



Prediction of Air Quality Index using Machine Learning and Deep Learning Techniques

Poornima Ravindra

10535945@mydbs.ie

Dissertation submitted in partial fulfilment of the requirements for the degree of
M.Sc. Data Analytics
at Dublin Business School

Supervisor: Obinna Izima

August 2020

Declaration

I declare that this applied project that I have submitted to Dublin Business School for the award of Master of Science in Data Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.'

Signed: Poornima Ravindra

Student Number: 10535945

Date: 25 Aug 2020

Acknowledgment

Firstly, I might want to offer my true thanks to my supervisor Obinna Izima for his persistent help and guidance throughout the process of implementation and documentation.

I also thank my parents for their constant support and encouragement towards my master's.

Lastly, I would like to thank the Dublin Business School department for providing me the opportunity to conduct the research in my field of interest.

Abstract

Over the past few decades air emissions and its mitigation are the fundamental challenges faced by the world. The mortality rate is increased as the air pollution as adversely affected respiratory and cardiovascular system. Lots of efforts has been taken by the government to predict air quality levels which aims at improving public health. This research scientifically contributes to the air pollution challenges so that preventive measures can be taken by the people. Since this dataset doesn't have any label so K-means Cluster is used to find the class for this dataset. Artificial Neural Network and other four Machine Learning techniques are used for classification and it is compared to check which gives more accuracy in predicting the Air Quality Index. Auto Regression is used to find next three months air quality. The dataset contains air quality data from various stations across multiple cities in India.

Table of Contents

1	Introduction	8
1.1	Hypothesis.....	9
1.2	Research Questions	9
■	Literature Review	10
2.1	Introduction	10
2.2	Air pollution monitoring and prediction using Machine Learning and Deep Learning Techniques.....	10
2.3	Air pollution monitoring and prediction using Internet of Things (IOT)	17
■	Background	18
3.1	Machine Learning and Deep Learning process	18
3.2	Comparison between Deep Learning and Machine Learning Techniques.....	20
3.3	How does Artificial Intelligence help improve air quality?	21
■	Research Methodology	22
4.1	Data	22
4.2	<i>Software Used</i>	22
4.3	<i>Data Preparation</i>	23
■	Modelling	24
5.1	K-means clustering.....	24
5.1.1	Overview	24
5.1.2	Implementation	25
5.2	Artificial Neural Network	25
5.2.1	Overview	25
5.2.2	Implementation	27
5.3	SVM	27
5.3.1	Overview	27
5.3.2	Implementation	28
5.4	Random Forest.....	28
5.4.1	Overview	28
5.4.2	Implementation	30
5.5	Logistic Regression.....	30
5.5.1	Overview	30
5.5.2	Implementation	31
5.6	Naïve Bayes	31
5.6.1	Overview	31
5.6.2	Implementation	32

5.7	Autoregression.....	32
5.7.1	Overview	32
5.7.2	Implementation	32
■	Evaluation	33
6.1	Confusion Matrix.....	33
6.2	Accuracy.....	34
6.3	Recall.....	34
6.4	Precision.....	34
6.5	F1 Score.....	34
■	Results - Model Comparison	35
7.1	Results of Artificial Neural Network.....	35
7.1.1	Resultant Classification Report	35
7.1.2	Resultant Confusion Matrix	35
7.2	Results of SVM	36
7.2.1	Resultant Classification Report	36
7.2.2	Resultant Confusion Matrix	36
7.3	Results of Logistic Regression	37
7.3.1	Resultant Classification Report	37
7.3.2	Resultant Confusion Matrix	37
7.4	Results of Naïve Bayes	38
7.4.1	Resultant Classification Report	38
7.4.2	Resultant Confusion Matrix	38
7.5	Results of Random Forest	39
7.5.1	Resultant Classification Report	39
7.5.2	Resultant Confusion Matrix	39
7.6	Accuracy comparison of Machine Learning Models.....	40
■	Conclusion and Future Enhancements	41
■	References	42

List of figures

Figure 1:Machine Learning Process	18
Figure 2:Deep Learning Process.....	19
Figure 3:Schematic Representation of a Perceptron.....	26
Figure 4:Execution Window of ANN	27
Figure 5:Depicting the hyperplanes of Support Vector Machines.....	28
Figure 6:Depiction of Logistic Regression graphically.....	30
Figure 7:Confusion Matrix	33
Figure 8:Classification Report of ANN.....	35
Figure 9:Confusion Matrix of ANN.....	35
Figure 10:Classification Report of SVM.....	36
Figure 11:Confusion Matrix of SVM.....	36
Figure 12:Classification Report of Logistic Regression.....	37
Figure 13:Confusion Matrix of Logistic Regression.....	37
Figure 14: Classification Report of Naïve Bayes.....	38
Figure 15:Confusion Matrix of Naïve Bayes.....	38
Figure 16:Classification Report of Random Forest	39
Figure 17: Confusion Matrix of Random Forest.....	39
Figure 18:Comparison of Machine Learning Models.....	40

1 Introduction

Air pollution can cause fatal diseases to human beings, damage to other living things and damage to the environment. When particulate matter, harmful chemical compounds or any other toxic products are released to the atmosphere then it is called as Air Pollution. Monitoring and maintaining air quality along with other essential activities pertaining to preservation of air quality can be deemed as the need of the hour in many industrial and urban areas today. Human activities such as transportation, burning of fossil fuels for the needs of the society have negative effects on air quality.

Nitrogen oxides (NO_x) is one of the extremely toxic gases and Nitrogen dioxide (NO₂) is released when products are burnt at high temperatures in power plants or it is easily generated from the vehicle emissions. Existing research information shows that NO₂ exposure leads to harmful respiratory consequences such as airway swelling in healthy individuals and it intensifies the asthma problems. Carbon monoxide has no color or odor yet extremely dangerous and poisonous gas. Vehicle exhaust and incomplete burning of oil, fuel, wood or coal releases Carbon Monoxide.

Numerous Industrial processes and volcanoes release SO₂. Sulphur dioxide is emitted by the combustion of coal and petroleum as it comprises of sulphur compounds. Oxidation of SO₂ leads to acid rain when a catalyst such as NO₂ is involved. The economy will be imbalanced if the air quality level is bad as it effects the wellbeing of the person. Implementation of effective models for measuring air quality is very necessary and those models should gather information on concentrations of air contaminants and it should give the estimation of air pollution in each region. The analysis, tracking, monitoring and prediction of air quality has become a significant field of research.

Air pollution in India is a major concern. Fuel combustion, adulteration, heavy traffic, industrial release of burnt gases etc are also affected due to Asian Brown cloud which is caused by air pollution. Biomass combustion is a prime producer of greenhouse emission which leads to change in climate. In 2009, according to International Energy Agency reports it is seen that India produces about 1.4 tons of gas per person and worldwide average released per person is 5.3 ton. India was third highest emitter of CO₂ in 2009. India alone contributes up to 5% carbon dioxide emission in the world.

Several environmental scholars have contributed their work to this topic using traditional methods. Air quality is influenced by time, place and other variables. Now-a-days big data analytic methods have been used to analyse assess and forecast air quality owing to the advances in big data technologies. Effective modelling methods to classify and verify data obtained from various sources on air emissions is one of the key challenges for researchers. There is an assumption that past is a strong predictor of future. So, different approaches such as ‘Climatology’ has been used for the prediction of air quality. Usually these methods are used to estimate air pollution level concentrations when it exceeds the threshold levels. So, more progress is still required in this area for prediction analysis. Models fail to substantially estimate the emission rates due to insufficient data.

An air quality index (AQI) is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. Public health risks increase as the AQI rises. Different countries have their own air quality indices, corresponding to different national air quality standards.

In the previous researches the implementation of some of the Machine Learning algorithms has already been done. Although Machine Learning and Deep Learning techniques has been implemented but none of them have compared their results as one research under the same constraints and with the same data.

1.1 Hypothesis

Machine Learning models would work better in accordance with accuracy, precision and recall when compared to Deep Learning model such as Artificial Neural Network.

1.2 Research Questions

- Which algorithms will work better in classification; Traditional or Deep Learning algorithms?
- How will Air Quality Index prediction help people.?
- How can we apply this work in real-time?

Literature Review

2.1 Introduction

The key goal of this research is to use machine learning algorithms such as Artificial Neural Network model (ANN) for air quality identification. Predicting air quality using multiple techniques has been proposed and implemented by many researchers. These approaches comprise all conventional methods as well as modern machine learning and deep learning algorithms. So, the discussion here will be about different types of algorithms developed and implemented to predict and identify the air quality by the researchers over a period of time.

2.2 Air pollution monitoring and prediction using Machine Learning and Deep Learning Techniques

Machine Learning provides good approaches for prediction and analysis of air pollution. Long Short Term Memory performed well on data that was collected from multiple web-based resources. Artificial Neural Network and Logistic Regression did not predict PM_{2.5} with good accuracy. But the Long Short Term Memory problem concerns the slow process of training compared to other algorithms (Ziyue Guan, 2018).

In complex learning tasks Deep Learning techniques are used as strong tools to show outstanding efficiency. Predicting the urban air quality is now a big alternative to reduce its adverse effects on human beings. Urbanisation and enhanced lifestyles have greatly increased the air emission in urban areas. Meteorological features pertaining to the atmosphere is used for air quality prediction as this data is easy and readily available and accessible for metropolitan areas (Chavi Srivastava, 2018). Throughout this study, numerous Machine Learning models were used to predict the air contaminants such as Stochastic Gradient Descent (SGD) Regression, Linear Regression (LR), Random Forest Regression (RFR), Decision Tree Regression (DTR), Multi-layer Perceptron (MLP) or neural Networks, Support Vector Regression (SGR), Adaptive Boosting Regression (ABR) and Gradient Boosting Regression (GBR). Compared to all the Machine Learning techniques used, Multi-layer Perceptron (MLP) provided good accuracy with less errors. The authorities can take effective steps with the help of this model and provide the details to the general population for their safety. The performance of this model is limited as the dataset used consists data of shorter period. Meteorological variables employed were less so it impacted the performance of the device.

