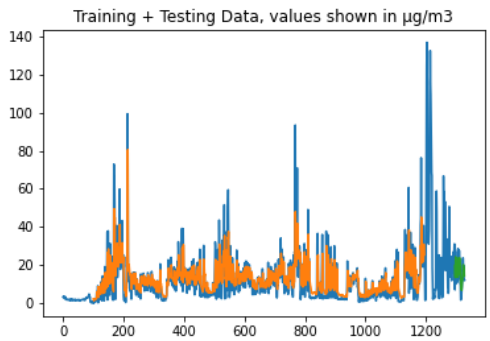


NO: prediction probability (April 18th to May 18th 2021)

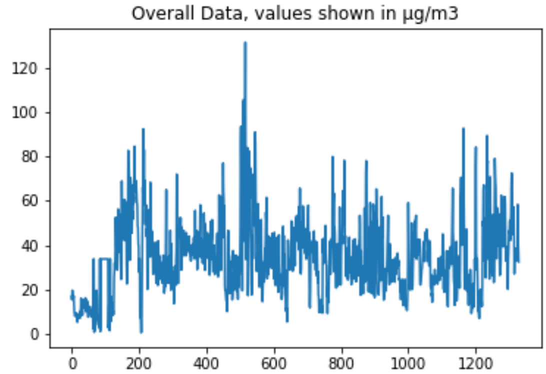


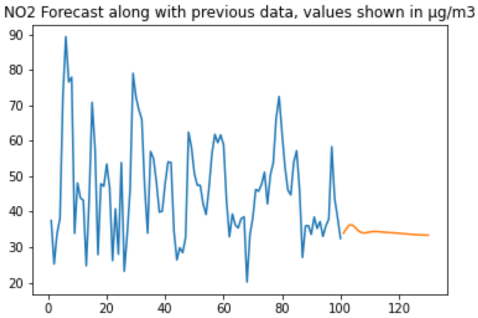


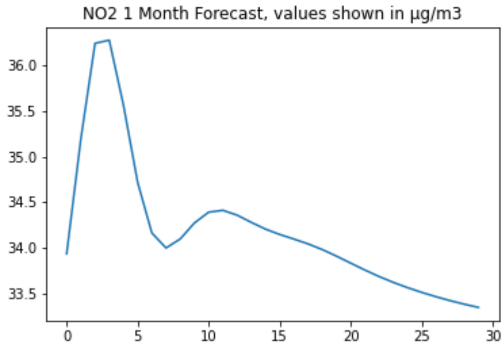




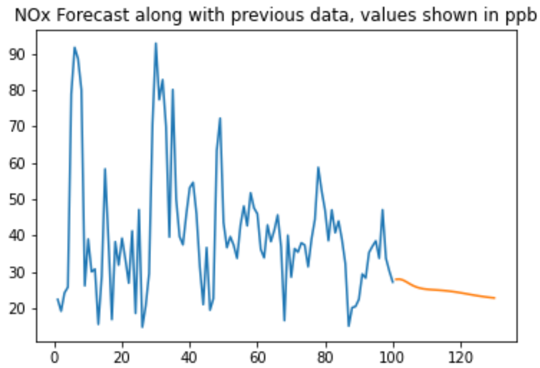
NO2: prediction probability (April 18th to May 18th 2021)

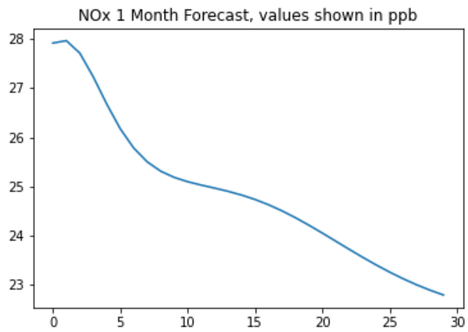


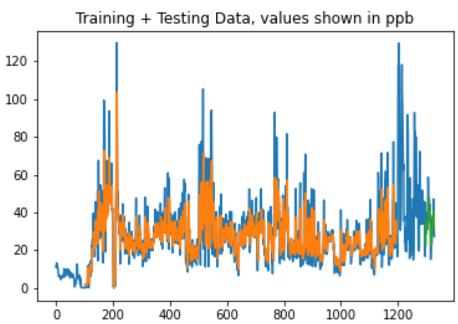


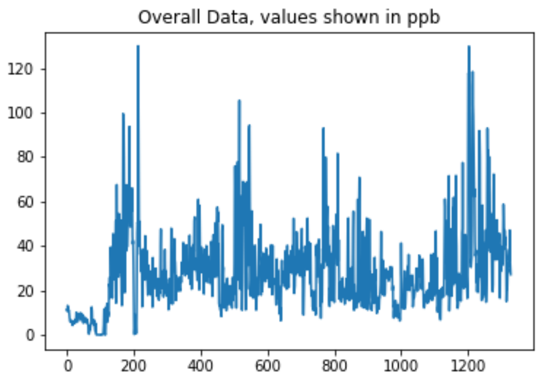


NOX: (prediction probability April 18th to May 18th 2021)

