TRAIN AND DEPLOY A MACHINE LEARNING MODEL IN AWS CLOUD – PREDICTION OF AIR QUALITY

Team Number	10	
Team Names	Kumar Rajesh. K	Naga Srinivas. B
	Sai Swetha. K	Manju. T
Mentor Name	Rajesh / Shaifu	
Team SPOC Number	8688852062	
Field: Machine Learning, HTML, Flask, Cloud.		

- Introduction:

"**AIR - THE PRIMARY FUEL OF LIFE**".

The necessity of healthy air has always been of great importance. As air is vital for all living beings on earth, it is our responsibility to keep the air clear. Forecasting air quality is useful for preventing and reducing pollution. This research focuses on a multi-time forecasting model that was applied to data from Italian's air quality.

The Air Quality Index (AQI) is a statistic for assessing the quality of the air in our immediate surroundings. It measures how air pollution can affect an individual's health within a specified period of time.

The common air pollutants are Particulate Matter (PM2.5, PM10), Nitrogen Dioxide (NO2) Carbon Monoxide (CO), Sulphur Dioxide (SO2) and Ozone (O3). The datasets related to these components are collected through the IOT sensors.

HTML and Flask technologies are included here, HTML is for the user interface and Flask is for running the machine learning code successfully

In this thesis, we explore air quality prediction with a particular focus on applied machine learning. Promising machine learning approaches from the literature are implemented and evaluated for performance.

The goal is to use state-of-the-art research results with a special focus on machine learning methods about air quality prediction, to evaluate and apply ideas and algorithms from these studies in the context of time series prediction. The ultimate goal of air quality prediction is to enable national and global decision-makers, communities, and individuals to proactively take measures to reduce the health hazards caused by air pollution.

- Scope of the project:

As mentioned previously that the air is fuel of life it has an unforeseen impact on past present future and forever on our lives and it holds the top priority on every small aspect.so the condition of the air i.e, the quality of the air need to be checked continuously so as to calculate or predict further coming scenarios.

The condition of the air in future can be predicted by Machine Learning Algorithm. And this future prediction is based on the previously collected data. By cosidering the predicted value and comparing it with the past values the state of air's condition can be analysed and can progress with further obligatory actions.

Based on the predicted value, the improvements should be made in order to maintain good health of people like decreasing the vehicle traffic, maintaining greenery, and to decrease plastic wastes. This project mainly works for some years very effectively. In order to increase life cycle of the project it should be updated with the dataset part. Hardly, no other columns will be added. However, changes should be made in the data part to increase the lifecycle of the project.

Steps need to be followed:
Some of the steps need to be followed to determine the life cycle is

- Identifying Problems of Air Pollution
- Collection of data sets from time series
- Applying Machine Learning

- Splitting the datasets
- Training the Machine Learning Algorithm
- Obtaining the predicted values from algorithm
- Uploading the trained algorithm in AWS cloud
- Analysing the Air Quality Conditions and taking necessary measures

- Strategy to accomplish the project

Machine Learning:

Machine learning is a branch which is set out of artificial intelligence. Its goal is to enable the computer to learn by itself without being explicitly programmed the rules. A machine learning algorithm can identify and learn underlying patterns in observed data to model and predict the world.

There are three kinds of machine learning techniques: reinforcement learning, unsupervised learning, and supervised learning. In reinforcement learning, the algorithm receives feedback based on performance as it navigates its problem space. Tasks such as playing a game or driving a car are examples where reinforcement learning is suitable. Unsupervised learning is an approach that learns from data that is unlabeled or classified. Instead of responding to feedback as in reinforcement learning, unsupervised learning identifies shared attributes and characteristics from the data.

Unsupervised learning algorithms include association problems, which tries to describe parts of the data, and clustering problems, that seek to identify natural groupings. In supervised learning, the algorithm attempts to learn from informative examples of labeled data. Such algorithms can be described as a data-driven approach, where historical data is used for predictions of the future. Air quality prediction is often solved with supervised methods, as time series can convert to labeled pairs of input and output, where the output target is the ground truth of the next value in the data sequence. The machine learning models presented in this work are of the kind supervised learning.