Chemical emissions and in particular air contamination have attracted considerable attention from many decades due to the reality that factories produce more and more extremely dangerous contaminants. To predict the levels of two pollutants CO_x and NO_x non-linear Auto Regressive model dependent on Artificial Neural Network was used. The dataset used had meteorological variables such as wind direction, wind speed, temperature, relative humidity and levels of pollutants in the petrochemical plants. To increase the forecasting efficiency and precision of computational approximation, Meteorological factors with concentration of contaminants was used (Nadjet Djebbri, 2017). This is an effective method to predict the level of pollution but the downside is that unclear pattern and wide variations of air pollution influence the efficiency and accuracy of the model.

Multiple neural networks were used to predict air quality for 2 days (Ping Wei Soh, 2018). These neural networks included Convolution Neural Network, Artificial Neural Network and Long Short Term Memory to extract spatial-temporal relations. Data relevant to elevation space with meteorological evidence was used to extract the effect of terrain. Several patterns were derived from different locations which was extracted from the association between neighbouring locations. The first hour prediction was improved and Long Short Term Memory was added. Convolution Neural Network was used for longer period prediction. Disadvantage of this model is that when all the locations' data was included it generated more noise and thus reduced the prediction performance.

The prediction performance of the model decreases when the noise in the model is increases. In this paper (Sankar Ganesh, 2017), the model is built to predict the air pollutants based on Multiple Linear Regression and Support Vector Regression. When compared to other models, these two models have the good accuracy and precision to monitor the air quality. The air quality monitoring is done by using environmental data of different cities. Gradient descent, stochastic gradient descent and mini-batch gradient is computed by using Multiple Linear Regression and Support Vector Regression. The quality of air depends on the concentration of following pollutants, NO₂, CO, O₃, PM_{2.5}, PM₁₀ and SO₂. Support Vector Regression has the highest precision when compared to other regression models. Since the data samples are separated in Support Vector Regression, it reduces the generalization error and helps to increase the accuracy.

Effective solutions to monitor air quality draft in several big cities can be modelled using Support Vector Regression (SVR) and Multiple Linear Regression (MLR). The one

disadvantage of Support Vector Regression is that it does not give much accuracy when large dataset is used. To overcome this issue Random Forest algorithm is used (Masih, 2019). Random Forest algorithm takes the emission and meteorological parameters as inputs. The prediction of concentrated levels of NO₂ in atmosphere was done by Random Forest algorithm. The outcome of this model was contrasted with Support Vector Regression model. It is noticed that Random Forest predicts NO₂ concentration levels better than Support Vector Regression model.

For managing air quality index in large cities, neural network models can be used as it is one of the powerful methods with high recall and less error (Kaminski, 2008). This can be done by investigating the possibility of designing a prognostic device for managing air quality in large cities. Incorrect prognosis rate was 1.4 percent in training and 1.9 percent in testing sequence.

A model for the production prediction of inorganic airborne pollutants was proposed for the danger region and another area from Contasta (Barbes, 2009). The prediction model had data on NO-NO₂-NO_x, H₂S-SO₂, CO₂ and PM₁₀. PM₁₀ is nothing but the particles with aerodynamic diameter which is 10µm. A slight cumulative absolute error of 0.42 was generated by contrasting the outcomes from the actual measured evidence from the urban area and the outcome of calculated values. This showcases the efficiency of the suggested model which predicts the concentrations of airborne contaminants.

There was a high correlation between results expected and reported for all contaminants when Convolution Neural Network was used. The range was between 0.54 and 0.87. The Convolution Neural Network model could predict SO₂ levels with good accuracy compared to PM₁₀. From the results it was found that the toxin concentrates were predicted effectively in winter (Sahin, 2011). Convolution Neural Network model can estimate the concentrations of air pollution for the days when quality data is not recorded. This is done by the new method found in Convolution Neural Network which is used for daily forecasting.

Air quality predictions in metropolitan environments using neuro fuzzy controller which was applied on the historical evidence for forecasting O₃ was implemented (Soni A, 2012). The implemented model mainly targets on the pollution induced by vehicle in urban areas and the dust accumulated in the vicinity of Jabalpur.

Artificial Neural Network was used to predict atmospheric air temperature on daily and monthly basis. For Feed Forward Network and Elman Network the input data was mean, min and max atmospheric air temperature (Afzali, 2012). The evaluation of the results was done using RMSE (root mean squared error) and MAE (mean absolute error) variables on both networks showed Artificial Neural Network is a suitable model for atmospheric air temperature predictions.

Based on the current and previous relevant information a systematic Machine Learning model was defined for interpretation, analysis and decision making for further study. Mean square error is used for performance evaluation in both testing and training levels of the model. When Artificial Neural Network and mentioned MLP network is compared, then it provides the evidence that Artificial Neural Network is preferably better with reasonable precision and efficient for CO pollutant estimation (Mahmoudzadeh, 2012).

Effectively Artificial Neural Network model can be implemented for decision making and problem solving for efficient control and management of the environmental pollution. (Azid, 2013) With the data seven Malaysian air monitoring stations Artificial Neural Network model was implemented to predict Air Pollutant Index (API). The implementation of Principal Component Analysis (PCA) was focused by Peninsular Malaysia to forecast and analyse Air Pollutant Index.

Sequential issues with forecasts can be effectively executed by Long Short-Term Memory (LSTM) networks. The data was collected from CPCB in the Indira Gandhi International Airport and sufficient set of tests was performed to pick the best functionality available and utilization of Talos python package was done to execute hyperparameter optimization. PM2.5 concentration was estimated using deep neural network model. Evaluation of performance is done by using Root Mean Squared Error (RMSE). RMSE of less than 1.2 was obtained by the proposed model which was the highest accuracy of prediction reported till date (Pratyush Singh, 2019).

Neural network dependent air quality prediction model was evaluated in the traditional coastal town of Macau using five years meteorological data that was collected at the ambient air quality monitoring station (K. I. Hoi, 2010). The efficiency of Artificial Neural Network model was progressively increased by rising the hidden neurons in the phase of training. But

it failed to be responsive in the prediction phase. Thus, it can be noticed that the efficiency of the Artificial Neural Network model was due to the overfitting of the data.

Alternatively, various numbers of hidden neurons have been examined for the posterior PDF of the parameter vector relative to the training dataset. Appropriate parameter vector could be found by Levenberg-Marquardt backpropagation algorithm. But this parameter vector may include unnecessary parameters and it fails to locate the most optimal parameter vector. When the model class became complicated the parametric space was not universally recognizable. Whereas, the parametric space for a basic Artificial Neural Network model was universally identified. Hence, more complicated MLP model that fits the data is not better than a basic model.

Large scale optimization algorithms are used to predict the concentration of air pollutants. Training a model on big data can be effectively done by machine learning. Forecasting of hourly concentrations of air emissions are confined to simple regression models in most of the prior studies. Prediction of hourly air pollution is done by Multi-Tasking Learning (MTL) on the grounds of previous days' meteorological data (Dixian Zhu, 2018). Thus, a suitable model with different regularization techniques can be found. A convenient regularization technique is proposed by using prediction models of every hour and then it is compared with some standard regularizations for Multi-Tasking learning. Some of the regularizations for Multi-tasking learning includes Frobenius norm, $\ell_{2,1}$ -norm and, nuclear norm. Improved outcomes are provided by the successive hour-related regularization compared with the conventional regularization and regression models.

The regularized Multi-Tasking Learning is the solution approach which uses specialized optimization algorithms. The focus is on reducing model uncertainty and increasing the efficiency. This can be done by decreasing the parameters of the model and by the usage of systematic regularizer respectively. The findings indicate that the performance of the predictions can be improved if regularization is done by implementing prediction models for two hours to be close (Dixian Zhu, 2018). The light formulation suggested produces much greater efficiency than the prior model formulations. Advanced optimization methods are essential for enhancing integration of optimization and it will accelerate the training phase for big data.

Considering the growing attention paid to the environmental governance in the recent times, a prediction model based on Long Short Term Memory is used to predict Air Quality Index (BARAN, 2019). Data from the department of Environmental Protection is utilized to predict the Air Quality Index. The data comprises the values of PM_{2.5}, PM₁₀, SO₂, NO₂, CO, O₃, wind direction and temperature which was recorded in the department. Environmental model for predicting air quality was implemented after analysing the background, state of progress, technical features and the challenges found during the monitoring process. Finally, after the prediction model is implemented and when the error of prediction output is studied, it is found that LSTM can predict the air quality index efficiently.

Time series data analysis is done by LSTM-Kalman model. The special memory function of LSTM is used to store the details collected in the pre-order records with long term and short term characteristics. LSTM is then used to extract the corresponding time series of the prediction problem in post processing modified and revised predicted value is obtained after Kalman Filter changes the simple time data series produced by the LSTM computation.

The LSTM recurrent neural network is used as a standard model for prediction and it is also used to summarise the difference of values from past records. Using Kalman Filtering algorithm the data is modified depending on the predicted values of LSTM recurrent neural network (Xijuan Song, 2019). Observation factors in the air quality data set is used to train the LSTM-Kalman model and test the efficiency of LSTM-Kalman model. When RMSE (Root Mean Squared Error) and R-squared values of LSTM-Kalman model predictions is compared to the LSTM model predictions, the findings indicate that LSTM-Kalman model gives more accurate predictions when compared to LSTM model.