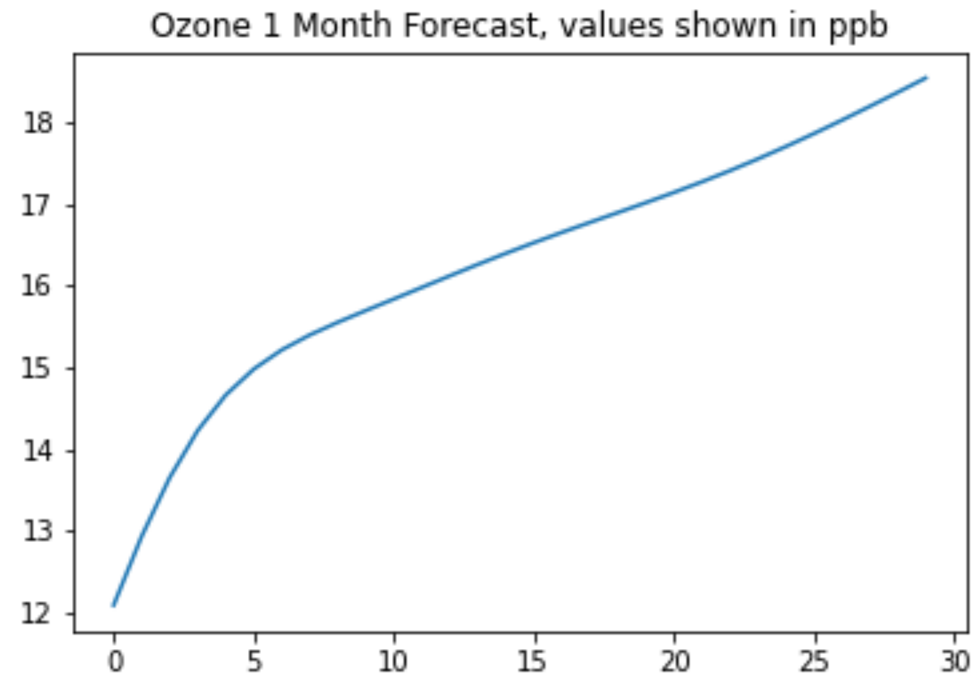


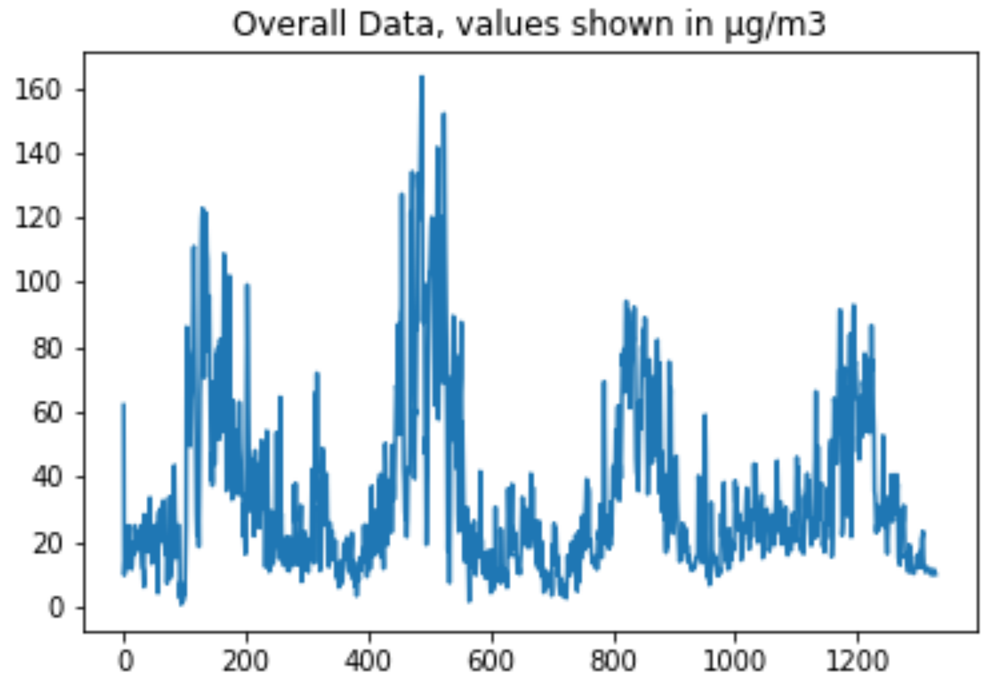


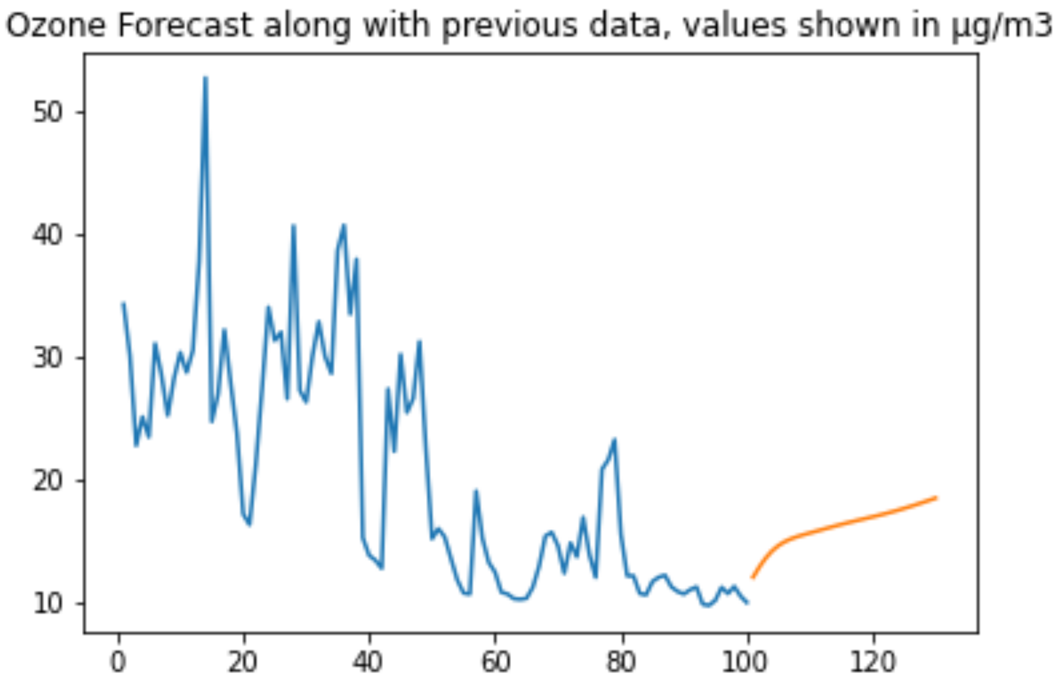




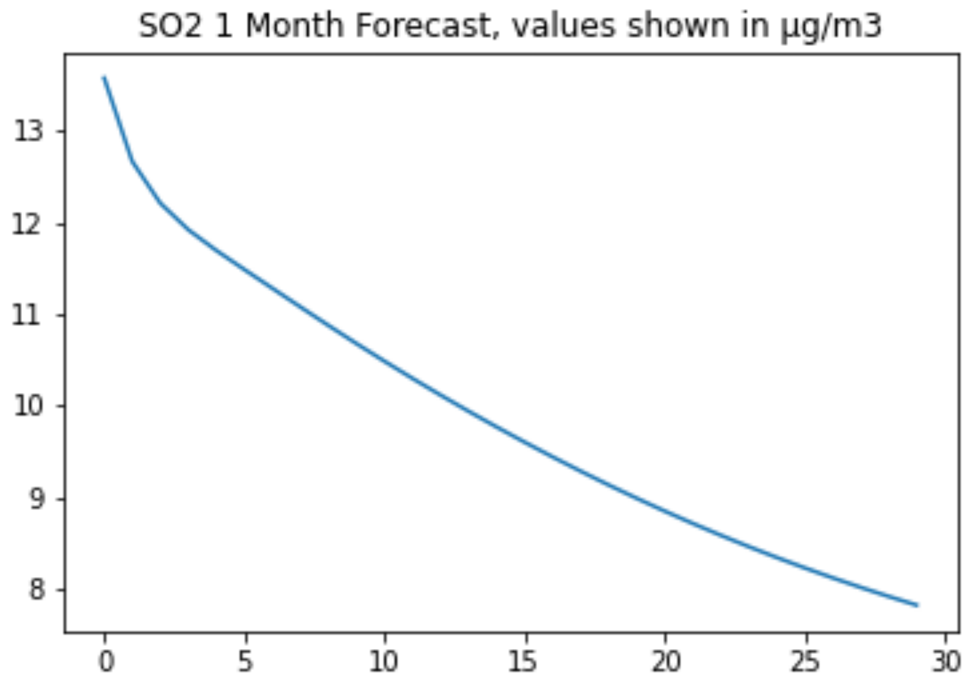
OZONE: (**prediction probability** April 18th to May 18th 2021)

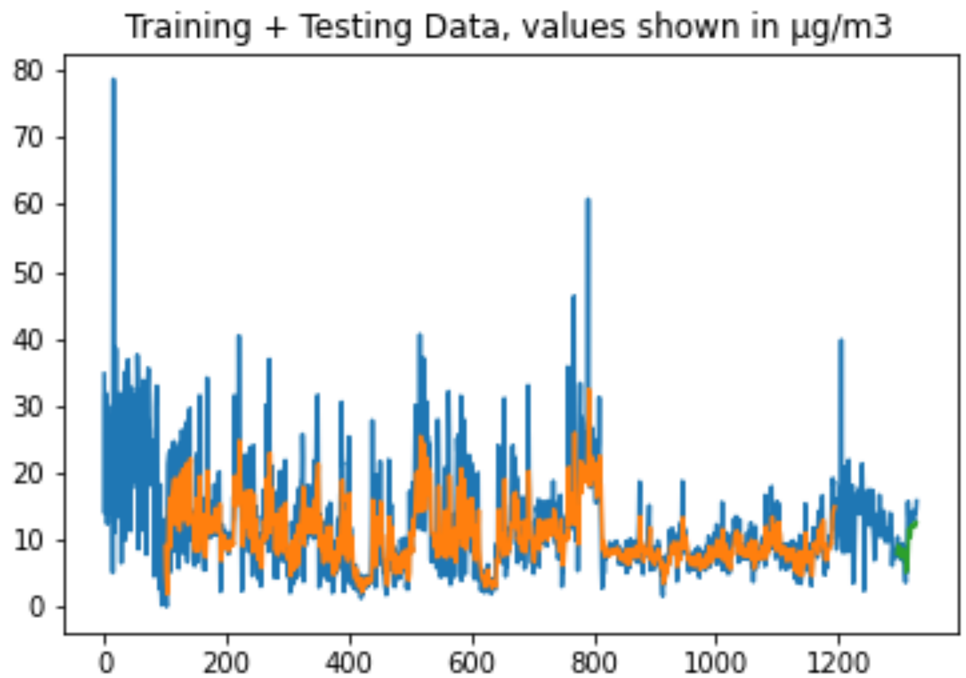


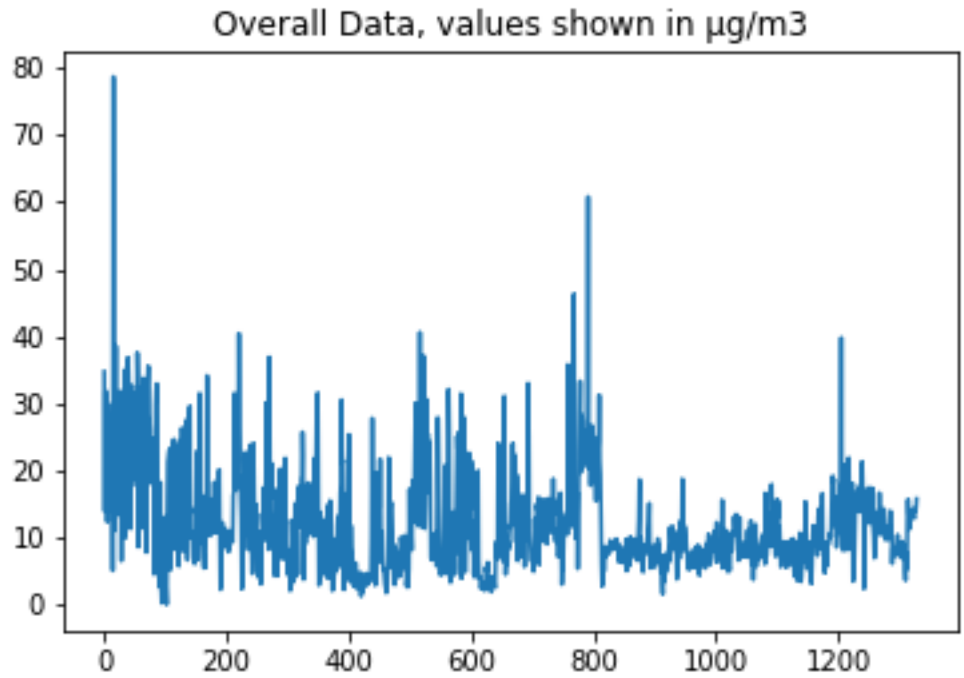

Overall Data, values shown in µg/m3
0
20
40
60
80
100
120
140
160
0
200
400
600
800
1000
1200


