SMARTBRIDGE
Let's Bridge the Gap

INTERNSHIP REPORT APPROVAL FORM

July 1, 2019

With immense pleasure, this is to approve that the students of Sona College of Technology, i.e.,

Arunbalaji S (1516102011),

Deekshitha I S (1516102021),

Dharshini P G (1516102030) and

Neelashkumar P (1516102076)

successfully completed their Project and Project Report on **“Air Quality Prediction”** under our guidance.

We are highly impressed with the work that they have done and commend them on their quick grasping skills. They have shown good intent to learn and have put the knowledge gained into application in the from of this project. We appreciate the hard work and commitment shown by them.

We, hereby approve that this document is completely checked and accepted by SmartBridge Technical Team. Its been an absolute pleasure to educate and mentor these students. We hope that this document will also serve as a Letter of Recommendation, to whomsover applied.

We wish them success in all future endeavors and a great career ahead.

GD Abhishek

AI Developer

ABSTRACT

Awareness of daily levels of air pollution is important to the citizens, especially for those who suffer from illnesses caused by exposure to air pollution. Further, the success of a nation to improve air quality depends on the support of its citizens who are well-informed about local and national air pollution problems and about the progress of mitigation efforts. Air quality index (AQI) or air pollution index (API) is commonly used to report the level of severity of air pollution to the public. A number of methods were developed in the past by various researchers/environmental agencies for determination of AQI or API but there is no universally accepted method exists, which is appropriate for all situations. The intended use of the air quality index is to identify the vulnerable zone. There are mainly two approaches viz. single pollutant index and multi-pollutant index to determine the air quality index. Every index has its own characteristic strengths and weaknesses that affect its suitability for particular applications. In this paper, we use single pollutant index, to tackle air quality forecasting by using machine learning approaches to predict the concentration of air pollutant i.e., Ozone(O_3). Machine learning, as one of the most popular techniques, is able to efficiently train a model on big data by using large-scale optimization algorithms. Although there exist some works applying machine learning to air quality prediction, most of the prior studies are restricted to several-year data and simply train standard regression models (linear or nonlinear) to predict the hourly air pollution concentration. In this work, we propose refined models like Random Forest Algorithm, to predict the air pollution concentration on the basis of historical data of previous days. Our experiments have shown that the proposed system achieve better performance than existing standard regression models and existing regularizations.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Air pollution is a global environmental problem that influences mostly health of urban population. Over the past few decades, epidemiological studies have demonstrated adverse health effects due to higher ambient levels of air pollution. Studies have indicated that repeated exposures to ambient air pollutants over a prolonged period of time increases the risk of being susceptible to air borne diseases such as cardiovascular disease, respiratory disease, and lung cancer (WHO, 2009). Air pollution has been consistently linked to substantial burden of ill-health in developed and developing countries. Globally, many cities continuously assess air quality using monitoring networks designed to measure and record air pollutant concentrations at several points deemed to represent exposure of the population to these pollutants. Current research indicates that guidelines of recommended pollution values cannot be regarded as threshold values below which a zero adverse response may be expected. Therefore, the simplistic comparison of observed values against guidelines may mislead unless suitably quantified. In recent years, air quality information are provided by governments to the public comes in a number of forms like annual reports, environmental reviews, and site or subject specific analyses/report. These are generally having available or access to limited audiences and also require time, interest and necessary background to digest its contents. Presently, governments throughout the world have also started to use real-time access to sophisticated database management programs to provide their citizens with access to site-specific air quality index/air pollution index and its probable health consequences. Thus, a more sophisticated tool has been developed to communicate the health risk of ambient concentrations using air pollution index (API) or air quality index (AQI). The World Health Organization (WHO) estimates that 25% of all deaths in the developing world can be directly attributed to environmental factors (WHO, 2006). The problem of air pollution and its corresponding adverse health impacts have been aggravated due to increasing industrial and other developmental activities.