Tree based learning models was designed and Principal Component Analysis was conducted on meteorological datasets with which urban air quality of Lucknow was predicted and identification of sources of air pollution was done (Anon., 2013). According to Principal Component Analysis automobile emissions was identified as the main source of air pollution. Throughout the winter and summer, the air quality index indicated more pollution. Some models were developed and their predictive efficiency was analysed considering statistical parameter and it was compared with Support Vector Machines. The tree based ensemble models which were compared are Decision Tree Boost, Decision Tree Forest, and Single Decision Tree. In Regression and Classification, the Decision Tree Forest and Decision Tree Boost models performed better than Support Vector Machine due to the boosting algorithm

incorporated. These models have predicted urban air quality almost accurately so it can be used as powerful techniques to manage and control environmental pollution.

Countries are experiencing severe air pollution due to the expansion in urbanization and industrialisation. This is affecting human health so people have raised their voice and showed their concern regarding air pollution. Existing techniques to predict air quality yielded poor performance so deep learning architecture models such as a novel Spatiotemporal Deep Learning (STDL) was tried (Xiang Li, 2016). Stacked Autoencoder (SAE) Model shows temporal stability and the training is done layer-wise. The proposed model performs better than Support Vector Regression (SVR), Auto Regression Moving Average (ARMA) and Spatiotemporal Artificial Neural Network (STANN). SAE comprises Logistic Regression and stacked autoencoder model which is used for real-value regression and unsupervised-feature extraction respectively.

Urban air quality computation includes prediction, interpolation and feature analysis. Urban air computing can help support air pollution control which creates significant social and technological impacts. All the urban air computation problems have been solved separately from the existing known work. So Deep Air Learning is proposed which provides efficient solution to interpolation, prediction and feature analysis in a single paradigm. Semi-supervised learning and feature selection is incorporated into various layers of deep learning network. To improve the performance, details relating to unlabeled spatio-temporal data is used (Zhongang Qi, 2010). Using these models, the inner dynamics of deep-network blackbox models can be exposed. Deleting unnecessary or insignificant features is done by executing feature selection in the input layer of neural network. The significance of various input features to the neural network predictions is revealed by association analysis. Experiments are done on real data sources which shows that deep air learning has better performance when compared to other models in urban air computation.

Feature similarity is utilized by KNN to determine the new data points. Depending on how similarly the data points fit in the training phase, the value to the new data points is allocated. In the industrial sector KNN is used for predictive problems. Regression as well as classification uses KNN algorithm. KNN is a lazy and non-parametric learning algorithm. In classification all the data is utilized in the training phase and a specific training process is not defined in the KNN algorithm. As the assumption regarding the underlying data isn't done by KNN it can be called non-parametric learning algorithm. Interpretation of the KNN algorithm

is really easy and clear. KNN algorithm does not assume anything regarding the data so it works well on non-linear data. It is a versatile, dynamic and fairly reliable algorithm but there are far stronger guided learning models compared to KNN. KNN saves all the training details so it is computationally pricy algorithm and it requires more memory space when the number of N is very big and the predictions becomes slow. In this paper, the technique is used in the air quality forecast domain in order to predict the value of the air quality index. This index is used to categorize the pollution level and to inform the population about some possible episodes of pollution (Dragomir, 2010).

2.3 Air pollution monitoring and prediction using Internet of Things (IOT)

Smart cities (SCs) help to raise awareness, interactivity, efficiency and help in the perception of a situation or fact (Ibrahim KOK, 2017). Interrelated computing systems with sensors that has the capacity to transmit data across a network without human contact is known as Internet of Things. The Internet of Things device tracks air quality by sensing pollution and predicting the Air Quality Index. It helps to handle and monitor fresh air inflow in an indoor air management system. Many communities cannot escape the issues caused by the air quality problems due to the massive shift in the environmental patterns. So, Air Quality Index can be used to show people how good or unhealthy is the air quality. Using Deep Learning concept that can be extended to Internet of Things (IOT) data in smart cities called bipartite is the key innovation of Ibrahim Kok. Then Long Short Term Model is implemented to predict air quality which will make it simpler to address potential air pollution challenges in smart cities. Better predictive precision is achieved with just a basic network structure using Long Short Term Memory prediction technique.

An Internet of Things based air prediction technique using Recurrent Neural Network (RNN) was proposed by Temesegan Walelign Ayele. Recurrent neural network is a machine learning algorithm used for IOT based air prediction model. The real time humidity and digital temperature is collected from the dataset and it gave a better accuracy with less number if training cycles. DHT11 sensor helps in the data collection and produces digital temperature and humidity. TensorFlow backend was carried out to this implement this model. From the findings, it is clear that the training phase takes less time and delivers good accuracy (Temesegan Walelign Ayele, 2018).

Background

3.1 Machine Learning and Deep Learning process

Machine Learning is a field when Artificial Intelligence can be applied on any data. Here it is applied on data collected from the environment to design a model. One of the explanations why Machine Learning was preferred to predict air quality index was the adaptability function of Machine Learning.

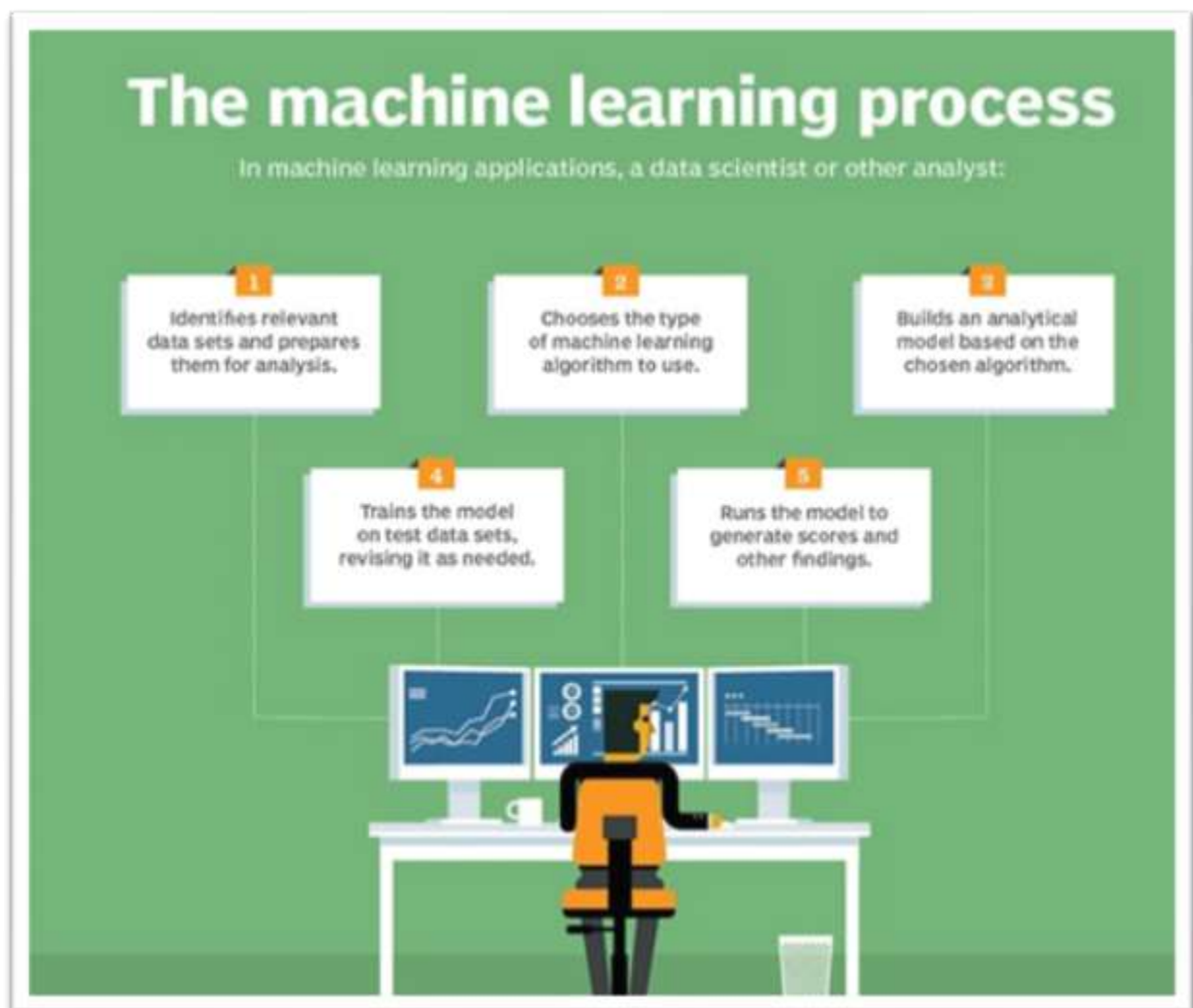


Figure 1: Machine Learning Process

Deep learning plays a major role in data science. Vast amount of labeled data is involved in Deep learning algorithms. Deep learning algorithms helps in extraction of higher-level features from raw data utilizing several layers. Feature extraction problems can be tackled by Deep Learning. Deep learning techniques are smart enough to learn to look for the right

features on their own with really no programmer guidance. Fundamentally, deep learning resembles the way our brain operates. Deep learning produces improved outcomes by taking immediate decisions and anticipates the outputs of any system on the basis of the dataset.

A complex abstraction can be classified by constructing a hierarchy where in the current abstraction level is developed from the information that is derived from the previous level of hierarchy. The input data goes through a non-linear transition to generate a logical model as output. Iterations will not stop until appropriate accuracy is achieved. The word deep in Deep Learning is motivated from the amount of processing layers the data passes through to achieve good accuracy. It has an essential aspect in analysis of data, predictive modeling and neural networks. Deep Learning resembles the way human acquire certain knowledge and take decisions. Automation of predictive analytics can be done by Deep Learning. Standard Machine Learning techniques are linear where as Deep Learning models are piled into a hierarchical structure of rising complexities.

