An accurate Estimation of Air Quality using Linear Regression model of Machine Learning

Amit Kumar, Mahesh Vaishnav, Akhilesh Kaushik,

Sai Jeevna, Kamla Shree, Nidhi Lal

Dept. of Computer Science and Engineering
IIIT Nagpur, India

amit.kumar@cse.iiitn.ac.in, mahesh.vaishnav@cse.iiitn.ac.in , akhilesh.kaushik@cse.iiitn.ac.in ,
jeevanasai6013@gmail.com, kamala4sree@gmail.com , nidhi.lal@cse.iiitn.ac.in

Abstract— **Air Quality Index (AQI) is a standardized summary measure of ambient air quality used to express the level of health risk related to particulate and gaseous air pollution. There is no warning alarm system in many countries yet, environmental warning system exists in Poland, although some test-trials took place in Katowice area and the city of Gdansk. The aim of the estimation is to get an accurate estimation of Air Quality and to confront AQI categories with local air quality, also in terms of health impact on the population. The number of deaths due to cardiovascular and respiratory diseases in elderly population (aged 65 and more) has increased now a day. Further we propose an accurate index of air quality so that we can predict the quality of the air and thus we can take precaution by getting proper mask on our face that day.**

***Keywords*— Ari Quality Index, Warning Alarm System, Precaution, Healthcare.**

I. Introduction

Worsening air quality is one of the major global causes of premature mortality and is the main environmental risk claiming seven million deaths every year [1]. Nearly all urban areas do not comply with air quality guidelines of the World Health Organization (WHO) [2,3]. The risk populations that suffer from the negative effects of air pollution the most are children, elderly, and people with respiratory and cardiovascular problems. These health complications can be avoided or diminished through raising the awareness of air quality conditions in urban areas, which could allow citizens to limit their daily activities in the cases of elevated pollution episodes, by using models to forecast or estimate air quality in regions lacking monitoring data.

Air pollution modelling is based on a comprehensive understanding of interactions between emissions, deposition, atmospheric concentrations and characteristics, meteorology, among others; and is an indispensable tool in regulatory, research, and forensic applications [4]. These models calculate and predict physical processes and the transport within the atmosphere [5]. Therefore, they are widely used in estimating and forecasting the levels of atmospheric pollution and assessing its impact on human and environmental health and economy [6–9]. In addition, air pollution modelling is used in science to help understand the relevant processes between emissions and concentrations, and understand the interaction of air pollutants with each other and with weather [10] and terrain [11,12] conditions. Modelling is not only important in helping to detect the causes of air pollution but also the consequences of past and future mitigation scenarios and the determination of their effectiveness [4]. There are a few approaches to air quality modelling –atmospheric chemistry, dispersion (chemically inert species), and machine learning with a huge data set. These models are based on assumption of continuous emission, steady state condition and conservation of mass. Here we

II. Related work

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.

Source of the data set was taken from Saverio De Vito (saverio.devito '@' enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development. They used linear model of regression using machine learning and calculated R^2 value as 0.9991371797127734. They used ratio of training and testing of the data set as 70:30 percentage. Figure 1 and figure 2 shows the predicted linear regression graphs.