1.1 PROBLEM STATEMENT

In order to predict air pollution conditions, it is essential to handle and manage historical data sets of the parameters measured. Considering the vast amount of data available and to distinguish the pattern and extent of relationships for useful and efficient extraction of knowledge, there is a need for using data analysis techniques. Much of the spatial data obtained are sparse in nature, for example, pollution levels at different locations. Thus, a method for obtaining a continuous data set from a sparse data set is a practically useful need. The present project examines the air pollution data in India and attempts to forecast the values. More specifically, this project uses data analysis tools and techniques of artificial intelligence like Random Forest Regression models, to forecast the air pollutant Ozone(O_3) in the monitored location.

1.2 OBJECTIVE

The primary objective of the project is to offer scientific analysis of pollution data to predict the level of ozone and thus determine the Air Quality Index (AQI). The analysis is done by applying machine learning and data mining techniques to the historical pollution data.

The following are some of the other objectives:

- To monitor meteorological data using IoT
- To adapt methods from machine learning and data mining of spatial data and apply them to pollution data analysis
- To develop a model for estimating the air pollution level
- To integrate methods of spatial pattern analysis using exploratory statistics, spatial data analysis and artificial intelligence
- To develop an application to implement the prediction module.
- To identify the Air Quality Index based on the predictions of the air pollutants.

Chapter 2

REVIEW OF LITERATURE

The prediction of air quality is becoming essential for minimizing the environmental imbalances further effectively addresses the air pollution. There are different types of numerical as well as statistical tools for the prediction and analysis of air pollution. The emergence of advanced computing/analysis techniques from traditional computing methods to recent soft computing techniques are effectively addresses the air quality prediction[1]. The traditional approach for air quality prediction uses mathematical and statistical techniques. In these techniques, initially a physical model was designed and then the data is coded with mathematical differential equations. But such methods suffers from disadvantages like they provide limited accuracy as they were unable to predict the extreme points i.e. the pollution maximum and minimum cut-offs cannot be determined using such approach. Also, such methods were lengthy and inefficient approach for better output prediction. But with the advancement in technology and research, an alternative to traditional methods has been proposed i.e. Artificial Intelligence (AI) techniques can be used for prediction purposes. Among various types of soft computing techniques, the following are the major air pollution predictive model techniques.

- Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)
- Fuzzy Logic (FL)
- Hidden Markov Model (HMM)
- Genetic Algorithm
- Particle Swarm Intelligence
- Hybrid soft computing techniques

Chapter 3

DATA COLLECTION AND PREPROCESSING

Data collection is one of the most important parts of building machine learning models. Because no matter how well designed our model is, it won't learn anything useful if the training data is invalid. We collected the meteorological data from “kaggle.com” which includes attributes like Temperature, Solar Radiation, Wind Speed and Ozone at a particular period. We paired the collected meteorological data and air pollutant data on the basis of time to obtain the required data format for applying the machine learning methods. In particular, for each variable, we formed one value for each day. However, the original data may have contained multiple records or missing values at some days. To preprocess the data, we calculated the daily mean value of each numeric variable if there were multiple observed records within a day. Thus the data collected is completely preprocessed for the data modelling stage.

Chapter 4

PROPOSED SYSTEM

4.1 ARCHITECTURE DIAGRAM

The below figure (Fig. 4.1.1) depicts the architecture of the proposed system.