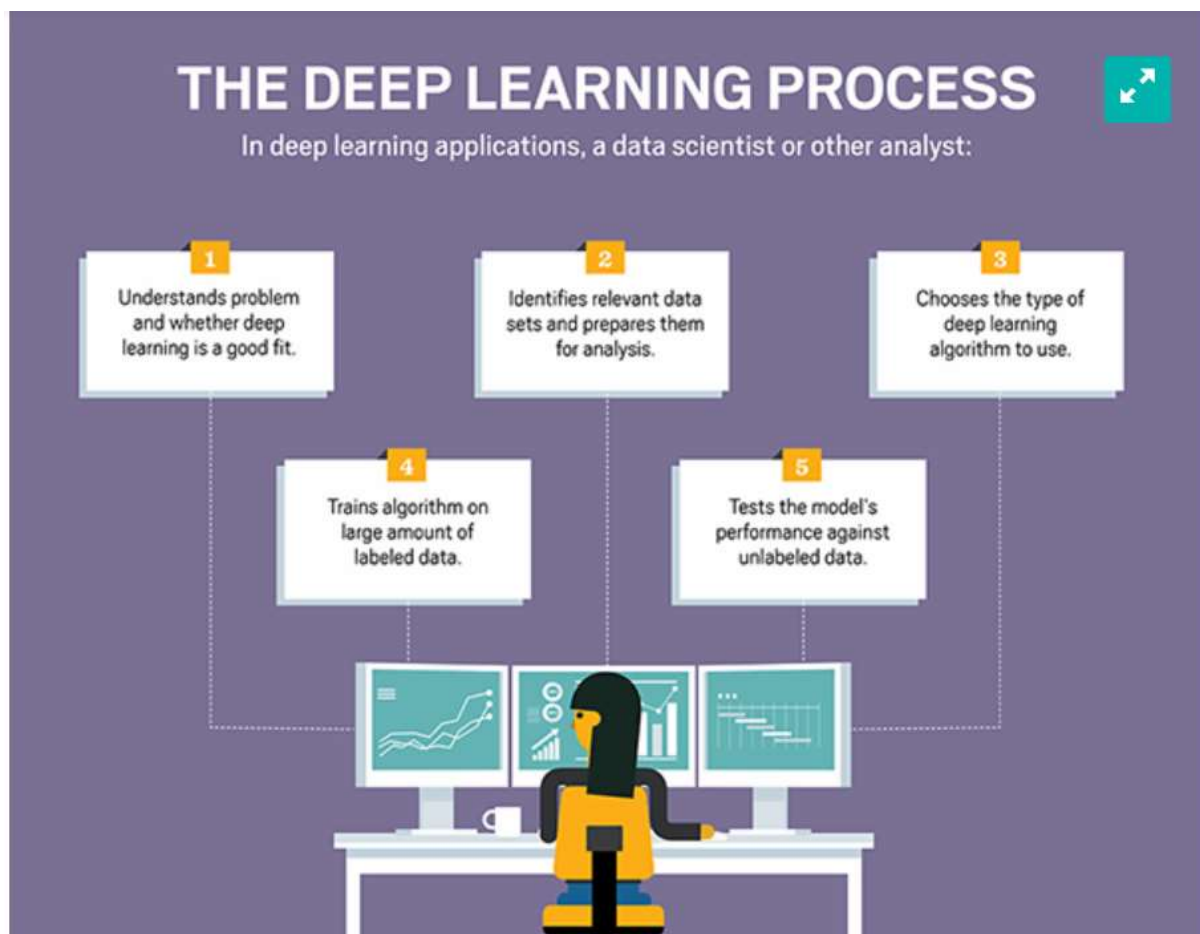


Figure 2: Deep Learning Process

3.2 Comparison between Deep Learning and Machine Learning Techniques

The learning phase in the conventional Machine Learning is supervised and the programmer should be highly precise while giving the instruction. The success and accuracy rate purely depend on the programmer's ability to specifically identify the feature set. Whereas in the Deep Learning algorithms the feature set can be built by itself without any supervision. Deep Learning algorithms are not only quicker but also more accurate and reliable.

Before the period of big data and cloud computing it was very tough to achieve good accuracy as it needs exposure to large volumes of training data. Deep Learning has the capability of generating efficient predictive model from large amounts of unstructured and unlabeled data.

To build efficient Deep Learning models various specific approaches may be used. Deep Learning models learn from the previous layer hierarchy observation which can be a major limitation. So, this implies that if the Deep Learning algorithms are trained with limited amount of information or if the data is from one source which does not fully represent the wider functional area then the model cannot be fully reliable as it does not learn in a generalisable manner. Machine Learning and Deep Learning can be distinguished from each other by the way it works and solves different problems. Domain specialists are required in Machine Learning whereas Deep Learning eliminates the necessity for domain expertise.

The training time required for Deep Learning algorithm is more compared to Machine Learning algorithms but again the opposite happens during testing i.e., Machine Learning algorithms requires more time when compared to Deep Learning algorithms. The test duration rises with the size of data.

Deep Learning demands high performance, expensive Graphics Processing Unit (GPU) which is not necessary for Machine Learning algorithms. When the information or data is limited then Machine Learning algorithms are suggested. The programmers prefer conventional Machine Learning over Deep Learning due to its high interpretability, usability and its analysis capability. Deep Learning is preferred when there is vast volumes of information, need of domain comprehension for feature examination or complex issues.

3.3 How does Artificial Intelligence help improve air quality?

Technology has been a facilitator for helping combat air pollution. It enables accurate calculation, identifies root causes, creates strategies, estimates and uses analysis to solve problems. It can be used by the government or the other organisation to automate some air quality predicting models which will help reduce the impact of air pollution on the society.

Air emissions can be handled efficiently due to air quality prediction. So, it can be employed to track, control, identify, predict, forecast, and manage air pollution.

Air quality prediction is a base for numerous things. When someone thinks about air pollution then Delhi, Beijing or Shanghai comes to mind. Only one out of ten people inhale unpolluted air according to the information given by the World Health Organisation.

Air pollution in the metropolitan areas in India has a negative effect on quality of life and it has led to substantial decline of atmospheric air quality. It can help war the citizens specifically rates of air pollution by mapping the existing emissions rate with regard to personal safety measures. Regular air pollution prediction reports can help spot safe plans.

Machine Learning and Artificial Intelligence have developed rapidly over the past few years. The decision-making capability of Artificial Intelligence has influenced almost all aspects of our life. Artificial Intelligence makes its own choices rather than working purely on the instruction provided by the programmers. Beginning from the start-up firms and winding up with major technology providers, Machine Learning and Artificial Intelligence have both been core fields of focus.

Research Methodology

4.1 Data

The dataset contains air quality data from various stations across multiple cities in India from 2015 to 2020. The cities include Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam, Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru. The concentration levels of 12 gases has been recorded and has been made publicly available by the Central Pollution Control Board. The dataset contains 14705 records.

4.2 *Software Used*

Python is a very powerful high-level and structured language with simple syntax and flexible semantics. Compared to other languages Python is very simple. Coding with Python is very convenient and requires less effort as there are many inbuilt libraries.

It is derived or developed from the language that that people use to interact in day to day life. Anyone can attempt to understand this Python code without the pressure of digging loads of computer code. Python employs indentation instead of braces to the code looks well organised and structured. Even with a smaller code good efficiency can be achieved. Python is open source and it supports many libraries. The main libraries used to carry out the research are:

Pandas

This module is used for data analysis. It should be manually installed as it is not bundled with Python. It gives descriptive and adaptable intelligent mechanisms such as information structures that enables data management and its control simple and clear across various objects. One such structure in data frame.

Keras

Keras is a Python library built on top of the TensorFlow framework. It is flexible and convenient to learn. Keras is created to emphasize on recognizing the methods of deep learning.

For quick prototyping and guided research keras is used. Keras has a simple interface so it is easy to construct. The errors are clearly mentioned. It has few limitations as it combines configurable structure squares custom structure squares built can be used to articulate new ideas for research.

4.3 Data Preparation

4.3.1.1 Correlation Matrix with Heatmap

Correlation shows whether the characteristics are relevant to each other. Correlation can be positive or negative. Heatmap makes it easier to recognise which features should be selected to train the model. Seaborn library is used to plot the correlation heatmap.

4.3.1.2 Data Pre-processing

The method which is used to convert raw data to an understandable and logically usable format is known as data pre-processing. The documents and information found as data in the real-world is partially incomplete, ambiguous, inconsistent and may contain many errors or null values. So, an established way to solve these matters are data-pre-processing. It is very important to accurately and consistently predict air quality to shield humanity from air contamination. This will also help in air quality regulation and as a consequence it can obtain major technological and social impacts.

4.3.1.3 Feature Selection

To reduce over-fitting and to enhance precision or accuracy it is very necessary to pick the right features. This is the first crucial phase which will help improve the accuracy. Insignificant or partially prominent features can have a detrimental effect on the efficiency of the model. It partially reduces the computational complexity and time. Minimizes the chances of overfitting. Allows more accurate prediction as the judgements won't be based on noisy data or outliers.

Modelling

5.1 K-means clustering

5.1.1 Overview

Clustering can be termed as the most essential unsupervised learning method. It involves in finding a pattern or a structure in a set of unlabelled data. The concept of clustering in grouping objects whose elements are related in some way. Clustering algorithm should satisfy some criteria like it should be able to handle noise and outliers, large dimensionality, interpreting capability etc. There is a wide range of clustering problems like: Present clustering methods fail to resolve all the requirements simultaneously and accurately. Due to time complexity it is difficult to work with high dimensionality and big data. The output of the clustering algorithms can be portrayed and viewed in various forms.

K-Means algorithm has an iterative approach. In this algorithm the data is clustered into k set of non-overlapping sub-units. The distance between the inner point of the cluster will be kept as close as possible while attempting to maintain distance between the clusters. K-Means algorithm assigns data points to the cluster in such a way that the total squared distance between the data point and centroid of the cluster is reasonably minimal. Less variation should be seen within a cluster. More variability in the cluster means the data points inside the cluster are non-identical or heterogeneous.

