A comparative study of machine learning algorithms for the analysis and prediction of air quality parameters in Kolkata.

Synopsis: Air quality is measured with the help of AQI(Air Quality Index).Depending on the range, our predicted AQI falls in, people will be given some message regarding AQI’s impact on health(ex: Hazardous, Unhealthy etc).Final AQI is calculated on a daily basis taking the maximum value from all AQIs, calculated using different air quality parameters. Depending on the data availability we have chosen three air quality parameters (PM10, NO2, SO2) and six meteorological parameters(Air temperature, Pressure station level, Pressure mean sea level, Relative humidity, Horizontal visibility, Dew point temperature).We have implemented different machine learning algorithms to comparatively study the changes in behavior of the parameters and have taken the final models for each parameter based on their optimum performances. Those meteorological parameters are our input/independent variables and air quality parameters are our output/dependent variables. As we are dealing with continuous data, our problem is regression problem. The algorithms we have implemented so far are : a) Single variable linear regression, b) Multivariable linear regression, c) Polynomial regression, d) Multivariable polynomial regression, e) Support vector regression, f) Decision tree, g) Random forest and h) some neural network models(LSTM, GRU, simple MLP and wide deep MLP). The models which have given good performances are Support vector regression, Decision tree, Random forest and neural network models. We have chosen four metrics (Mean absolute error, Mean squared error, Root mean squared error, Score value) to evaluate the performance of each model. We have also calculated accuracy of a model by finding out percentage of predictions having less than 20% error. To maintain a non over-fitted model we have split our dataset into three parts(Training set, Validation set, Test set). Except some neural network models, no other models have given over-fitted prediction in our case. To get better accuracy we have introduced some artificial parameters(Day no(1-365), label encoded Date parameter, previous day’s parameter values etc) and it has improved. We have used correlation matrix to get the better understanding of AQI and it’s dependency on air quality parameters. Examining different plots and correlation matrix, we have found that AQI value is mostly dependent on PM10 parameter, followed by NO2, followed by SO2.We have also plotted scatter plots of all models where we have graphically shown the performances of those models(actual value and predicted value scatter plots). Seeing different plots, we are also trying to give some predictive conclusion like which parameter will have less impact in near future etc. We have also designed a web application which uses django in it’s back-end. Here an user can choose a date and our model will give most accurate prediction of that particular day’s PM10, NO2, SO2 and AQI values and a message regarding health issues. Besides that if he wants to get some specific model’s prediction, he can choose that too. All aforementioned plots will be available there.