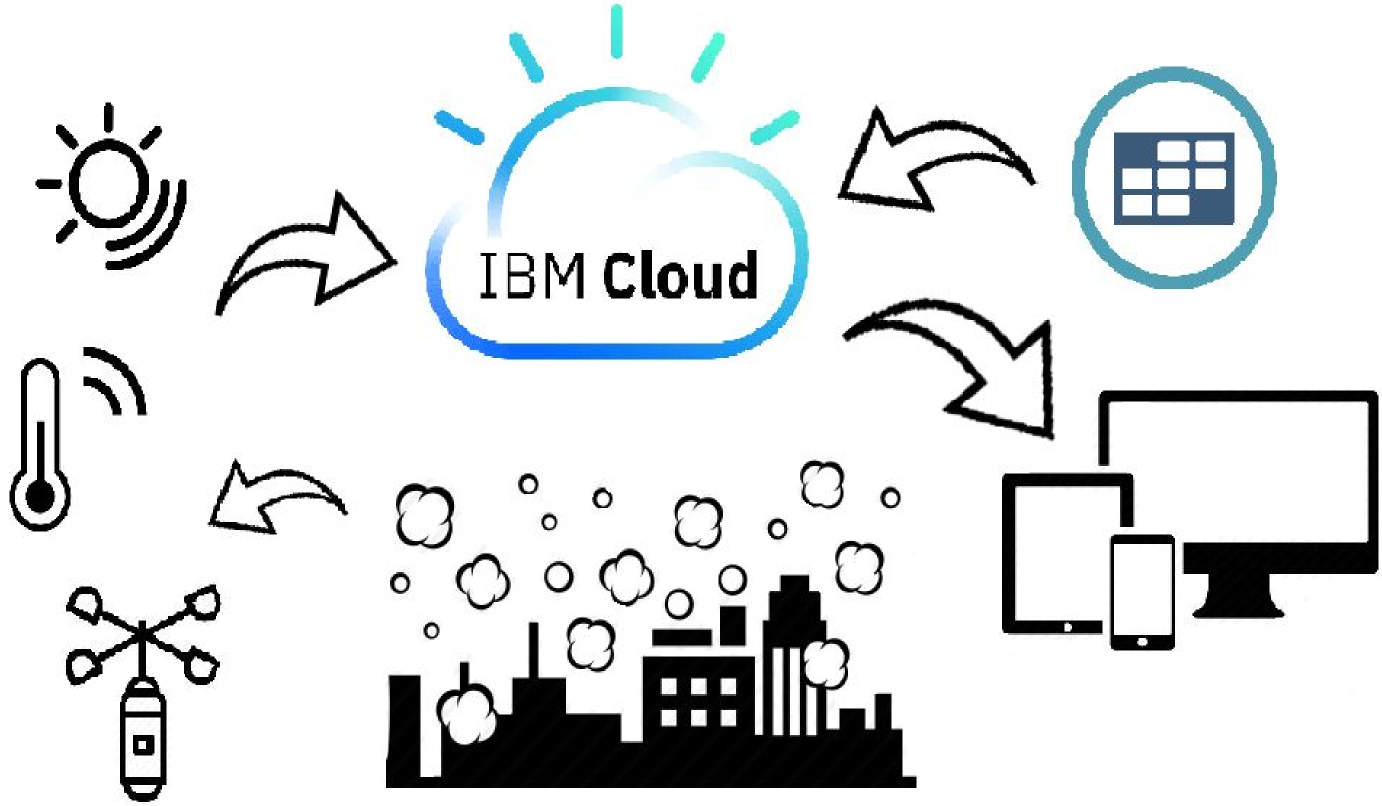


Fig 4.1.1 Architecture of the project

4.2 IoT MODEL

In this project, we use the level-6 of IoT. A level-6 IoT system has multiple independent nodes that perform sensing and/or actuation and send data to the cloud. Data is stored in the cloud and the application is cloud-based. The analytics component analyzes the data and stores the results in the cloud database. The results are visualized with the cloud-based application. The centralized controller is aware of the status of all the end nodes and sends control commands to the nodes. The figure given below (Fig. 4.2.2) describes the IoT model used in the project.

