MLP Model for Predicting the Air Quality

**Problem statement:**

Researchers have established that the multi-layer perceptron is able to store information of learning brief dependencies, and they have applied this capability to a variety of studies, including translation software, speech recognition, and the prediction of air pollution concentrations. In our proposed work, the time series-based air quality dataset of Beijing is used for prediction where MLP model is used developing the prediction capability. Finally, the actual and predicted data need to be identified by the MLP networks. In our work, every feature is trained with MLP and the error values were also be identified.

**Dataset Available**

The air quality dataset is gathered from the University of California, Irvine Repository.  The Beijing air quality dataset from the University of California, Irvine, which comprises meteorological data as well as statistics on PM2.5 (fine particulate matter) air pollution levels. The dataset is collected once every hour using a data interface based on the United States Embassy in Beijing, and it is then deduced from there using the interface. A number of variables are included in this trial clean air UCI data collection, including the date and time, the weather (including wind direction and humidity), the PM2.5 concentration, and other relevant information.

**Dataset Link:**

[**https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data**](https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data)

**Dataset Description:**

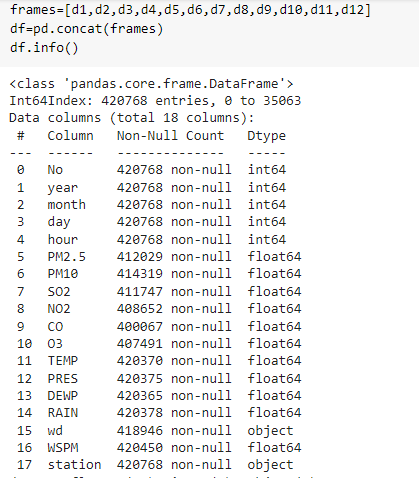
year: year of data in this row

month: month of data in this row  
day: day of data in this row  
hour: hour of data in this row  
PM2.5: PM2.5 concentration (ug/m^3)  
PM10: PM10 concentration (ug/m^3)  
SO2: SO2 concentration (ug/m^3)  
NO2: NO2 concentration (ug/m^3)  
CO: CO concentration (ug/m^3)  
O3: O3 concentration (ug/m^3)  
TEMP: temperature (degree Celsius)  
PRES: pressure (hPa)

DEWP: dew point temperature (degree Celsius)  
RAIN: precipitation (mm)  
wd: wind direction  
WSPM: wind speed (m/s)  
station: name of the air-quality monitoring site

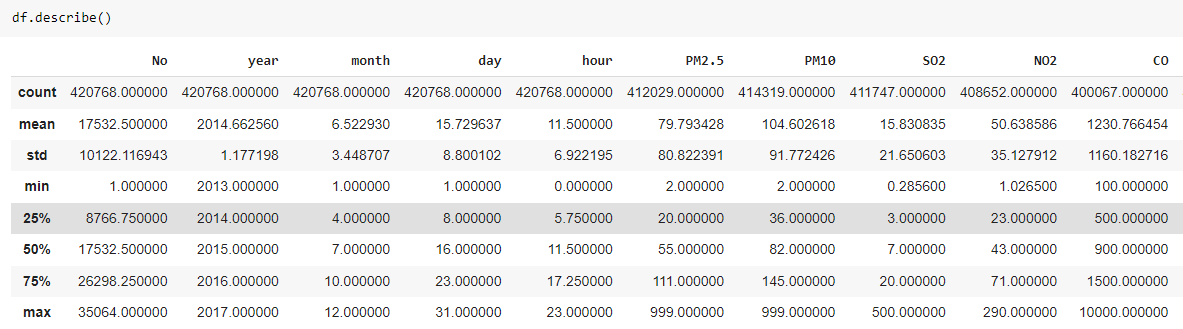
**Data Exploration:**

The Data exploration process is the first step in the procedure. Data exploration is an approach that is quite similar to first data analysis in that it looks for patterns in data. Instead of relying on data management systems to grasp the elements of a data and its attributes, it makes use of visual exploration to accomplish this. For example, the size, culmination, correctness, and potential connection between distinct data components or data files/tables are all characteristics that can be measured. We have 4,20,768 instances in our dataset, and the key features that were present in the dataset were 17 in number.