This algorithm requires a sample training dataset and the k value to be specified. The two methods which can be used to figure out the number of clusters or the K value are Elbow method and Purpose method. To understand elbow method first WSS should be known. WSS can be abbreviated as “within the sum of squares”. If a graph is drawn between the total WSS and the number of clusters K then a curve resembling a human elbow is seen. So, this elbow point helps to find the best possible number of clusters. The magnitude of WSS shifts very gradually after the elbow point. In purpose-based method the data is clustered on the basis of various criteria and then its performance is evaluated.

K-Means clustering will do well with special clusters. It forms tight clusters and enhances accuracy. Does not require more computational time. If there is strongly overlapping data then K-Means clustering does not yield good results. The variables can be unevenly weighted by Euclidean distance. It is always not possible to randomly choose the centroids to produce

positive outcomes. Outliers cannot be tackled by K-Means Clustering. There is absence of consistency and it does not function well on non-linear data.

5.1.2 Implementation

Since this dataset doesn't have any label so K-means Cluster is used to find the class for this dataset. Number of clusters i.e., k value is given as 5. Elbow method was used to determine the value of k. Maximum iteration given is 100. After the dataset is clustered it is saved into an Excel file.

5.2 Artificial Neural Network

5.2.1 Overview

To understand what is ANN it is very necessary to know what is a neural network. Neuron is a fundamental unit of nervous system thus a network of this type is called neural network. In simple words, if the strength of neuron network in a brain were to be imbibed in an abstract group of items that can replicate the same action then it can be called as Artificial Neural Network.

ANN can be used to do a specific set of tasks like classification, clustering etc. ANN can be interpreted as weighed guided graphs. In ANN the nodes are generated by artificial neurons and the relationship between their input and outputs can be expressed by weight-directed edges. The input signal for ANN is taken from the outside world as a pattern or a image as a vector and is represented mathematically by $x(n)$, where n is the number of inputs.

The weights are nothing but the information used by ANN to tackle an issue. In general, it reflects the neural interconnection. The weighted inputs are added within the processing machine, if the total equals zero then a bias is applied to render the outcome as a value other than zero. The input value to the bias is one. A threshold is set to maintain the answer within the limits as the total of inputs weight can vary between zero and positive infinity. This output is then passed through an activation function which includes a group of transfer functions used to get a required output.

Artificial Neural Network comprises of a significant number of artificial neurons which is organised in a sequence of layers.

The different layers present in the ANN are 1) Input layer, 2) Output layer and the 3) Hidden layer.

Input Layer

The actual learning on the network takes place in the input layer where the data is obtained from the outside world.

Output Layer

This layer includes units that responds to the input data fed to the device where we can notice if the system has learned the task or not.

Hidden Layer

This layer is concealed between the input and the output layer. The function of this layer is to convert the input data to some form of usable format which can be used by the output layer. After each iteration the ANN gets updated which helps maximum learning.

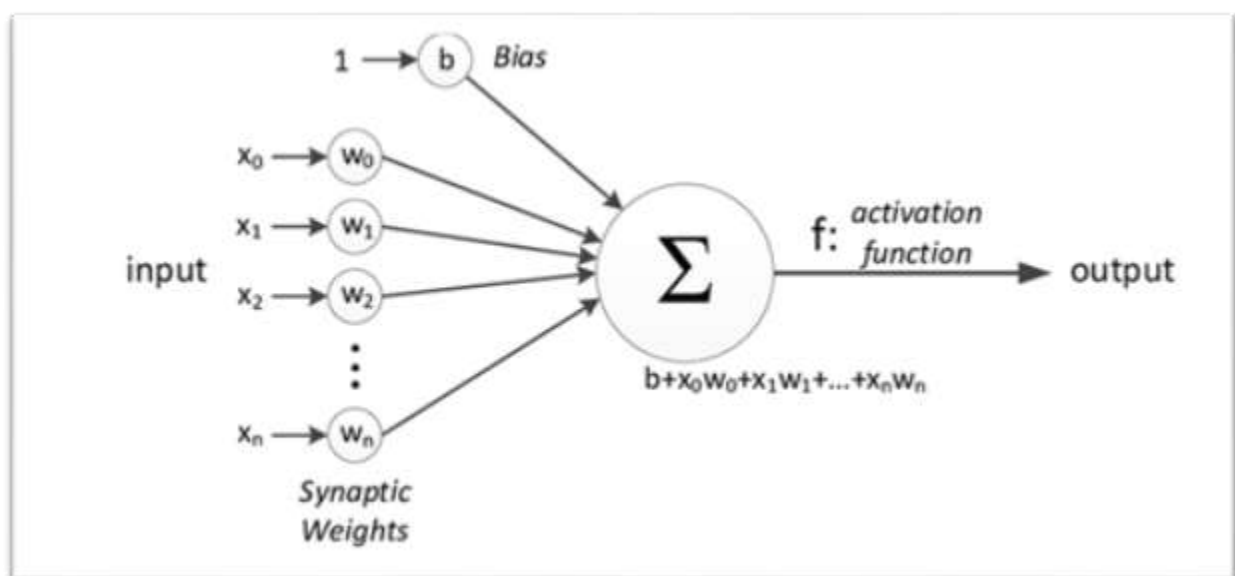


Figure 3: Schematic Representation of a Perceptron

Perception is a single layer of neural network which provides a single output.

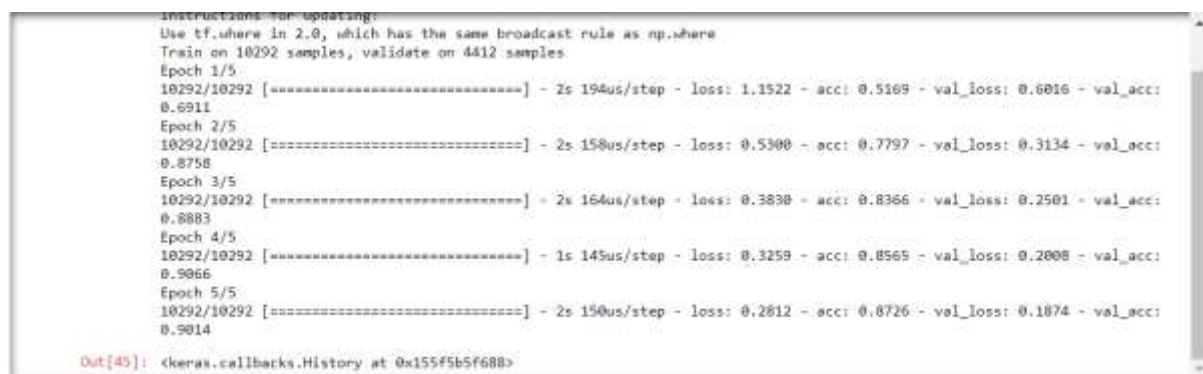
x_0 and $x(n)$ reflects the independent input values which is multiplied with synaptic weights i.e., w_0 to $w(n)$. Weight indicates the power of the node. b is the bias value which is used to switch the activation function high or low.

In general, the synaptic weights and bias are added and then fed to the activation function to produce a result which is given as the output.

5.2.2 Implementation

X and Y values are initialised. Y consists of dependent variables which can be numerical or categorical and X consists of dependent variables. Here X has the concentration levels of gases and Y has Class.

Epochs specifies number of times dataset should be iterated. So, 5 is the value given for Epochs. Seed makes components of algorithm to be dependent on random number generator seed for randomization. The seed value given is 121. Several types of activations are supported for hidden layers.



```
Instructions for updating:
Use tf.nn.softmax in 2.0, which has the same broadcast rule as np.softmax
Train on 10292 samples, validate on 4412 samples
Epoch 1/5
10292/10292 [=====] - 2s 194us/step - loss: 1.1522 - acc: 0.5169 - val_loss: 0.6016 - val_acc: 0.6911
Epoch 2/5
10292/10292 [=====] - 2s 158us/step - loss: 0.5300 - acc: 0.7797 - val_loss: 0.3134 - val_acc: 0.8758
Epoch 3/5
10292/10292 [=====] - 2s 164us/step - loss: 0.3830 - acc: 0.8366 - val_loss: 0.2501 - val_acc: 0.8883
Epoch 4/5
10292/10292 [=====] - 1s 145us/step - loss: 0.3259 - acc: 0.8565 - val_loss: 0.2008 - val_acc: 0.9066
Epoch 5/5
10292/10292 [=====] - 2s 150us/step - loss: 0.2612 - acc: 0.8726 - val_loss: 0.1874 - val_acc: 0.9014
Out[45]: <keras.callbacks.History at 0x155f5b5f688>
```

Figure 4: Execution Window of ANN

5.3 SVM

5.3.1 Overview

Classification as well as regression problems uses SVM which is one of the supervised learning algorithms. But SVM is mostly used for the classification problems. In SVM, every data object is plotted as a point in n-dimensional space. Support vectors are literally positioning or co-ordinates of specific individual findings or observations. The two groups are properly divided with a line or frontier in SVM classifier.

In high dimensional spaces SVM is very efficient. SVM fits decently when there is a good differentiation range. SVM works well when the total dimensions exceed the number of

items. Memory efficiency can be noticed in SVM as it uses support vectors i.e., subset of training points is utilised in the decision function.

When there is a huge dataset SVM cannot be used since the necessary training period is very high. The functioning of the SVM is poor when there is more noise. SVM does not include estimates of probability explicitly, instead they are determined using a five-fold cross validation.