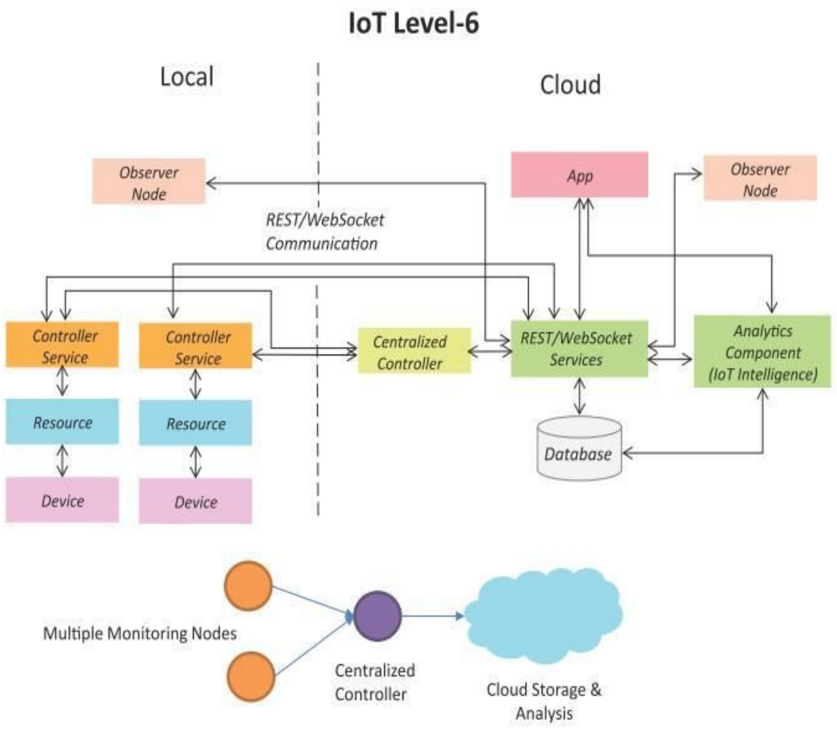


Fig 4.2.2 IoT Model

(Courtesy : Vijay Madisetti, Arshdeep Bahga (Authors))

4.3 DATA MODELLING

4.3.1 Random Forest Regression

The Data Model chosen for Air Quality Prediction is Random Forest Regression. **Random forests** or **random decision forests** are an ensemble learning methods for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees[2]. Random decision forests correct for decision trees habit of overfitting to their training set.

The **random forest** model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$$

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called **model ensembling**. In random forests, all the base models are constructed independently using a different **subsample** of the data.

4.3.2 Why Random Forest

Different kinds of models have different advantages. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target.

One important note is that tree based models are not designed to work with very sparse features. When dealing with sparse input data (e.g. categorical features with large dimension), we can either pre-process the sparse features to generate numerical statistics, or switch to a linear model, which is better suited for such scenarios.

4.3.3 Implementation of Algorithm

The dataset consists of meteorological attributes such as Temperature, Solar Radiation, Wind Speed and Ozone at a particular period. The dataset was first cleaned by applying many data wrangling steps. After analysing the data, it is found that Regression type algorithm should be used. Because the dependent variable is continuous.

The diagram given below (Fig. 4.3.3.1) is the analysis of data, that depicts how the level of ozone is affected by the parameters like Solar Radiation, Wind Speed, Temperature, Month and Day.

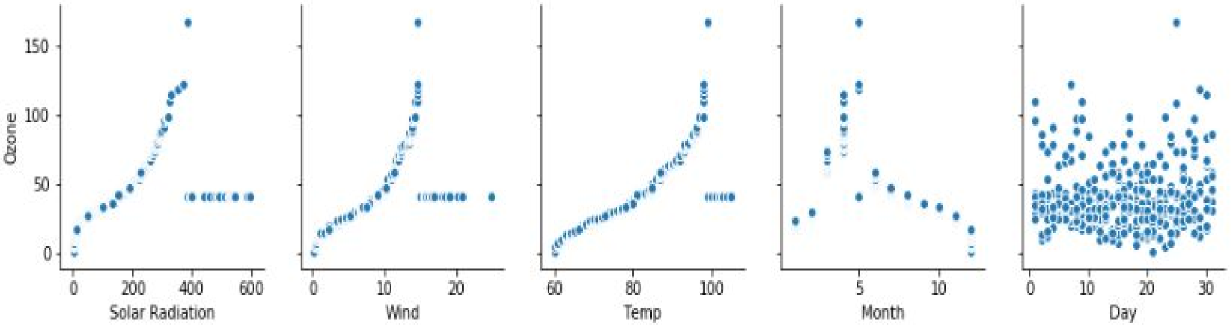


Fig 4.3.3.1 Pair Plot depicting relation between attributes

There are many Regression algorithms like Simple Linear Regression, Multiple Linear Regression, Polynomial Regression, Decision Tree, Random Forest, etc. After implementing all the algorithms for our dataset, we have found out the performance metrics by using "r2score" and "rmse". On the basis of the metrics, we come to a conclusion that Random Forest has high performance. So, our model is finally trained by using this algorithm. The below figure depicts the correlation between Ozone and the other factors.