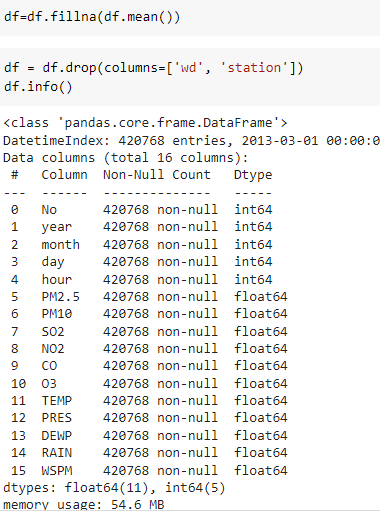


The Pandas DataFrame describe () method returns a list of 8 statistical attributes for each property.

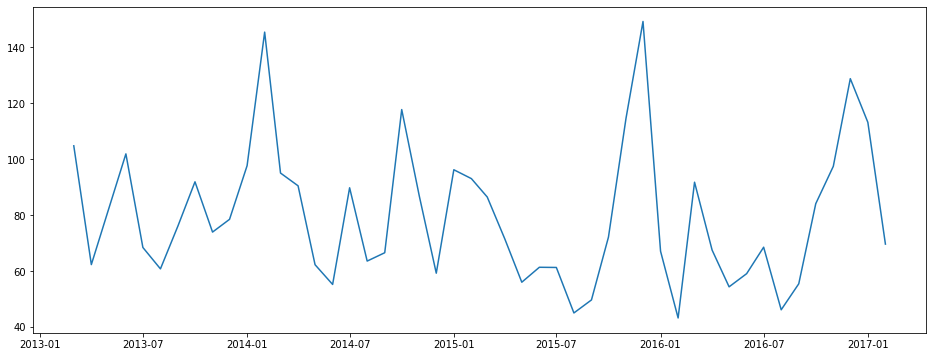
* Count
* Mean
* Std (Standard Deviation)
* Min (Minimum Value)
* 25% (25th Percentile)
* 50% (50th Percentile)
* 75% (75th Percentile)
* Max (Maximum Value)



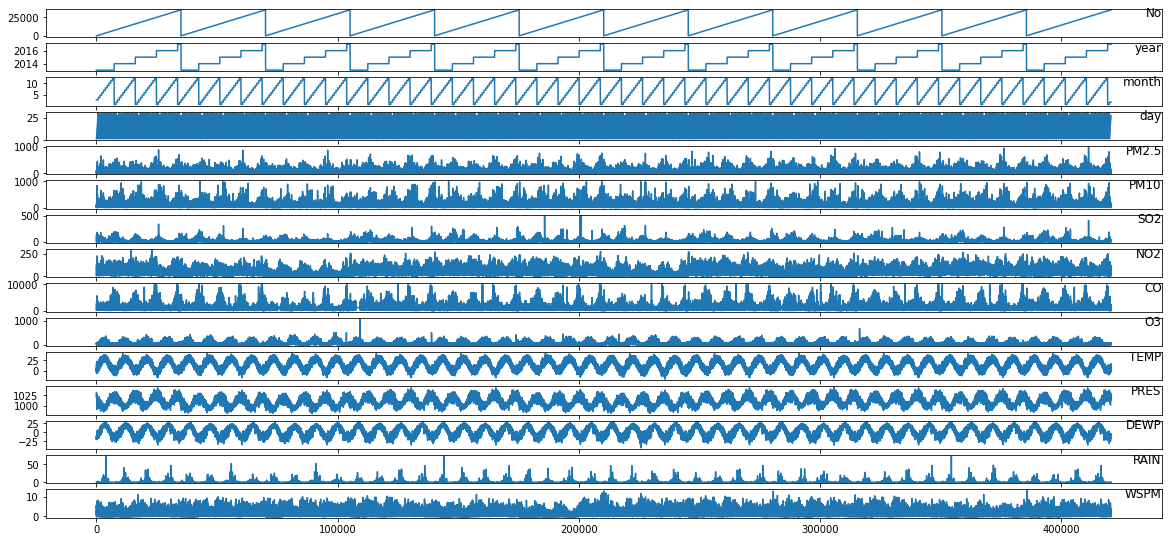
We can see that there are some missing data in the count when we look at the statistical summary. As a result, missing values must be determined prior to settling on the mean. Even if missing values are seen, they should be considered as such they can diminish the power / fit of such a model can result in a biased model if the behaviour and relationship with other parameters are not carefully examined. It has the potential to result in incorrect prediction or classification. The missing values are replaced with the mean, and the wd and station datatypes are used because they are object datatypes in the variable. As a result, they have been removed. The dataset now contains 15 characteristics and has a total size of 54.6MB+.