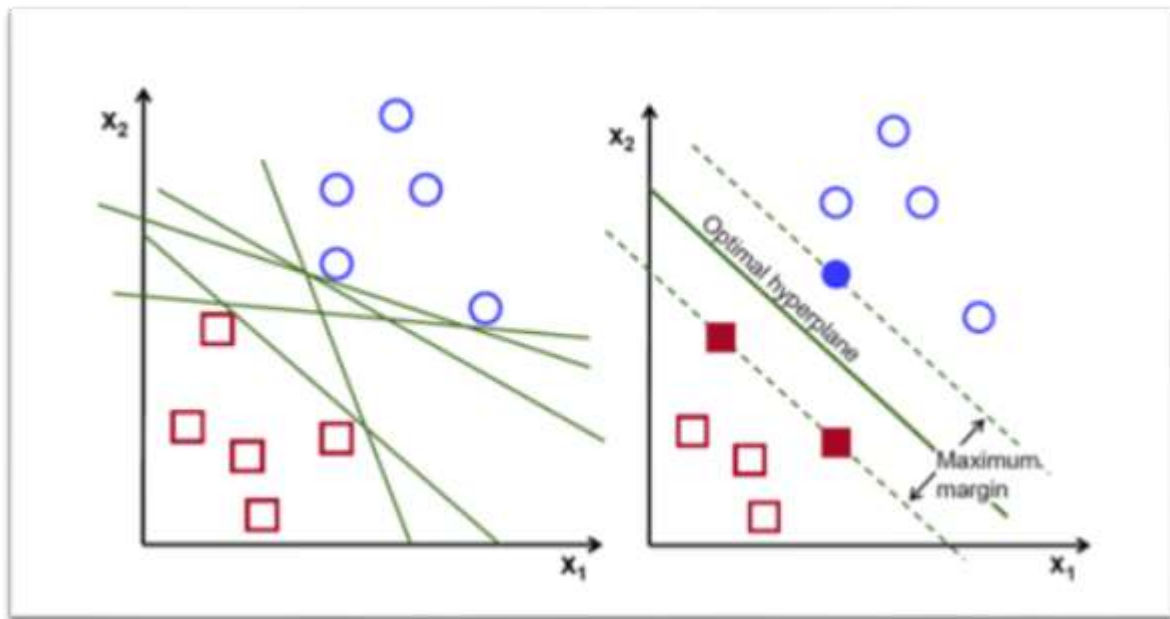


Figure 5: Depicting the hyperplanes of Support Vector Machines

5.3.2 Implementation

For implementing Support Vector Machine in Python standard libraries are imported and Scikit-Learn support vector classifier is used to train a Support Vector Machine model on the dataset. Here, linear kernel is used to fit Support Vector Machine.

5.4 Random Forest

5.4.1 Overview

Random Forest is a reliable, dynamic convenient to use machine learning algorithm that gives good results. Hyperparameter tuning is not necessary to get best results. It is perhaps one among the commonly employed algorithms owing to its flexibility and versatility which is used for both regression and classification.

The “Forest” that it creates is a series of decision trees, typically trained by the “bagging” process. The basic principle of this approach is that a mixture of models can maximize the total performance. In other words, several decision trees are developed which are combined together to render decision that are more reliable and efficient.

Same hyperparameters are seen in Random Forest as well as in decision tree. Instead of using the combination of decision tree and a bagging classifier, a simple random forest can be chosen. Similarly, Random Forest’s regressor can be utilized for any regression tasks. Significant randomness is added to the model as the trees increase. While splitting it looks for the best feature in the random subset which renders good results. Random thresholds can be used for each feature.

Difference between Decision Trees and Random Forest

There are several variations even though the random forest is a set of decision trees. In decision trees if the training data is given the input with labels and features then a series of rules will be created that could be used for predictions. In contrast to this Random Forest algorithm randomly chooses observations and characteristics to construct multiple decision trees and the average of this outcome is taken into consideration.

Overfitting is another problem in decision trees, whereas Random Forest avoids overfitting by constructing random subsets of features and with random subsets of features smaller trees are created and combined together. Based on number of trees built by the Random Forest algorithm the computational speed depends. The hyperparameters are often used to maximise the predictive ability or to speed up the algorithm in Random Forest.

Comprehending the hyperparameters are extremely easy because there aren’t many hyperparameters. The major disadvantage of this algorithm is that huge number of trees will eventually make the algorithm too sluggish and unreliable. In fact, this algorithm is easy and fast to train but little slow when it comes to making predictions. For an efficient and accurate prediction, huge number of trees are needed which in turn slows down the model. Other algorithms are comparatively better if there is a search for analysis of relationships in the data is required.

5.4.2 Implementation

The data is divided into training and test sets. The most important parameter in Random Forest Classifier is `n_estimators` which defines the number of trees in the random forest model. `N_estimator` value given is 600.

5.5 Logistic Regression

5.5.1 Overview

To understand logistic regression, it's important to know what is regression analysis. Regression analysis is used to determine relation between variables i.e., independent variable and a dependent variable. It is a form of predictive modelling methodology. To predict the result of dependent variable if a sequence of independent variables is used then it is called multiple regression. Logistic regression is programmed to complete or predict the likelihood of a conditional occurrence which is dependent on the independent variables. Conditional outcome should be categorical values which can even be binary where the only potential scenarios is zero or one. The parameters which affect the result or the dependent variables can be known as independent variables independent variables can be continuous or discrete. Multicollinearity should be observed in the model which is necessary for logistic regression and the size of the sample should be big.

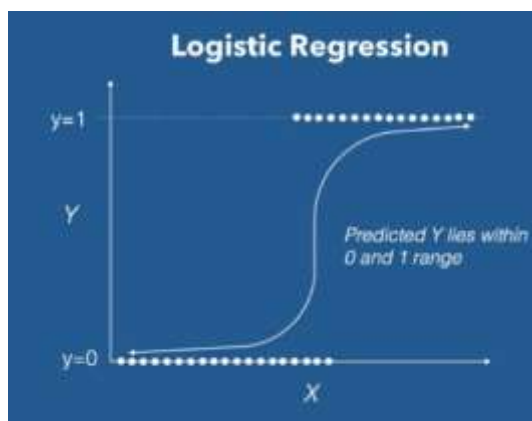


Figure 6: Depiction of Logistic Regression graphically.

In machine learning compared to other approaches logistic regression is very simple and implementation is quick and easy when the sample or input data is linearly separable then logistic regression operates best. If it is feasible to create a segment that will distinguish the groups then a dataset is considered to be linearly separable. Variable insights can be observed

clearly. Continuous values cannot be predicted using logistic regression model. Efficiency and accuracy of the model will be low if the size of the sample is small overfitting may occur if there a limited dataset as the model will not be able to identify the patterns.

5.5.2 Implementation

LogisticRegression is imported from `sklearn.linear_model`. The data is divided into training and test sets. C value given in 1.0 which is the Inverse of regularization strength and it must be a positive float.

5.6 Naïve Bayes

5.6.1 Overview

Naive Bayes classifiers rely on Bayes' theorem. It is a collection of algorithms which follow a similar concept. Each pair of characteristics or features are not dependent on each other. Naive Bayes classifiers perform really well in real world scenarios despite their seemingly oversimplified hypothesis. To approximate correct parameters very limited amount of training data is needed by Naïve Bayes classifiers.

These classifiers are incredibly fast relative to other advanced approaches. Every distribution can be individually calculated as a one-dimensional distribution, so it tends to relieve the difficulties resulting from the burden of dimensionality.

Gaussian Naïve Bayes

Gaussian Naïve Bayes is based on gaussian distribution. The assumption is that continuous values which are related to each function are distributed according to gaussian distribution. If a graph is plotted then it creates a bell-shaped curve that is symmetrical around the mean value of the feature values.

Multinomial Naïve Bayes

Multinomial Naïve Bayes is based on multinomial distribution where the frequencies indeed are represented by feature vectors which produces some events. Multinomial Naïve Bayes can be used for document classification.

Bernoulli Naïve Bayes

In this type of model, the properties are distinct Booleans representing inputs. Bernoulli Naïve Bayes is normally used for content classification activities i.e., to check if the word exists in the content or not instead of checking the frequency of the term.

The basic principle of Naïve Bayes is that every feature is not dependent and contributes equally to the result.

Naïve Bayes Classification does not need more training data. It is simple, easy to enforce and is strongly dynamic in design. It can predict probabilistic problems. Naïve Bayes Classification is very flexible as it can work on binary and other classification problems. It has the capability to manage categorical and continuous data.

High feature isolation is the most significant drawback as in the actual life it is nearly difficult to get a group of features that are totally independent. The problems of zero frequency is a major issue as it reduces the efficiency while making a prediction.

5.6.2 Implementation

Gaussian Naïve Bayes is used to predict the air quality. The dataset is divided into train and test test. GaussianNB is imported from `sklearn.naive_bayes` and it does not require initialisation of many parameters.

5.7 Autoregression

5.7.1 Overview

If linear combination of previous values is used to predict or estimate the variable value then it can be called an autoregression model. It is nothing but the regression of a parameter value against itself. It is found that resemblance or relation is observed between past and present values so autocorrelation is witnessed with such data. This helps in making a rough prediction. For example, by knowing the value of a commodity today and in past, a prediction can be made to see what will be its value in future.

5.7.2 Implementation

From the dataset monthly average value for all the gases is found and then Auto Regression is used to find next three months air quality. The monthly average value which is found for gases is given as the value of X. Now this is divided into train and test set.

Evaluation

Evaluation of machine learning model is essential task as it assists in truly judging a model. Classification accuracy is universally used parameter for measuring performance of the model, but it is not enough to evaluate model truly

6.1 Confusion Matrix

Output Assessment Methodology for classification models in Machine Learning is Confusion Matrix. The predicted outcomes of a classification problem can be described using a confusion matrix. The accuracy of the classification model can be visualised by confusion matrix as it compares the actual and predicted values. It illustrates the way the classification model gets confused when the predictions are made.

		Predicted	
		Positive	Negative
Observed	Positive	TP (# of TPs)	FN (# of FNs)
	Negative	FP (# of FPs)	TN (# of TNs)

Figure 7: Confusion Matrix

True Positive (TP): Estimated values are accurately predicted as positive.

False Positive (FP): Estimated values are inaccurately predicted as positive.