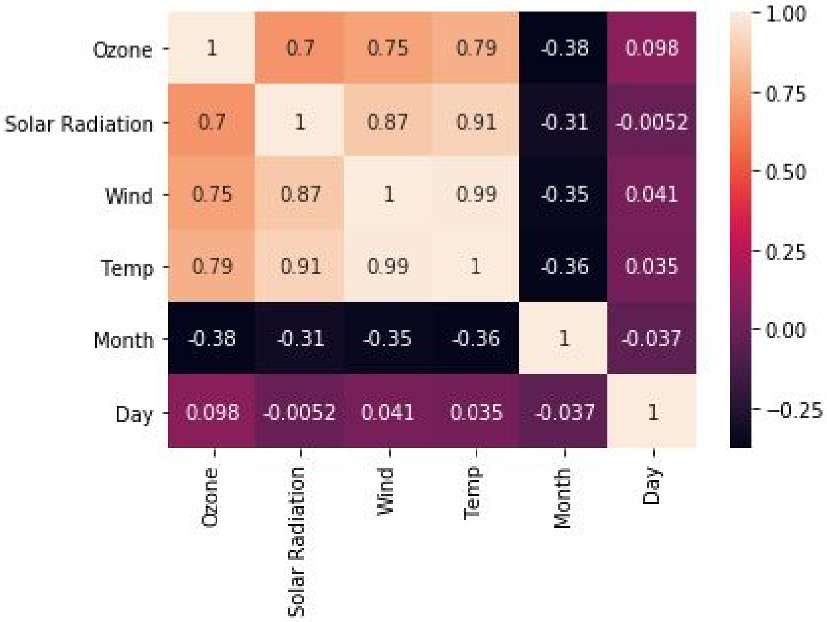
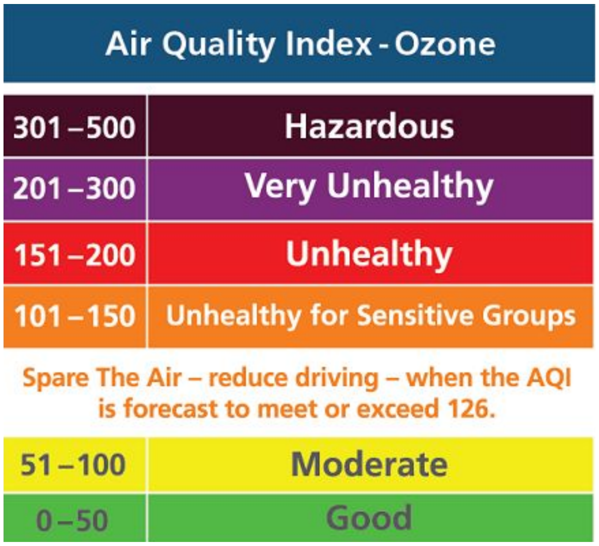


Fig 4.3.3.2 Heat Map

4.4 IMPLEMENTATION

IoT sensor network, which consists of three sensors, namely, Temperature sensor(LM 35), Photosensitive Light-Dependent Control Sensor(LM 393), Wind Speedometer is used to sense the meteorological data i.e., Temperature, Solar radiation and Wind speed. These parameters are chosen because it affects the Ozone level in atmosphere. This Ozone(O_3) is the major pollutant that affects the Air Quality. The figure below (Fig. 4.3.1) depicts the range of Air Quality Index with respect to Ozone[3].