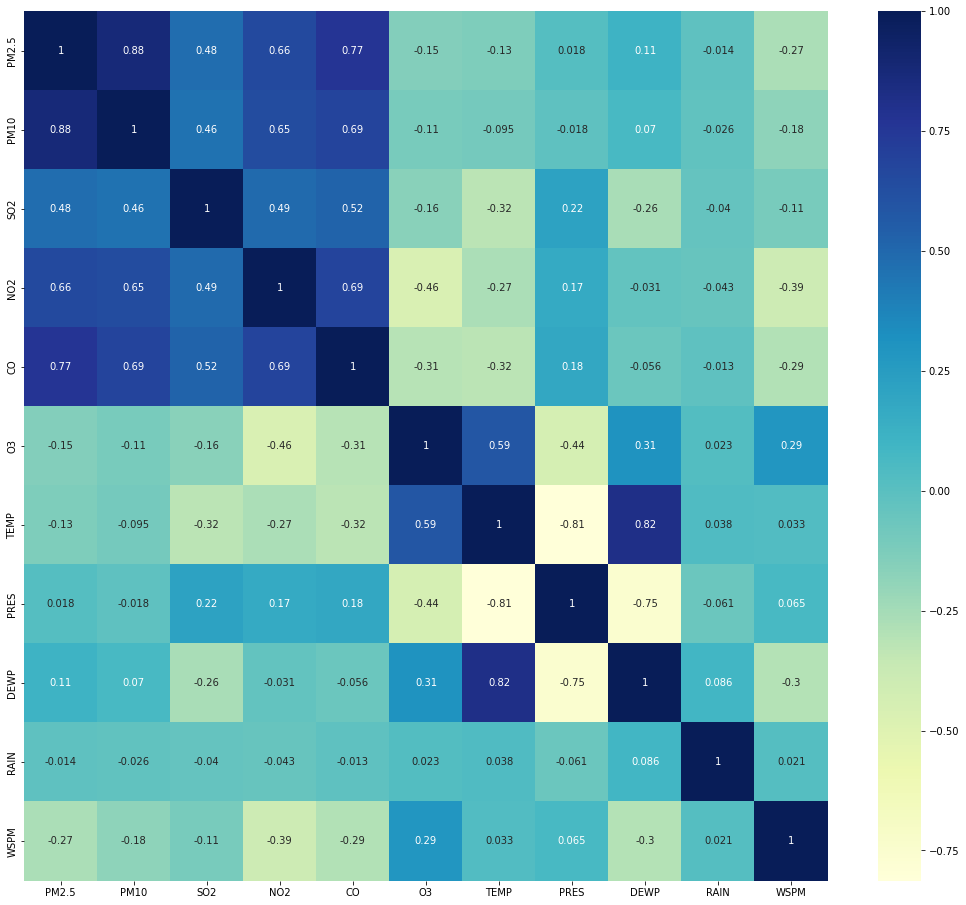
Another significant visualisation in EDA Analysis is the distribution of observations, which is illustrated in the figure below. The symmetrically or asymmetrically distributed observations in some linear time - series data forecasting algorithms are well-behaved, whereas others are not. As a result, the density plot is used to make year-by-year observations upon the data set being studied. The following is an example of a PM2.5 observation.