False Negative (FN): Positive values are inaccurately predicted as negative.

True Negative (TN): Estimated values are accurately predicted as negative.

6.2 Accuracy

Classification accuracy is ratio of total number of correct predictions to total number of data points used as input.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

6.3 Recall

Recall is the proportion of all the outcomes which are accurately predicted as positive divided by the sum of accurately predicted values as positive and the positive values which are predicted as negative. If the recall is high then the accuracy is good.

$$\text{Recall} = \frac{TP}{TP + FN}$$

6.4 Precision

Precision gives the accuracy of positive outcomes. It evaluates how probably the prediction of positive outcomes is accurate. The highest value of precision is one. This happens only when the classifier classifies all the positive value correctly. As the negative outcomes are ignored in precision, it is not really effective.

$$\text{Precision} = \frac{TP}{TP + FP}$$

6.5 F1 Score

The total average result of Recall and Precision is F1 score.

Results - Model Comparison

7.1 Results of Artificial Neural Network

7.1.1 Resultant Classification Report

	precision	recall	f1-score	support
0	0.95	0.99	0.97	1957
1	0.93	0.91	0.92	1464
2	0.73	0.88	0.80	650
3	0.93	0.93	0.93	137
4	0.00	0.00	0.00	204
accuracy			0.90	4412
macro avg	0.71	0.74	0.72	4412
weighted avg	0.86	0.90	0.88	4412

Figure 8: Classification Report of ANN

7.1.2 Resultant Confusion Matrix

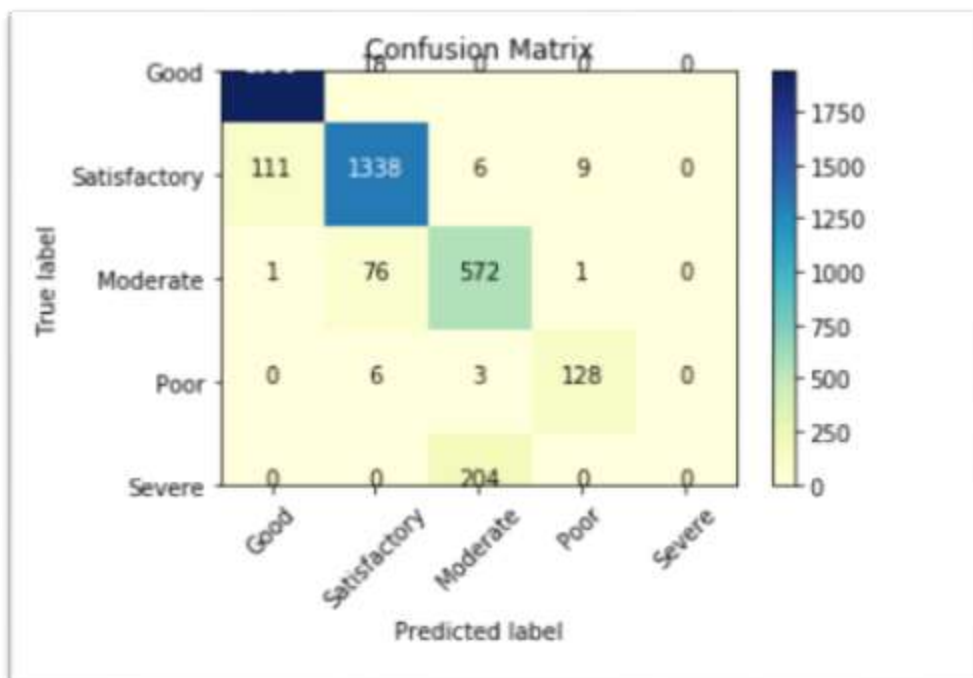


Figure 9: Confusion Matrix of ANN

7.2 Results of SVM

7.2.1 Resultant Classification Report

	precision	recall	f1-score	support
0	0.99	0.88	0.93	1957
1	0.56	0.99	0.72	1464
2	1.00	0.13	0.22	650
3	1.00	0.09	0.16	137
4	1.00	0.00	0.01	204
accuracy			0.74	4412
macro avg	0.91	0.42	0.41	4412
weighted avg	0.85	0.74	0.69	4412

Figure 10: Classification Report of SVM

7.2.2 Resultant Confusion Matrix

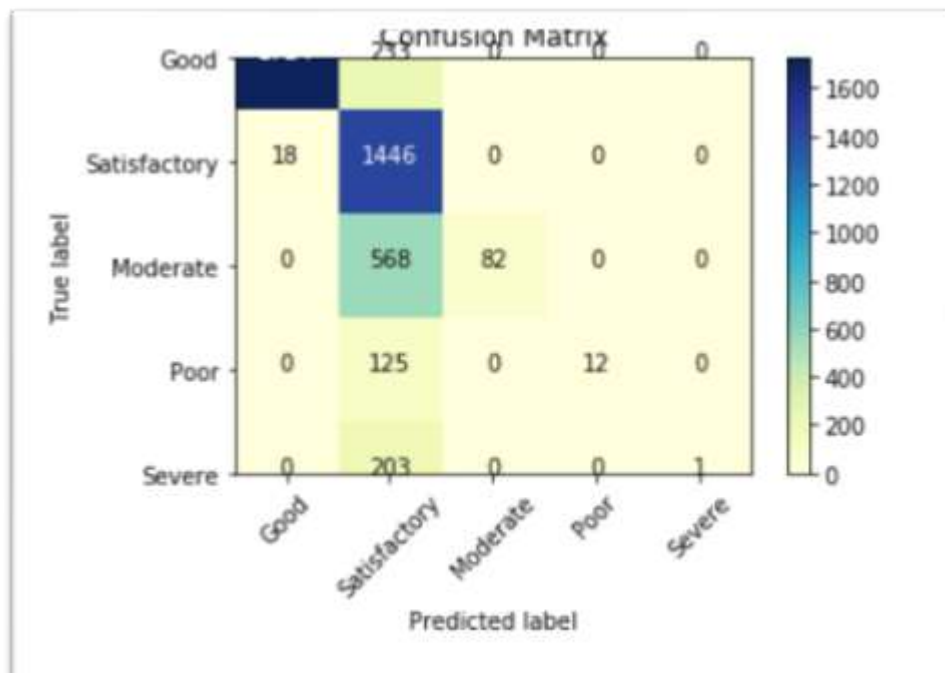


Figure 11: Confusion Matrix of SVM

7.3 Results of Logistic Regression

7.3.1 Resultant Classification Report

	precision	recall	f1-score	support
0	0.96	1.00	0.98	1957
1	0.82	0.93	0.87	1464
2	0.86	0.53	0.65	650
3	0.92	0.96	0.94	137
4	0.93	0.82	0.87	204
accuracy			0.90	4412
macro avg	0.90	0.85	0.86	4412
weighted avg	0.90	0.90	0.89	4412

Figure 12: Classification Report of Logistic Regression

7.3.2 Resultant Confusion Matrix

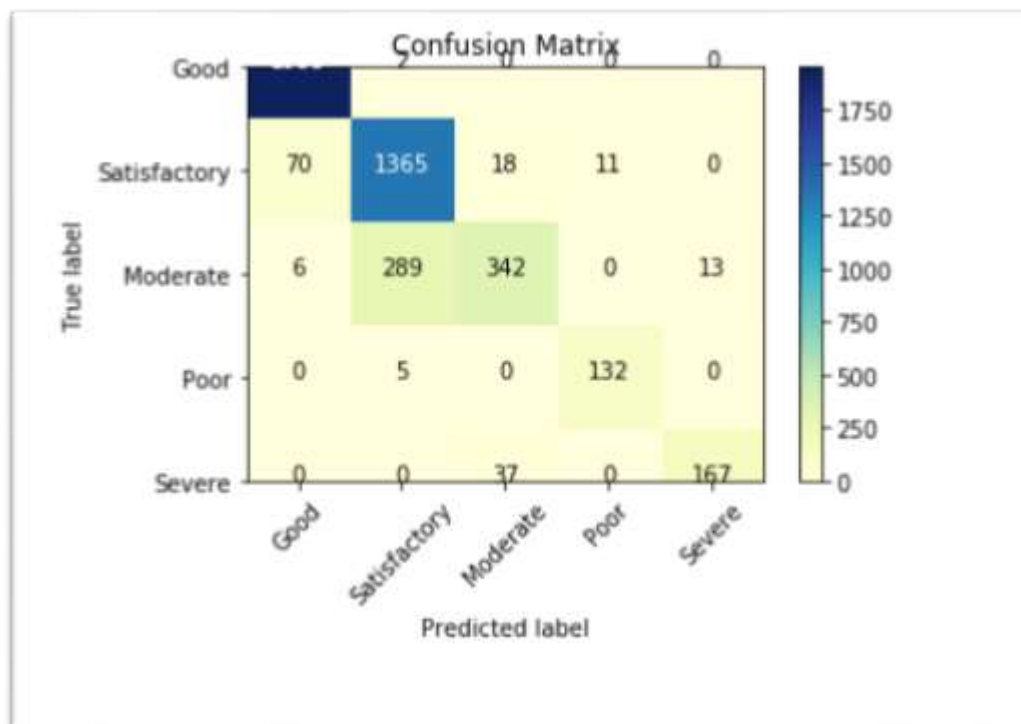


Figure 13: Confusion Matrix of Logistic Regression

7.4 Results of Naïve Bayes

7.4.1 Resultant Classification Report

	precision	recall	f1-score	support
0	0.98	0.89	0.93	1957
1	0.86	0.90	0.88	1464
2	0.85	0.90	0.87	650
3	0.74	0.93	0.83	137
4	0.82	0.95	0.88	204
accuracy			0.90	4412
macro avg	0.85	0.92	0.88	4412
weighted avg	0.91	0.90	0.90	4412