Air Quality Index - Ozone

301–500	Hazardous
201–300	Very Unhealthy
151–200	Unhealthy
101–150	Unhealthy for Sensitive Groups

Spare The Air – reduce driving – when the AQI is forecast to meet or exceed 126.

51–100	Moderate
0–50	Good

Fig 4.3.1 Air Quality Index

Sensed data from the sensor network, is then passed to the IBM Cloud through IoT services. There are two applications that is used to predict the AQI with the help of ML model. When the user requests for prediction, the data is fetched from the IBM cloud. This in turn is sent to the model for prediction. The predicted value is then displayed to the user in the Web or Android application.

Chapter 5

CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning model (Random forest regression) can be efficiently used to detect the quality of air. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities. In this paper, we have developed efficient machine learning methods for air pollutant prediction. We have focused on alleviating model complexity by reducing the number of model parameters and on improving the performance by using Random Forest model. Our results show that the proposed light formulation achieves much better performance than Linear Regression models by enforcing prediction models for consecutive days to be close can also boost the performance of predictions. For future work, we will further consider the commonalities between nearby meteorology stations and combine them in our framework, which may provide a further boost for the prediction.

REFERENCES

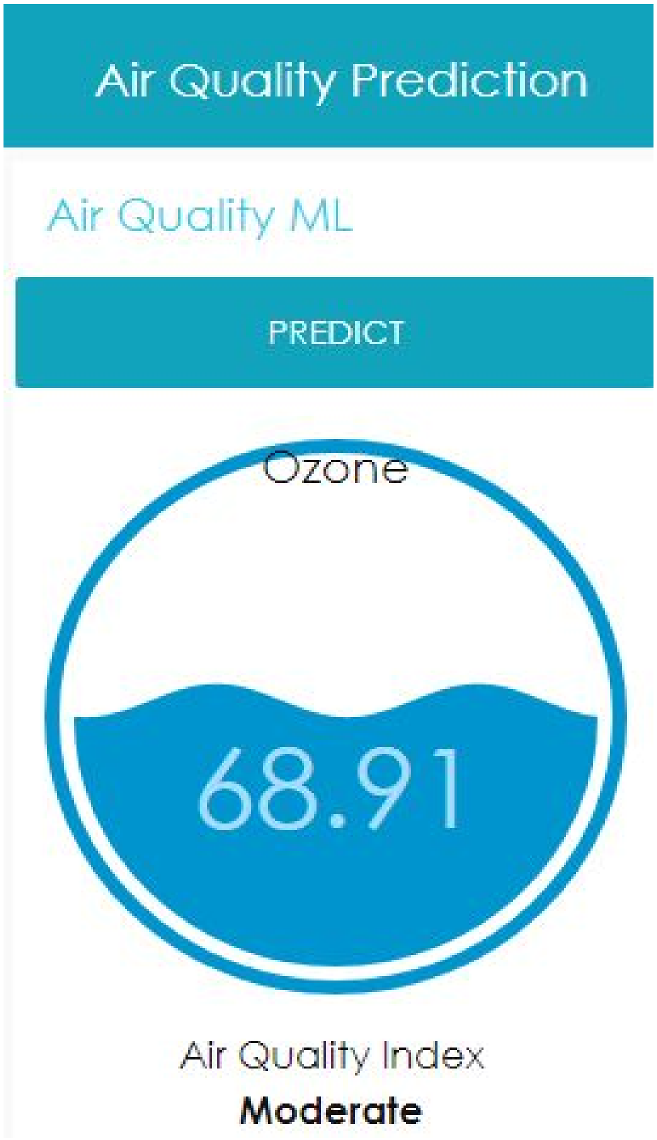
[1]https://pdfs.semanticscholar.org/a6a2/c12a0e90537c9c96f6b076cd53819ec4aabe.pdf