Ozone Forecast along with previous data, values shown in µg/m3
10
20
30
40
50
0
20
40
60
80
100
120

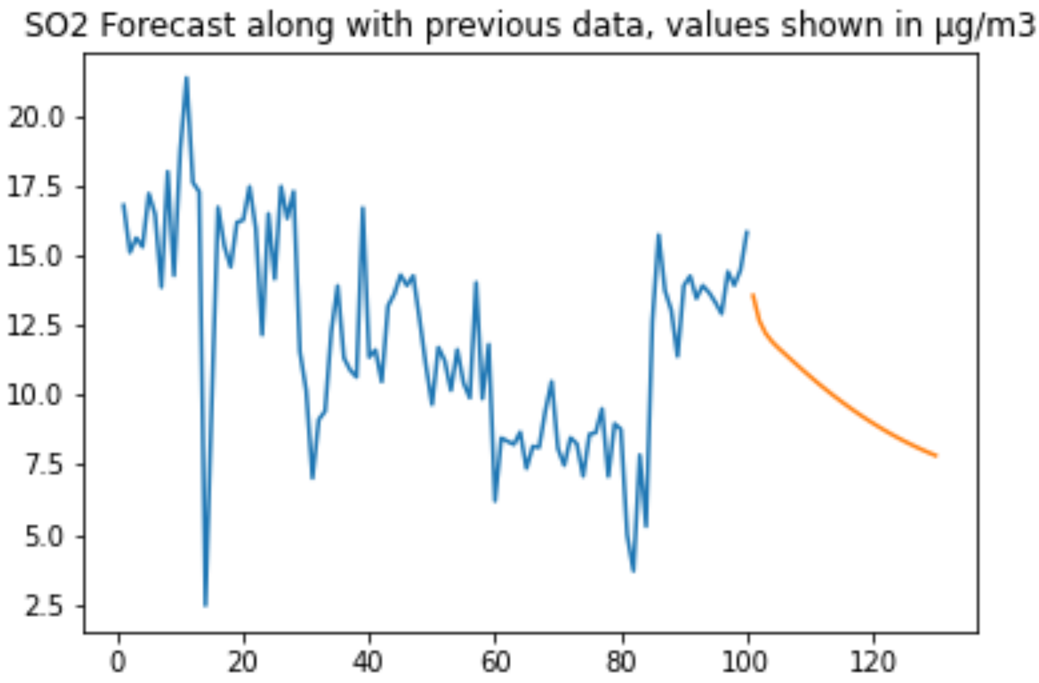
SO2: prediction probability (April 18th to May 18th 2021)





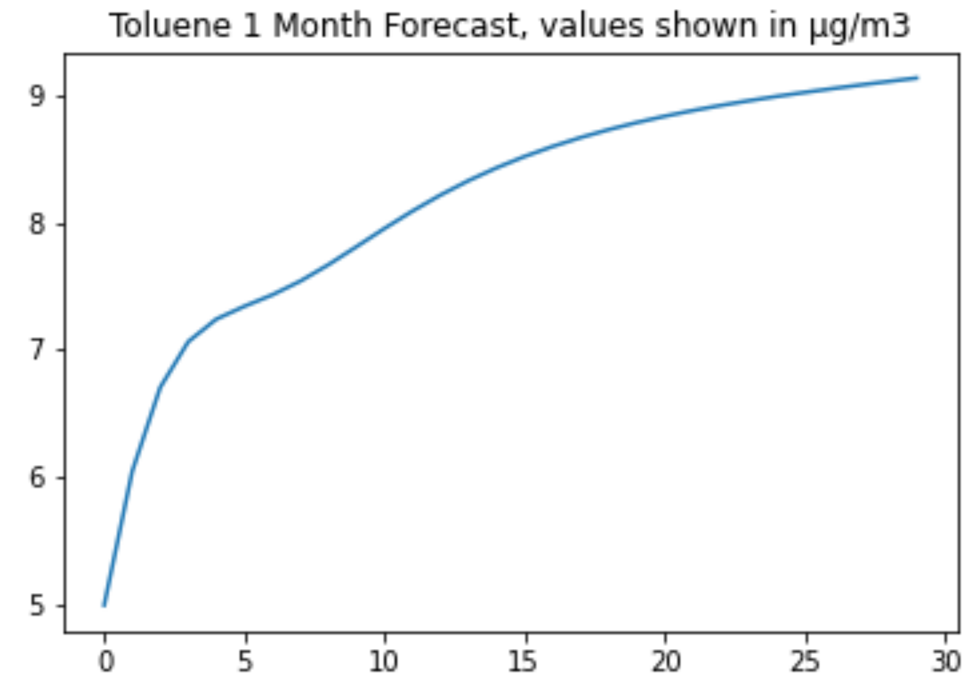


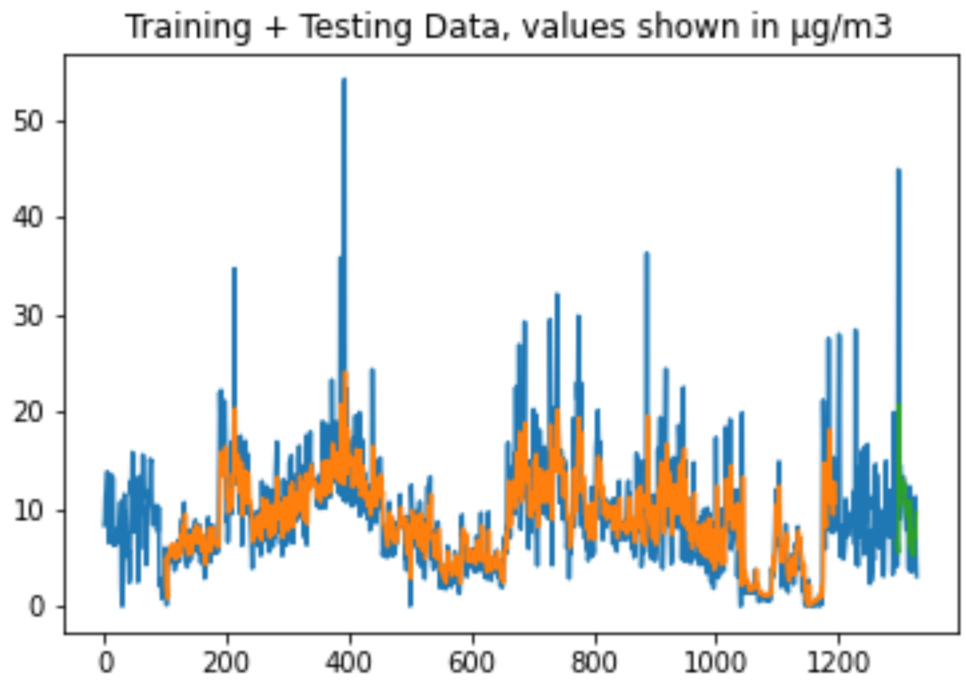
Overall Data, values shown in µg/m3
0
10
20
30
40
50
60
70
80
0
200
400
600
800
1000
1200