Cloud Computing:

In simple terms, cloud computing is a range of services delivered over the internet, or “the cloud.” It means using remote servers to store and access data instead of relying on local hard drives and private datacenters.

Before cloud computing existed, organizations had to purchase and maintain their own servers to meet business needs. This required buying enough server space to reduce the risk of downtime and outages, and to accommodate peak traffic volume. As a result, large amounts of server space went unused for much of the time. Today’s cloud service providers allow companies to reduce the need for onsite servers, maintenance personnel, and other costly IT resources.

HTML:

HTML is the World Wide Web's core markup language. Originally, HTML was primarily designed as a language for semantically describing scientific documents. Its general design, however, has enabled it to be adapted, over the subsequent years, to describe a number of other types of documents and even applications.

The scope of this specification is not to describe an entire operating system. In particular, hardware configuration software, image manipulation tools, and applications that users would be expected to use with high-end workstations on a daily basis are out of scope. In terms of applications, this specification is targeted specifically at applications that would be expected to be used by users on an occasional basis, or regularly but from disparate locations, with low CPU requirements. Examples of such applications include online purchasing systems, searching systems, games (especially multiplayer online games), public telephone books or address books, communications software (email clients, instant messaging clients, discussion software), document editing software, etc.

Flask:

Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Pocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine. Both are Pocco projects. Unlike the Django framework, Flask is very Pythonic. It’s easy to get started with Flask, because it doesn’t have a huge learning curve

It’s a microframework, but that doesn’t mean your whole app should be inside one single Python file. You can and should use many files for larger programs, to handle complexity.

Micro means that the Flask framework is simple but extensible. You may all the decisions: which database to use, do you want an ORM etc, Flask doesn't decide for you. Flask is one of the most popular web frameworks, meaning it's up-to-date and modern. You can easily extend it's functionality. You can scale it up for complex applications.

By considering Italian city as the basis for finding whether the particular city is polluted or not, the dataset thus collected by placing five different sensors at ground level. Used sensors are namely carbon monoxide sensor where the values acts as target variable and remaining sensors like sulpher dioxide, benzene and ozone and they acts as independent variables. Hourly responses of these sensors are taken for one year to prepare dataset for acuurate prediction of carbon monoxide.

Our project requires multiple input values which describes the concentration of the air molecules in that respective area and will predict an output "AQI" depending upon the allready trained dataset which is holding the previously collected air molicules data and the corresponding AQI values.

First we have selected the linear regression algorithm for our project implementation and applied it . And after preprocessing the data by removing the unwanted datasets and converting the data into relevant form,we have plotted a corellation graph for finding the relation between the independent variables,and then depending on the results we have finalised the dataset. While dividing the dataset for train&testing we have taken 80% of data for training and 20% for testing purposes .

While working with the linear regression algorithm we have got an error rate of 57. So we have changed our algorithm and applied Knearest neighbours algorithm to reduce the error rate and finally we reduced the error rate from 57 to 30. As the error rate decreases the code which is written will be used for future prediction.

Impact and outcome:

The goal is to use state-of-the-art research results with a special focus on machine learning methods about air quality prediction, to evaluate and apply ideas and algorithms from these studies in the context of time series prediction. The ultimate goal of air quality prediction is to enable national and global decision-makers, communities, and individuals to proactively take measures to reduce the health hazards caused by air pollution.

We can utilize the proposed approach to forecast data from other cities in the future.

We can also determine the contaminated area and the cause of the pollution using prediction. Some pollutants are hazardous to human health, posing a major threat in the future.

In this, data from different sensors are collected in order to frame data set. Data set includes The following entities like particulate matter (PM2.5, PM10), Nitrogen Dioxide (NO2) Carbon Monoxide (CO), Sulphur Dioxide (SO2) and Ozone (O3) in which CO should be predicted.

The dataset contains data collected from past years. For more accurate predictable value, more data is required. When there exists more distance between the value that need to be predicted and the values of dataset, accuracy decreses in case of crisis between those years. So, value that need to be predicted should be less distant from the values of dataset.