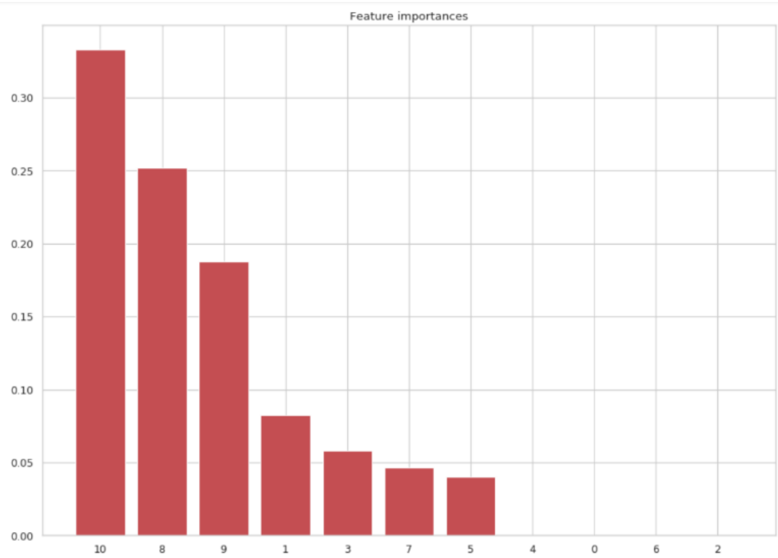

Figure 2

<matplotlib.collections.PathCollection at 0x7f7ce4ae09b0>

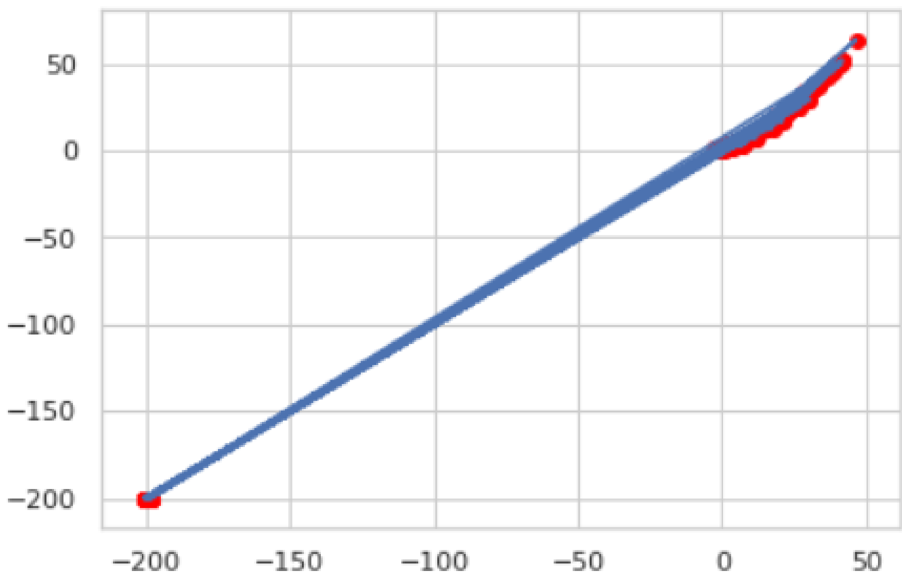

Figure 1

III. Proposed Work

In the project we use linear regression of sklearn module. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. If we plot the independent variable (x) on the x-axis and dependent variable (y) on the y-axis, linear regression gives us a straight line that best fits the data points, as shown in the figure below. By varying the dataset by careful selection and possible manipulation of our testing and training data's. And by carefully selection of the random state.

When we use iteration in our dataset then our R^2 value is Change from 0.9991371797127734 to 0.9992884871433486,
mean_squared_error from 1.4571943722667688 to 1.1867347957963301,
Mean_absolute_error from 0.8060244440070082 to 0.7909720964997654,
Root_mean_squared_error 1.2071430620546881 to 1.0893735795384107.

Figure 3 and Figure 4 shows the predicted but more accurate linear regression graphs.