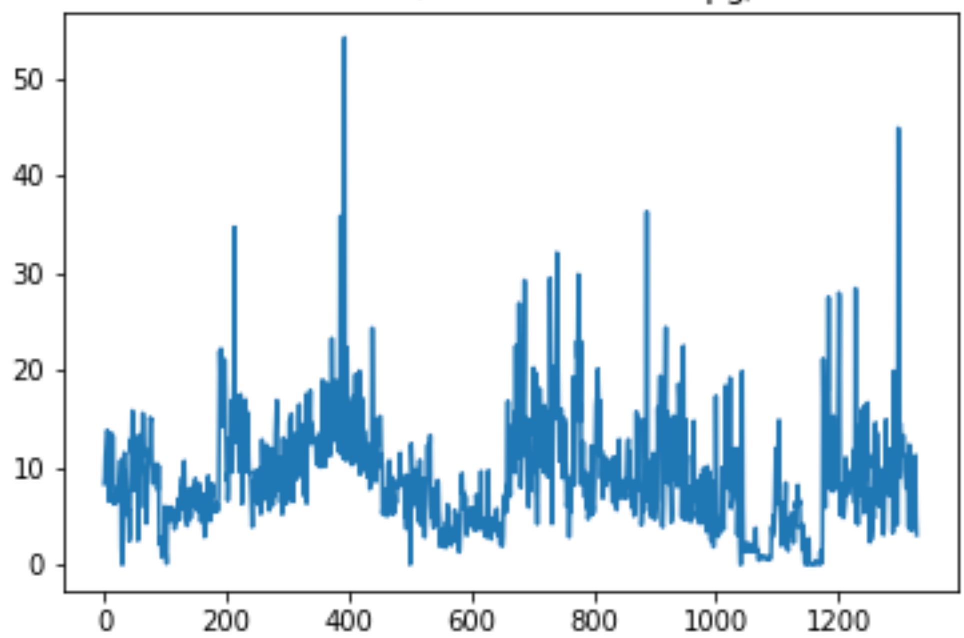
SO2 Forecast along with previous data, values shown in µg/m3
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
20
40
60
80
100
120

TOULENE: prediction probability (April 18th to May 18th 2021)