[2]https://en.wikipedia.org/wiki/Random_forest

[3]https://www.google.com/search?q=air+quality+ozone&safe=active&rlz=1C1GCEU_enIN851IN851&source=lnms&tbm=isch&sa=X&ved=0ahUKEwil_d-qy47jAhVELI8KHV3bAssQ_AUIESgC&biw=1600&bih=757#imgrc=Esr2cSTDNN31pM:

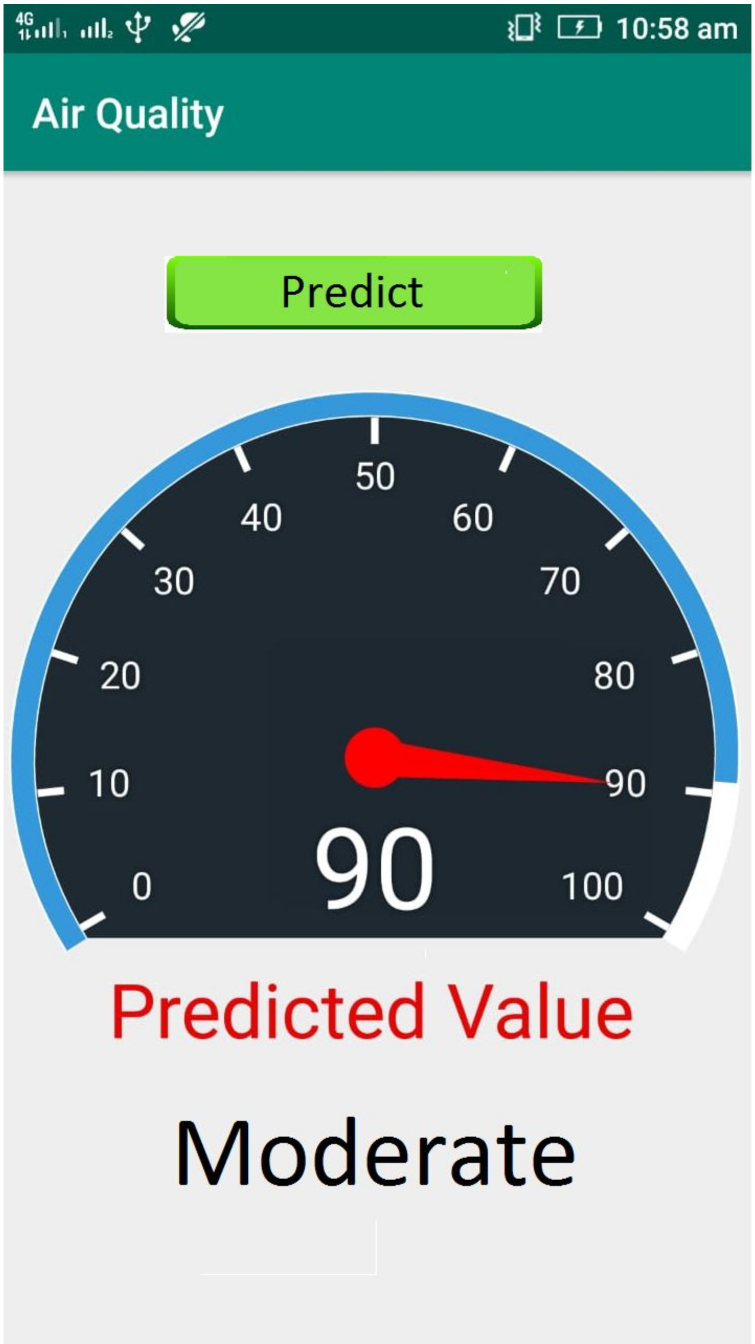
APPENDIX

Web Application



The above image is a screenshot of the Web application created using NodeRed Flow in IBM Watson. When the button "Predict" is clicked, the data from IoT is sent to the machine learning model and the output which is the level of Ozone is shown. And also the level of Air Quality is shown based on the Ozone value.

Android Application



The above image is a screenshot of the Android application created using Android Studio. When the button "Predict" is clicked, the data from IoT is sent to the machine learning model and the output which is the level of Ozone is shown. And also the level of Air Quality is shown based on the Ozone value.