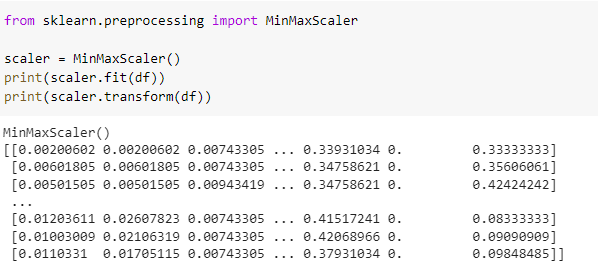
The graph that illustrates the content of various pollutants, can get a sense of the increase and decrease in their concentration levels in the environment. In the following image, a graph is displayed for each of the pollutants, with the x-axis reflecting the samples and the y-axis representing the concentration in micrograms per cubic metre of water.



In addition to that, the predictive link between the answer and the predictor variables; in the case of a strong positively or negatively correlation, the predictors can be used as feature for training the models. To determine multicollinearity, we must first understand the linear relation between predictor variables, which is accomplished using to determine **multicollinearity, hence we use Pearson correlation heat map**. If the correlation among predictor is more than 0.7 or less than 0.7, one of predictor variables can be eliminated from the model when it is being trained.

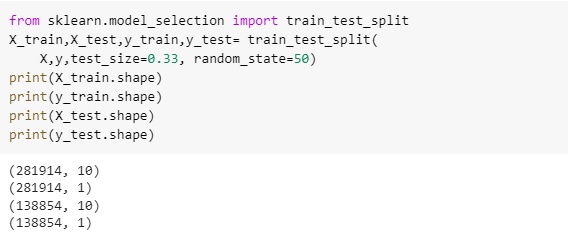


Following the completion of the EDA Study, the data preparation process is initiated. The term "data processing" refers to the manipulation of information by a computer. The following are examples of data processing functions: Validation is the process of confirming that data is accurate and relevant. Sorting and arranging elements in a sequential manner, it is vital to conduct data analysis, which includes data collection, organization, analysis, interpretation, and presentation. Min-max scaling normalization is employed during the pre-processing of data. Normalization is a special case of min-max scaling that involves eventually re the features to the range [0, 1]. Min-max scaling is used to standardize data across one or more feature columns.



**Train and Test Split**

The train-test split can assess a machine learning system's performance in real-world scenarios. It can be used for classification or regression problems, and with any supervised learning algorithm. The method divides the data into two groups using a mathematical equation.  According to our findings, 67% was spent on training and 33% testing.

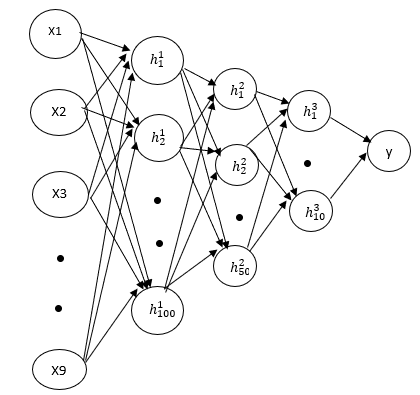


**MLP Modelling & Methodology**

The Multilayer Perceptron (MLP) is a sophisticated nonlinear neural network that can predict air pollution, which makes it a preferred choice over other networks. The MLP begins when the network receives the input parameters that are of relevance to the user. These input parameters generate input signals, which are through the network from the input layer to the output layer via the concealed layer and hidden layer. We can multiply the scaled input sequence by weights, introduced by neurons in each layer by a real-number quantity. The linear model still has the weighted sum data. The activation or transfer function of information or a model is nonlinear.

ReLu was the activation function that was used in this investigation. The input layer of a node is responsible for activating the node or producing the output for such a input in such a neural network. The rectified linear activation function (ReLU) is a piece - wise linear function that produces the input directly if it is positive and zero otherwise. It became the usual perceptron for many sorts of layers since it is better to handle and outperforms in many circumstances. Numerical features are removed from the dataset, which reduces the number of features to ten by deleting the columns of Numerical, Year, Month, Day, and Hour for modelling. As a result, nine features (out of ten) are considered inputs, whereas the prediction-required feature (PM2.5) is deemed an output.