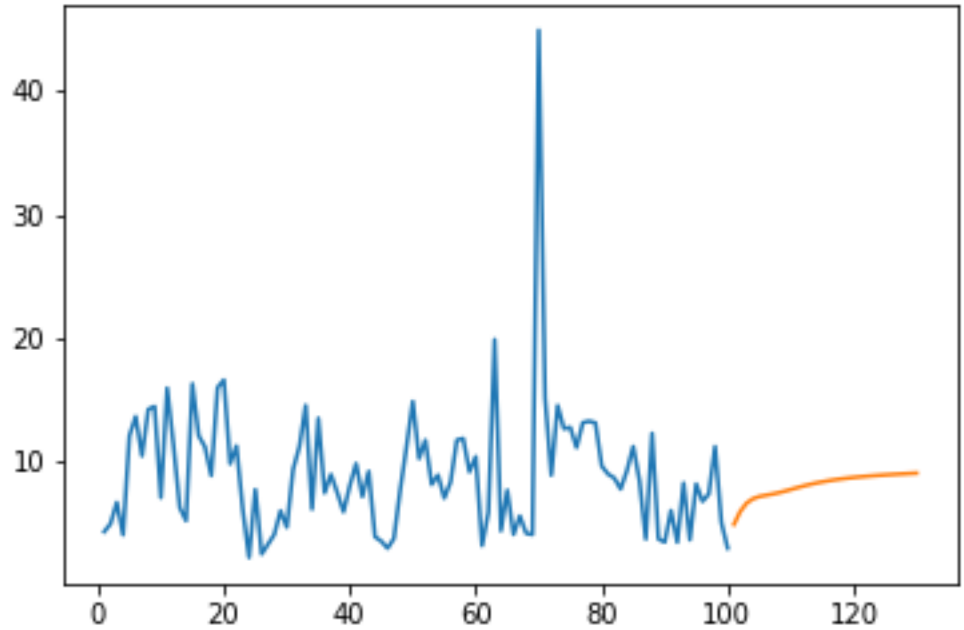




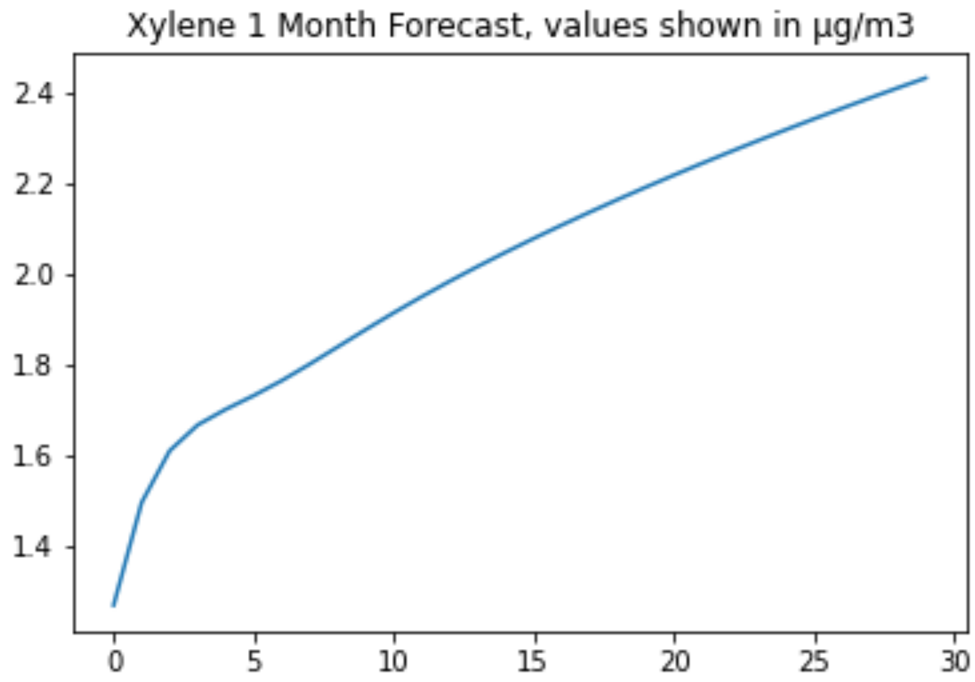
Overall Data, values shown in µg/m3

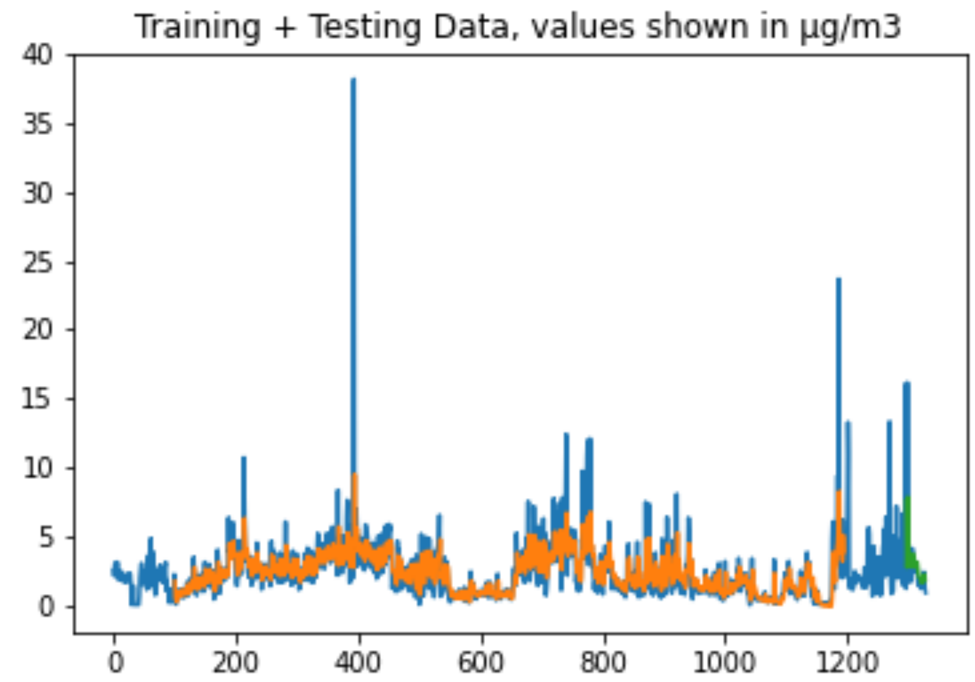


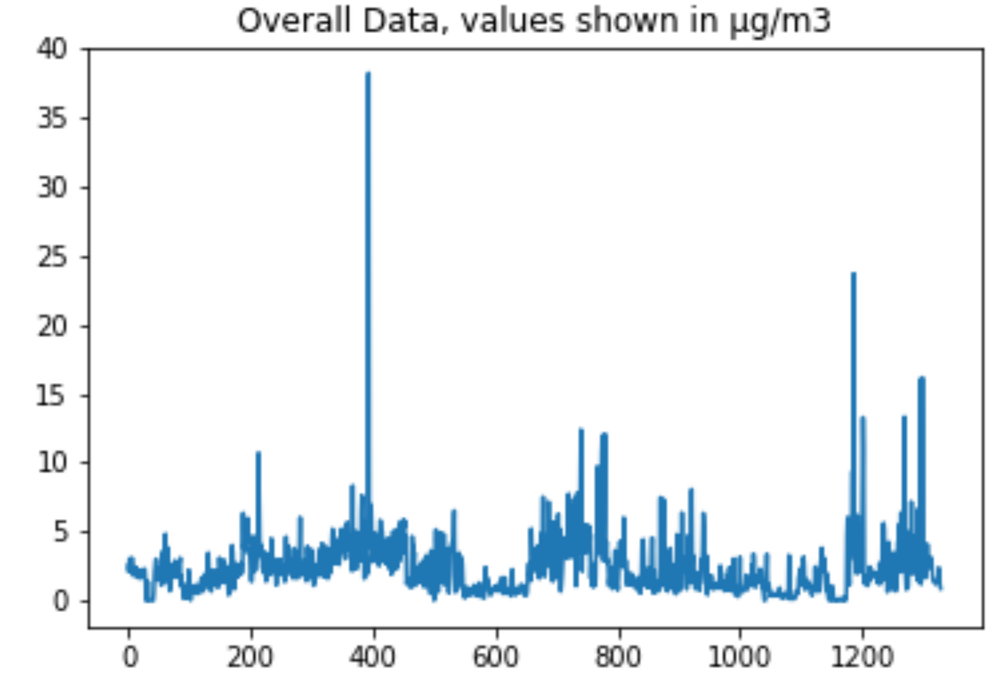
Toluene Forecast along with previous data, values shown in µg/m3

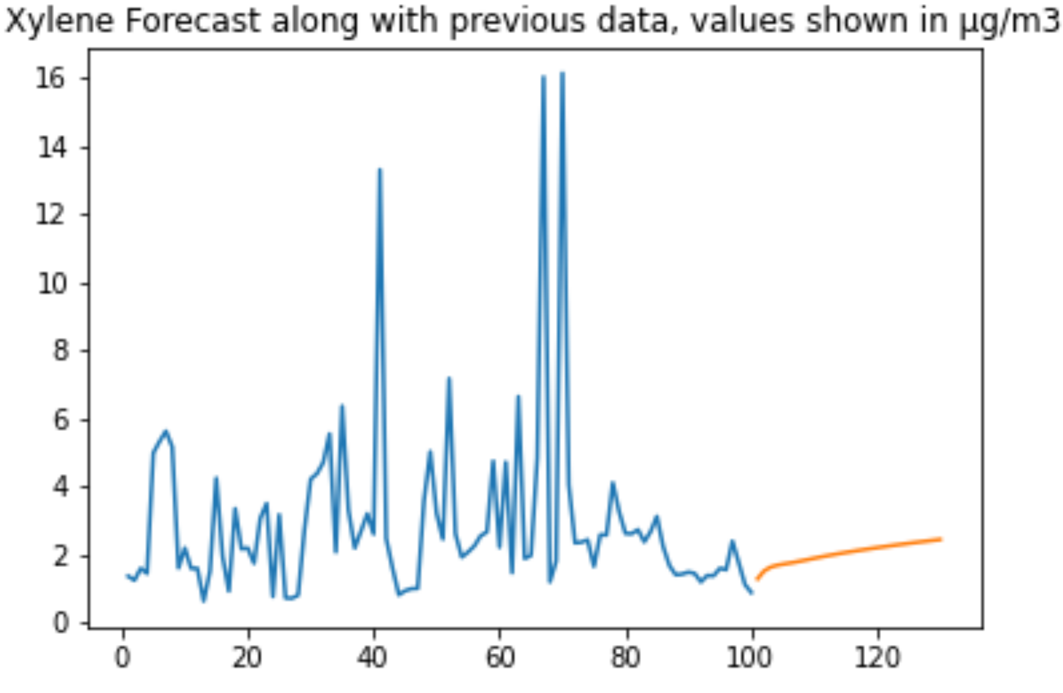


XYLENE: prediction probability (April 18th to May 18th 2021)