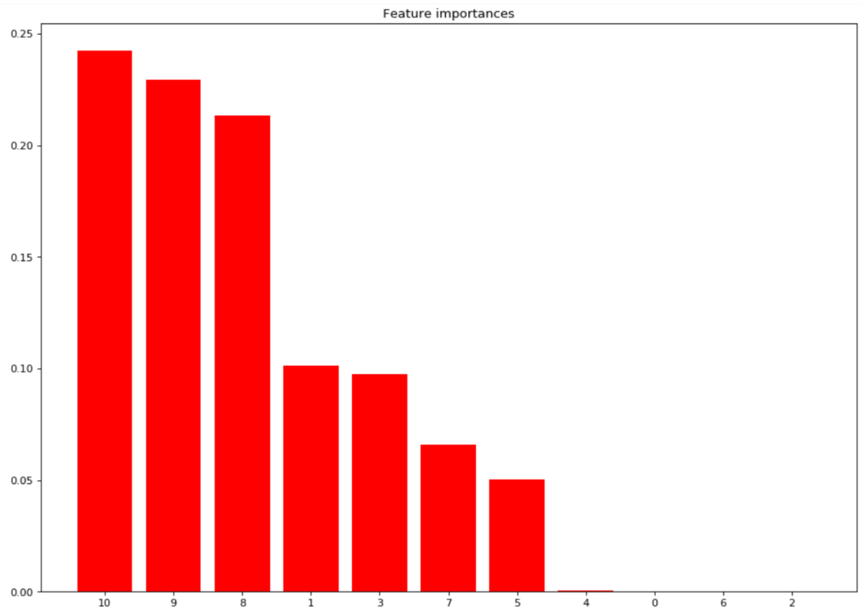

Figure 4

<matplotlib.collections.PathCollection at 0x7ff621a51550>

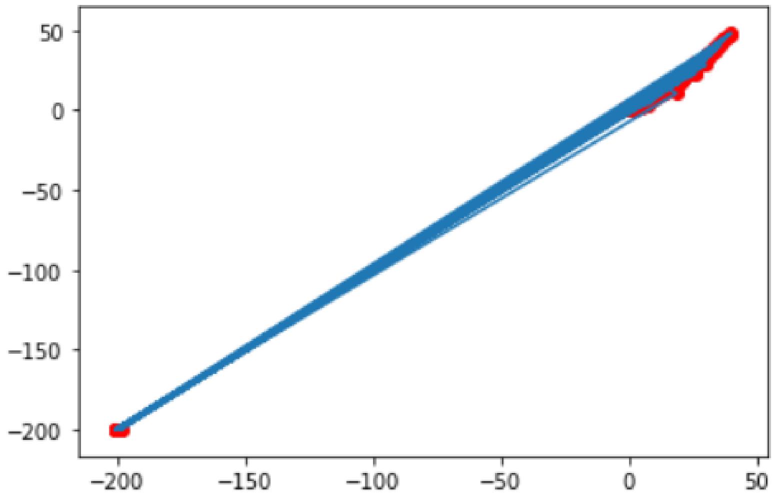

Figure 3

IV. Result and Discussion

The accuracy of the air quality is improved as a result of the project. The R^2 score of the native model is increased as well as MSE , RMSE , MAE values are decreased means error is decreased as result by careful selection and possible manipulation of our data's features. The purpose of this is to feed our model only the most optimal form of input. We can consistently give our model only the parts of the data it needs to make accurate predictions, then it doesn't have to deal with any extra noise that comes from the rest of the data. And after that taking condition so that the value of r^2 is maximum and all error are minimum for all of the data that we used in prediction.

V. CONCLUSIONS

Air pollution risk is a function of the hazard of the pollutant and the exposure to that pollutant. Air pollution exposure can be expressed for an individual, for certain groups (e.g. neighborhoods or children living in a country), or for entire populations. For example, one may want to calculate the exposure to a hazardous air pollutant for a geographic area, which includes the various microenvironments and toxic gases. WHO has convened a Global Platform on Air Pollution and Health with experts across academia and government, to improve methods of global, regional and national monitoring and surveillance of air pollution exposures, ensuring open-access to air quality data.

Our ratio of training and testing of the data set is 70:30 percentage. By the use of linear regression, we improve the accuracy and as a result our model can predict more accurate air quality index from given dataset. And finally reduced the error and improved the accuracy from the older version. In future we will try to get more accuracy by improving and optimizing our method.

VI. REFERENCES

[1] WHO. 7 Million Premature Deaths Annually Linked to Air Pollution Page 1 of 2 WHO|7 Million Premature Deaths Annually Linked to Air Pollution Page 2 of 2. Available online: http://www.who.int/mediacentre/ news/releases/2014/air-pollution/en/#.WqBfue47NRQ.mendeley.

[2] Limb, M. Half of wealthy and 98% of poorer cities breach air quality guidelines. BMJ 2016, 353. [CrossRef] [PubMed]

[3] WHO.AirPollutionLevelsRisinginManyoftheWorld'sPoorestCities. Availableonline: http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/en/#.WhOPG9ANlBk.mendeley.

[4] Daly, A.; Zannetti, P. Air pollution modeling—An overview. Ambient Air Pollut. 2007, I, 15–28.

[5] Met Office. Numerical Atmospheric-Dispersion Modelling Environment (NAME) Model. Available online: http://www-cast.ch.cam.ac.uk/cast_pics/WP_NAME.pdf (accessed on 5 December 2018).

[6] Cohen, A.J.; Brauer, M.; Burnett, R.; Anderson, H.R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. Lancet 2017, 389, 1907–1918. [CrossRef]

[7] Kinney, P.L. Climate Change, Air Quality, and Human Health. Am J. Prev. Med. 2008, 35, 459–467. [CrossRef] [PubMed]

[8] Lelieveld, J.; Evans, J.S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 2015, 525, 367. [CrossRef] [PubMed]
[9] Pannullo, F.; Lee, D.; Neal, L.; Dalvi, M.; Agnew, P.; O'Connor, F.M.; Mukhopadhyay, S.; Sahu, S.; Sarran, C. Quantifying the impact of current and future concentrations of air pollutants on respiratory disease risk in England. Environ. Health 2017, 16, 29. [CrossRef] [PubMed]
[10] Silva,R.A.;West,J.J.;Lamarque,J.-F.;Shindell,D.T.;Collins,W.J.;Faluvegi,G.;Folberth,G.A.;Horowitz,L.W.; Nagashima, T.; Naik, V.; et al. Future global mortality from changes in air pollution attributable to climate change. Nat. Clim. Chang. 2017, 7, 647–651. [CrossRef]
[11] Kim, D.; Stockwell, W.R. An online coupled meteorological and air quality modeling study of the effect of complexterrainontheregionaltransportandtransformationofairpollutantsovertheWesternUnitedStates. Atmos. Environ. 2008, 42, 4006–4021. [CrossRef].
[12] Grigoras, G.; Cuculeanu, V.; Ene, G.; Mocioaca, G.; Deneanu, A. Air pollution dispersion modeling in a polluted industrial area of complex terrain from Romania. Rom. Rep. Phys. 2012, 64, 173–186.