From the MLP model, the input units are 9, 3 layers of hidden network with the neurons of 100,50, and 10 are provided and the output is 1. Hence the structure will be

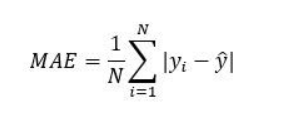


**Results:**

A set of Performance Indicators (PI) was utilized to determine the best-fitting MLP following the application of MLP in the modelling process. There are two factors which are taken into consideration when calculating PI, namely the error and accuracy. Whereas the accuracy measure suggests the best-fitted model was close to 1, the error best-fitted measure that is closer to 0. When the error measure suggests that the most effective model is closer to 0 Because our problem is based on regression, the error metrics that we have chosen are the mean absolute error and the R2 error.

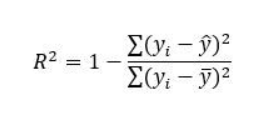
1. Mean Absolute Error

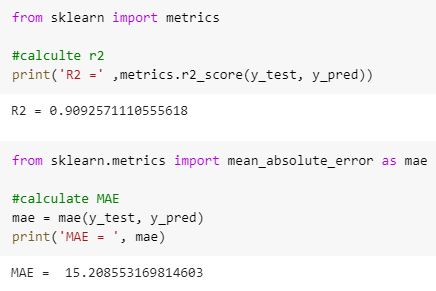
The Mean Absolute Error in a dataset shows the average of the absolute difference between actual and predicted values over the course of the data collection. As a result, it computes an averaged residuals as in dataset.



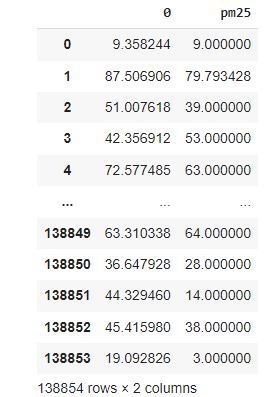
1. R Square error

According to the linear regression model, The R-squared reflects how much variability in the dependent variable that is explained by the model. It is a scale-free score, which means that regardless of whether the values are little or large, the r - square value will always be less than 1.





The actual and predict values obtained from the MLP is showed that



**Conclusion:**

According to the work presented, the air quality of the PM2.5 target observed from other input attributes are visualized. The data pre-processing, correlation and test train splits were also performed. With the multi layer perceptron technique, the hidden layers along with activation function was used to train the model. Since our work is based on regression problem, hence modelling metric errors such as r square and mean absolute error was calculated and the scores were 0.90 and 15.20. The actual and predicted values of PM2. 5 was also determined in our work.

**References:**

[1] International Agency for Research on Cancer, Carcinogen List, <http://samr.cfda.gov.cn/WS01/CL1991/215896.html>, 2017.

[2]Athanasiadis, I.N.; Kaburlasos, V.G.; Mitkas, P.A.; Petridis, V. Applying machine learning techniques on air quality data for real-time decision support. In Proceedings of the First international NAISO symposium on information technologies in environmental engineering (ITEE’2003), Gdansk, Poland, 24–27 June 2003.

[3] Famoso, F.; Wilson, J.; Monforte, P.; Lanzafame, R.; Brusca, S.; Lulla, V. Measurement and modeling of ground-level ozone concentration in Catania, Italy using biophysical remote sensing and GIS. Int. J. Appl. Eng. Res. 2017, 12, 10551–10562.

[4] Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 2008, 42, 7561–7578.

[5] Lee, J.-H.; Wu, C.-F.; Hoek, G.; De Hoogh, K.; Beelen, R.; Brunekreef, B.; Chan, C.-C. Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. Sci. Total. Environ. 2014, 472, 1163–1171

[6] Singh, K.P.; Gupta, S.; Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 2013, 80, 426–437.

[7]Corani, G. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 2005, 185, 513–529.