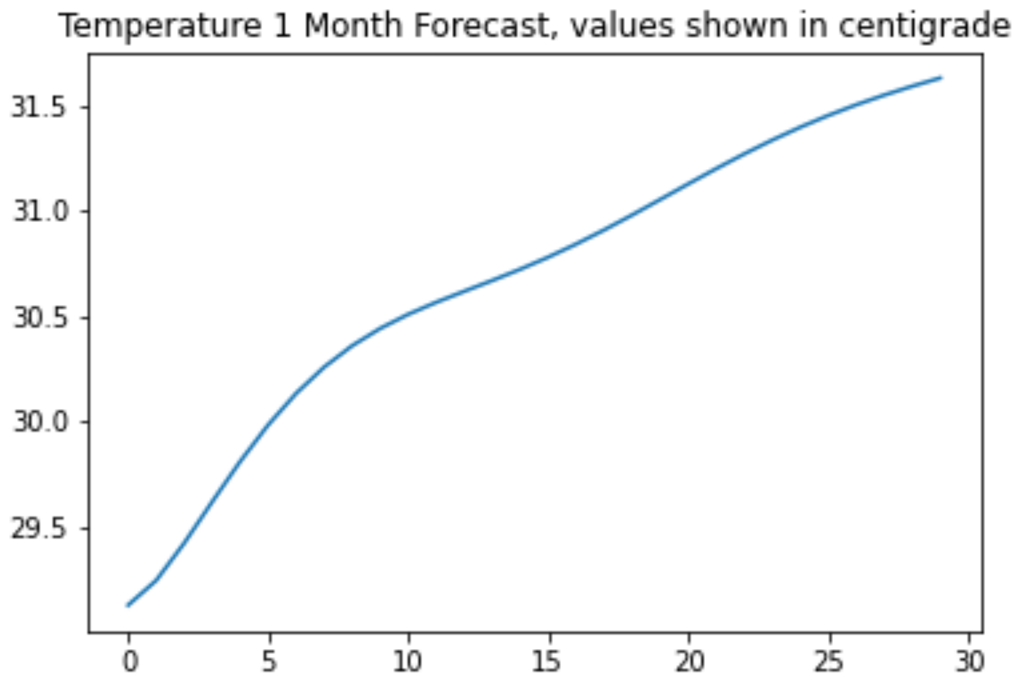


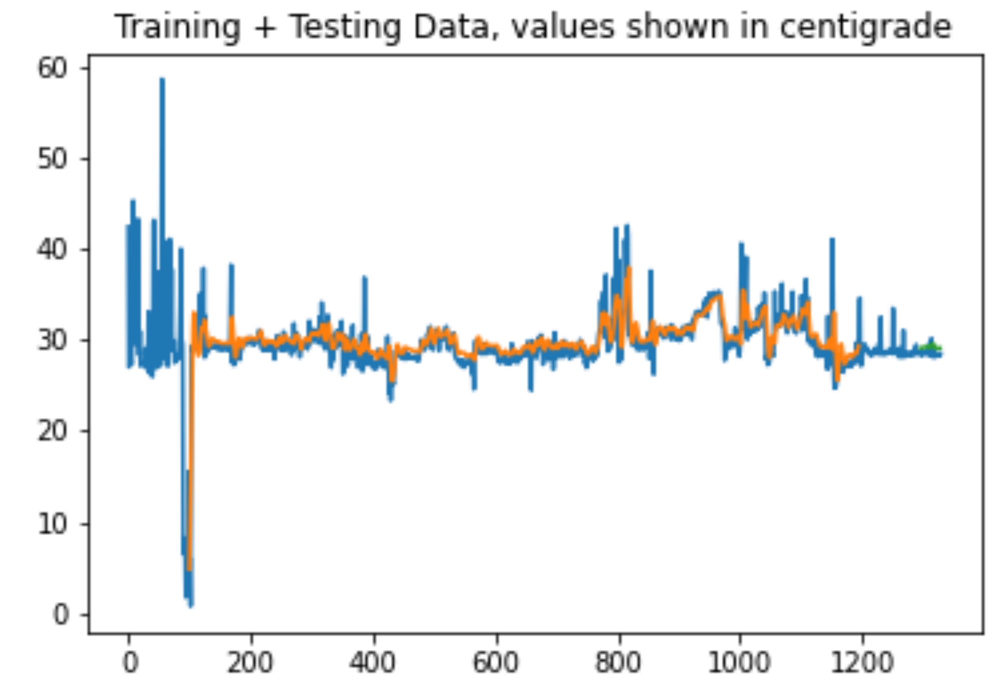


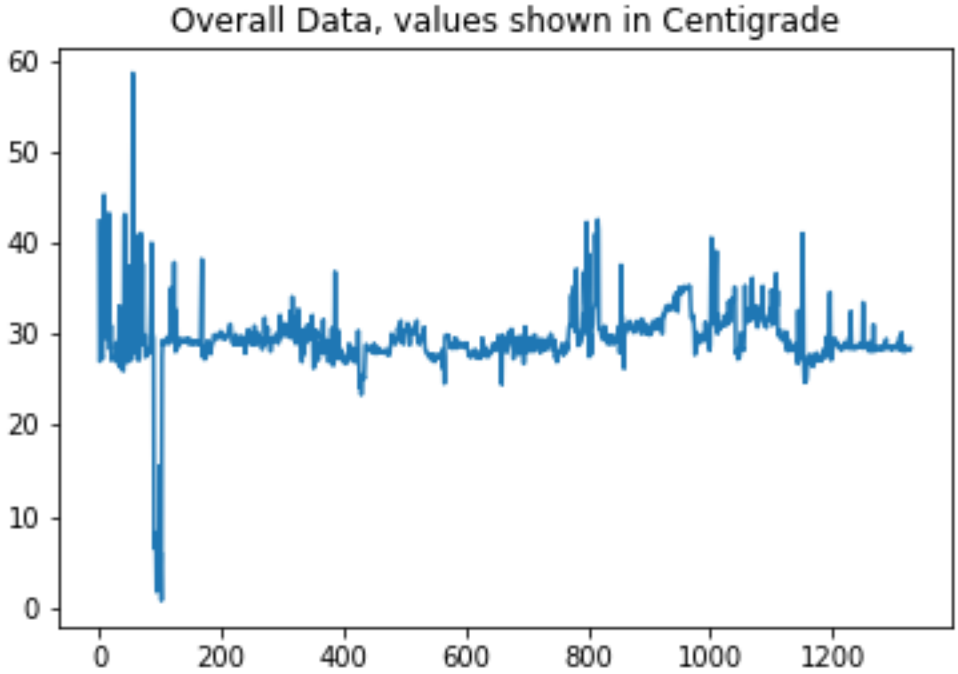

Overall Data, values shown in µg/m3
40
35
30
25
20
15
10
5
0
0
200
400
600
800
1000
1200

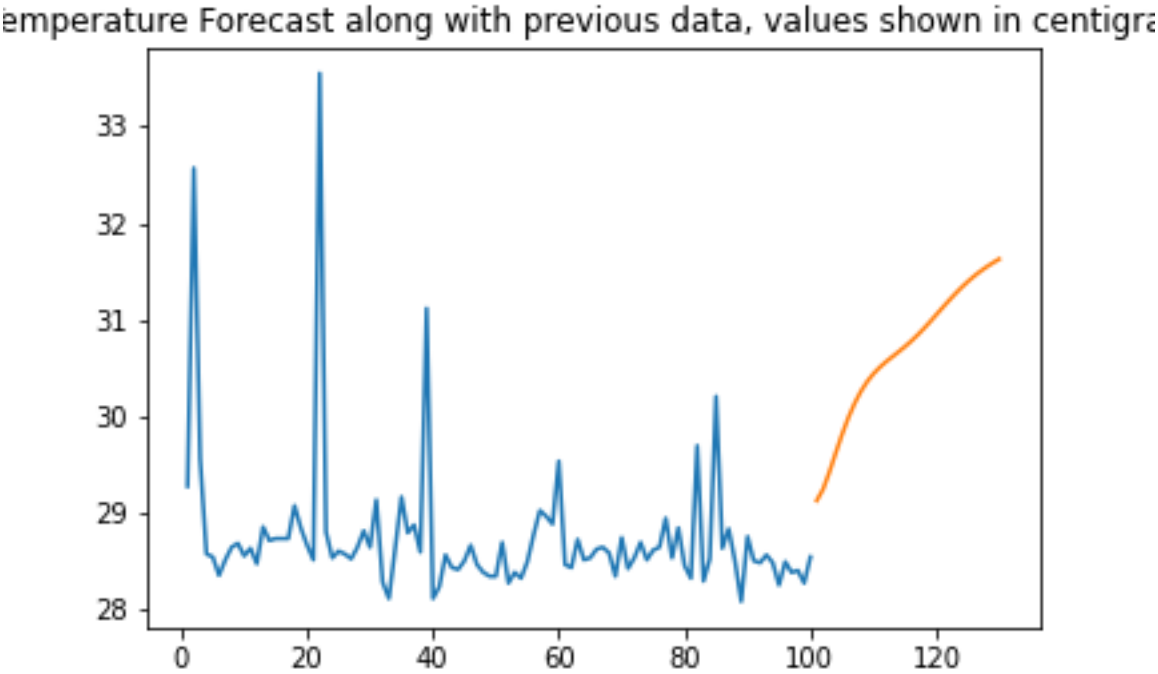

Xylene Forecast along with previous data, values shown in µg/m3
16
14
12
10
8
6
4
2
0
0
20
40
60
80
100
120

TEMPERATURE: prediction probability (April 18th to May 18th 2021)

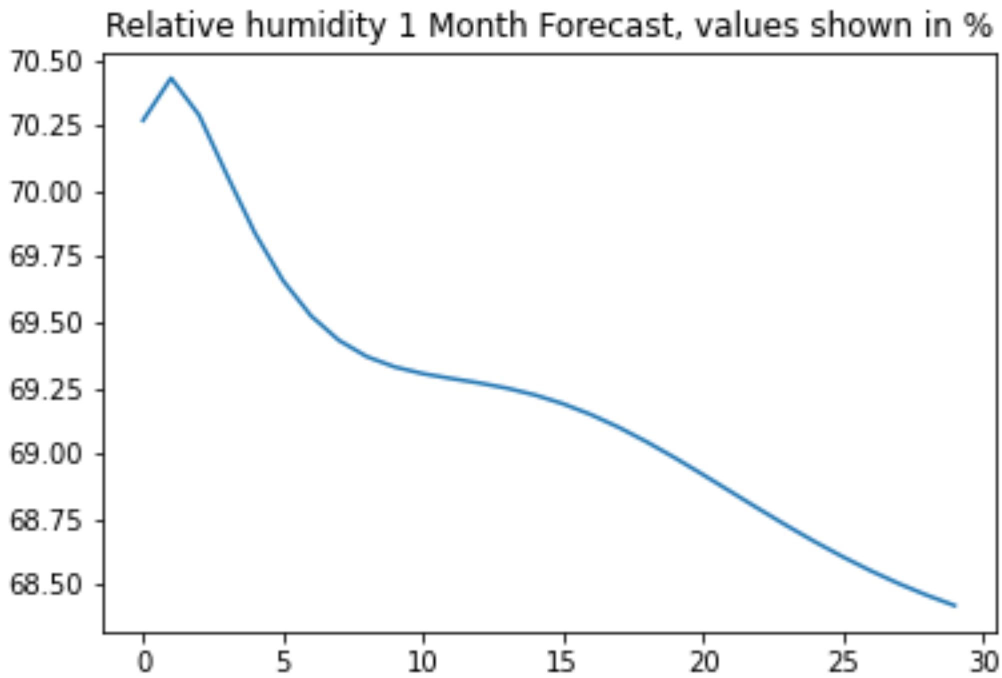


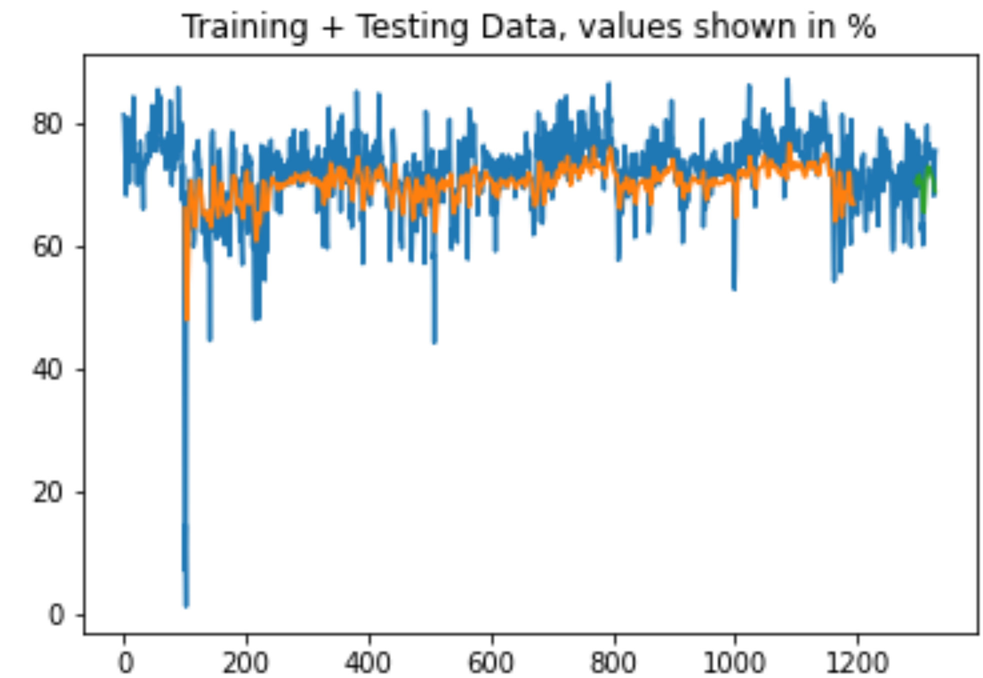


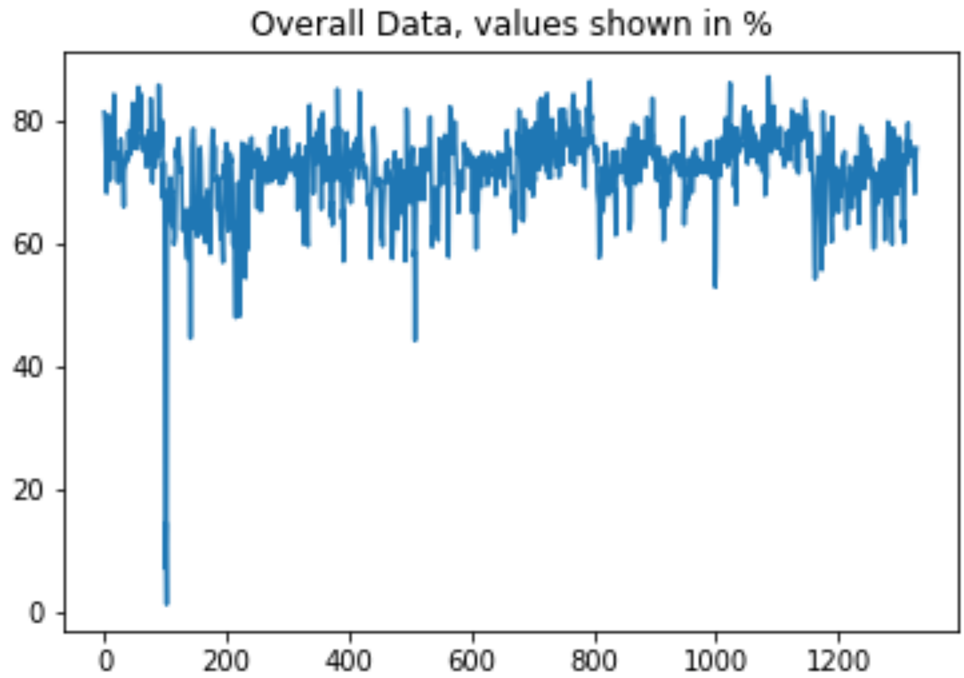

Overall Data, values shown in Centigrade
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200

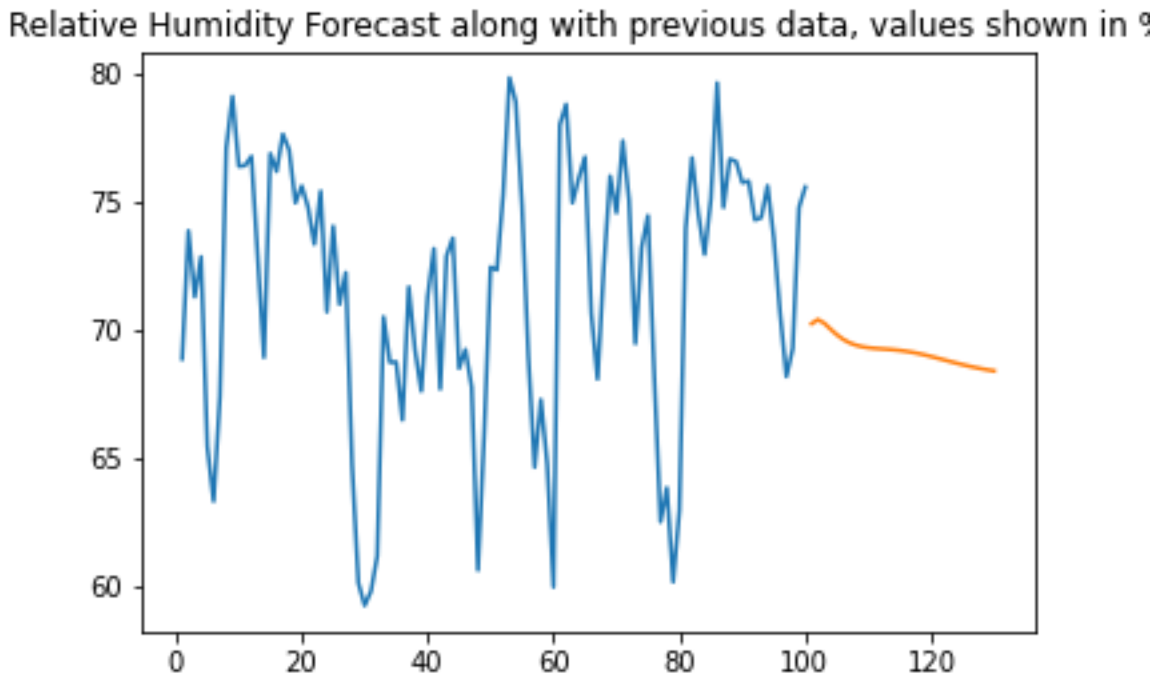

emperature Forecast along with previous data, values shown in centigra
33
32
31
30
29
28
0
20
40
60
80
100
120

RELATIVE HUMIDITY: prediction probability (April 18th to May 18th 2021)

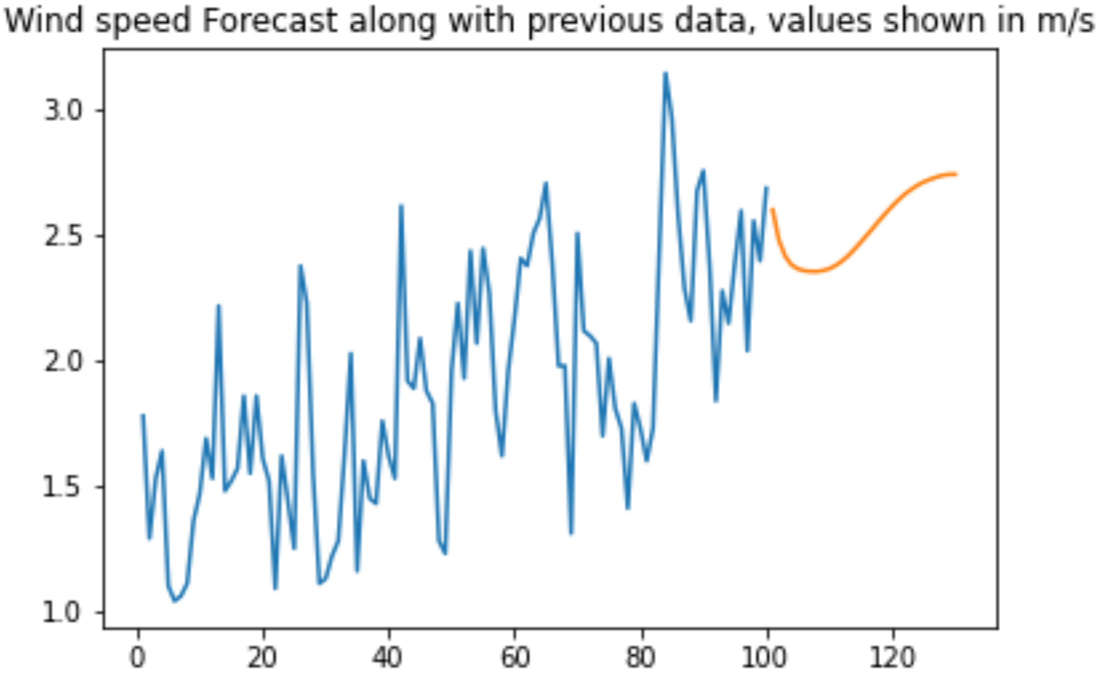


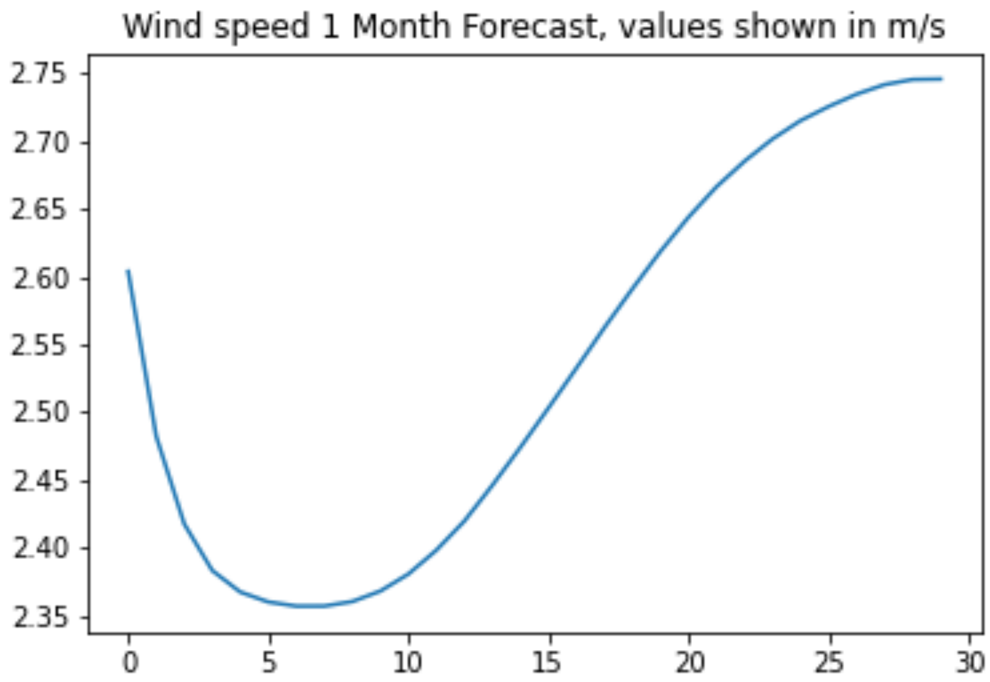


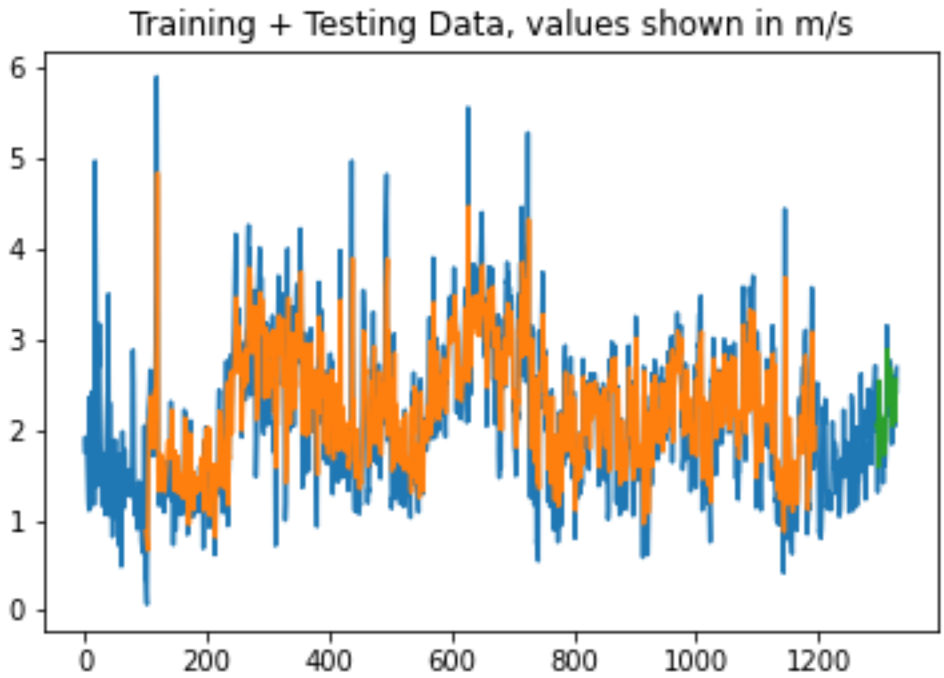

Overall Data, values shown in %
0
20
40
60
80
0
200
400
600
800
1000
1200

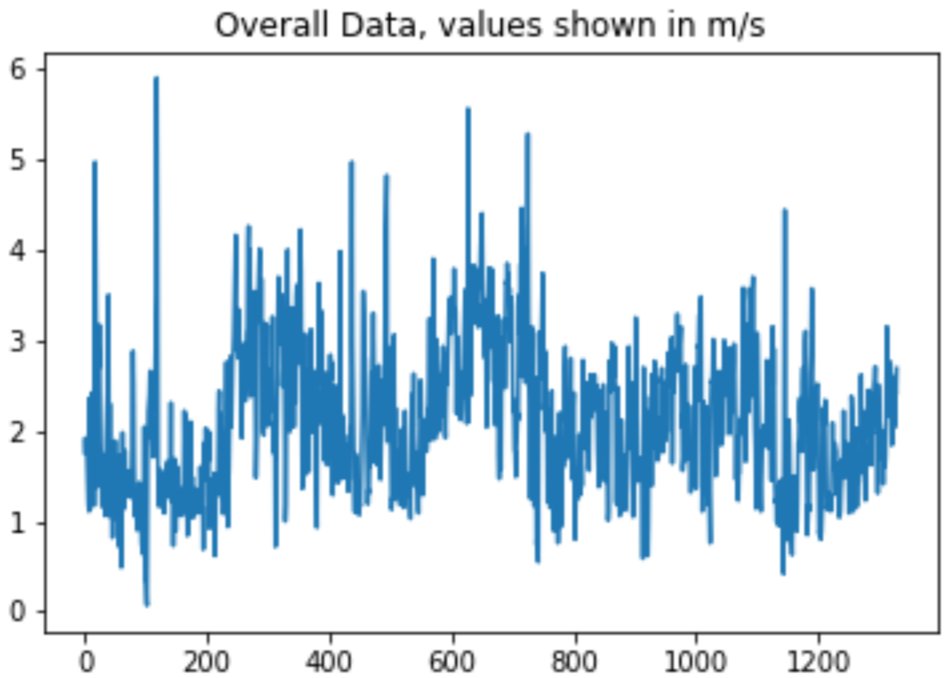

Relative Humidity Forecast along with previous data, values shown in %
60
65
70
75
80
0
20
40
60
80
100
120

WIND SPEED: prediction probability (April 18th to May 18th 2021)






Training + Testing Data, values shown in m/s
0
1
2
3
4
5
6
0
200
400
600
800
1000
1200


Overall Data, values shown in m/s
0
1
2
3
4
5
6
0
200
400
600
800
1000
1200

Visualisation link:

Dashboard for entire 1 year may 2021- April 2022

https://tabsoft.co/337exRs

CHAPTER 05

CONCLUSION

Air is the first necessity of our existence so monitoring and keeping track of what we breathe is important. Machine learning is a substitute of artificial intelligence which tries to emulate the thinking of a human mind. So we tried to use various ML models and zeroed in on LSTM because it gave better results for this specific use case we can see that the ML Model has performed good in terms of prediction and forecasting. The tableau dashboard is public and can be integrated with any website/app to make it more accessible. In the future, we aim to develop a standalone app containing all the above features and also analyse Pan India Data. We think this idea if provided time and funding can be a great asset to the citizens of Visakhapatnam. We hope this project is a right step towards making citizens Air Quality Aware and conscious of the impact of industrial and vehicle pollution.

References:

1. Kingsy Grace. R I , Manimegalai. R 2 , Geetha Devasena. M.SI , Rajathi. Si , Usha. KI , Raabiathul Baseria. N (2016 IEEE Region 10 Conference (TENCON))
2. Sangeetha, T. Amudha 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)\
3. (B. Ojeda-Magaña1, M. G. Cortina-Januchs2, J. M. Barrón-Adame3, J. Quintanilla-Domínguez2, W. Hernandez 4, A. Vega-Corona3, R. Ruelas1 and D. Andina) 2010 IEEE
4. Mihaela Oprea, Hai-Ying Liu, Chen Chen, Qiang Liu, Pei-Wen Liu, Yao-Sen Huang, Wei-Qing Li, Hao Luo (2017 International Conference on Network and Information Systems for Computers (ICNISC))
5. Shahid Ali, Sreenivas Sremath Tirumala, Abdolhossein Sarrafzadeh (2014 IEEE)
6. Utkarsh Kulshreshta, Surya Durbha
   IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium

7. Ntombikayise Koyana, Freestate, South Africa ; Elisha Didam Markus; Adnan M. Abu-Mahfouz
2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)
8. G Spandana, R Shanmughasundram
2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)
9. Bartolomeo Montrucchio; Edoardo Giusto; Mohammad Ghazi Vakili; Stefano Quer
IEEE Transactions on Vehicular Technology
10. Ranu Gadia, Shivania Sudhir, Kumar S, harmab Tuhin, Kumar Mandalb
Chemosphere Volume 221, April 2019, Pages 583-596