Figure 14: Classification Report of Naïve Bayes

7.4.2 Resultant Confusion Matrix

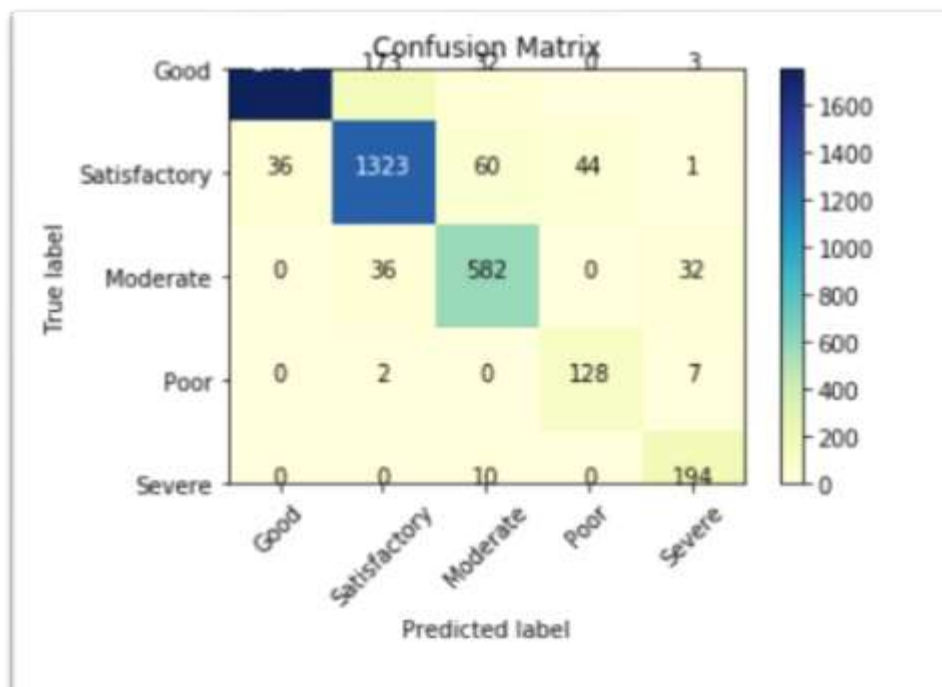


Figure 15: Confusion Matrix of Naïve Bayes

7.5 Results of Random Forest

7.5.1 Resultant Classification Report

	precision	recall	f1-score	support
0	0.98	0.99	0.99	1957
1	0.97	0.97	0.97	1464
2	0.97	0.97	0.97	650
3	0.97	0.96	0.97	137
4	0.96	0.98	0.97	204
accuracy			0.98	4412
macro avg	0.97	0.97	0.97	4412
weighted avg	0.98	0.98	0.98	4412

Figure 16: Classification Report of Random Forest

7.5.2 Resultant Confusion Matrix

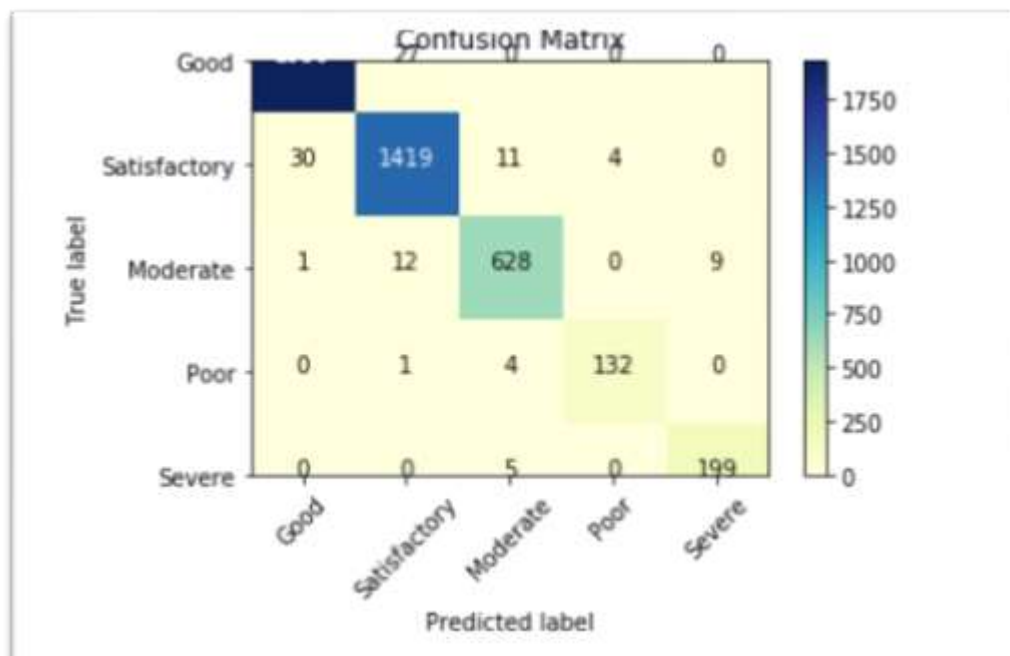


Figure 17: Confusion Matrix of Random Forest

7.6 Accuracy comparison of Machine Learning Models

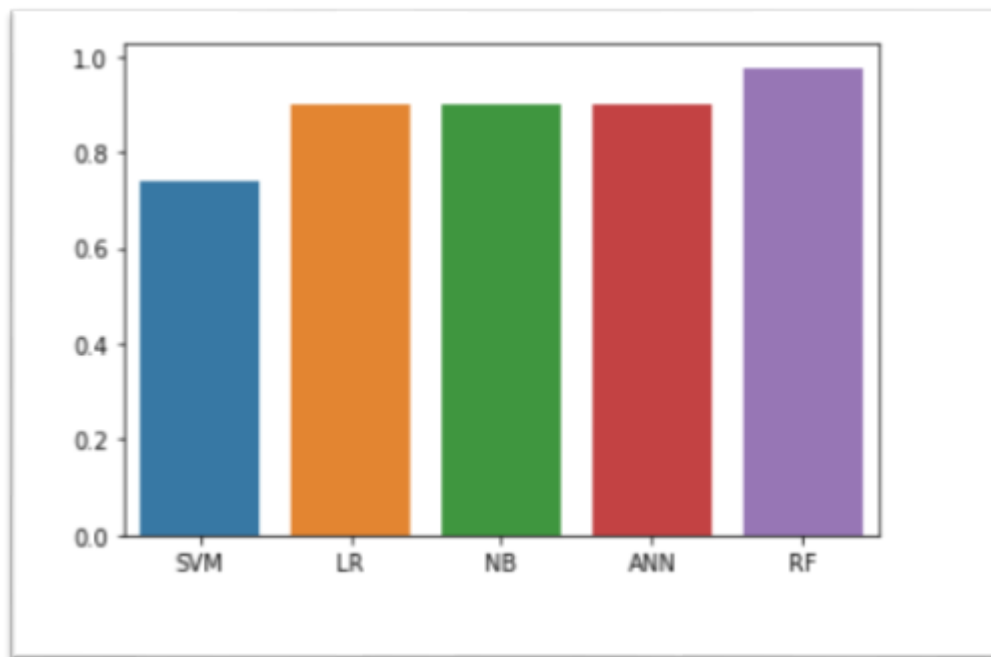


Figure 18: Comparison of Machine Learning Models

Random Forest performs well with the highest accuracy over other regression models. Models of the Machine Learning are found to be accurate when predicting concentrations of air pollutants.

Conclusion and Future Enhancements

Air pollution is a major concern so identifying and predicting the air quality will surely help in making decisions and will contribute in protecting the safety of residents. Human activities can be scheduled in a defined geographical area using air quality predictors and it also helps in reducing the adverse impact of emissions. Machine Learning and Deep Learning models were constructed for predicting the air quality in different cities of India using five years database. The model is designed in such a way that the air quality can be found for any of the cities present in the dataset. Based on average value the air quality is classified into Good, Satisfactory, Moderate, Poor and Severe. The outcomes are positive and it has been shown that the application of such algorithms is very efficient in predicting the air quality. Several statistical parameters were computed to evaluate the generalization and predictive abilities of the proposed models. These models can be used as tools in air quality prediction and management.

This model can be interfaced with the user's web application which can benefit from the work and take precautions to maintain minimal air pollution levels. Only five algorithms are tested and evaluated, lot of other tools and techniques can be applied on this data. It will help bridge the gap that existed in this model. Effective modelling methods to classify and verify data obtained from various sources on air emissions is one of the key challenges for researchers. So, more progress is still required in this area for prediction analysis as models fail to substantially estimate the emission rates due to insufficient data.

References

- Afzali, M. A. A. a. Z. G., 2012. The Potential of Artificial Neural Network Technique in Daily and Monthly Ambient Air Temperature Prediction.. *International Journal of Environmental Science and Development*., pp. 33-38.
- Anon., 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric environment*, pp. 426-437.
- Azid, A. J. H. L. T. Z. M. a. O., 2013. Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia. *Energy Journal of Environmental Protection*., pp. 1-10.
- BARAN, B., 2019. Prediction of Air Quality Index by Extreme Learning. p. 8.
- Barbes, L. N. C. M. L. I. V., 2009. The Use of Artificial Neural Network (ANN) for Prediction of Some Airborne Pollutants Concentration in Urban Areas. *Journal of Rev. Chem*, pp. 301-307.
- Chavi Srivastava, A. S. S. S., 2018. "Estimation of Air Pollution in Delhi using Machine Learning Techniques". *Institute of Electrical and Electronics Engineers International Conference, Noida india*, pp. 304-309.
- Dixian Zhu, C. C. T. Y. a. X. Z., 2018. A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. *Big Data and Cognitive Computing* , p. 15.
- Dragomir, E. G., 2010. Air Quality Index Prediction. *BULETINUL*, pp. 103-108.
- Ibrahim KOK, M. U. S. O., 2017. "A Deep Learning Model for Air Quality in Smart Cities". *Institute of Electrical and Electronics Engineers International Conference, Boston USA*, pp. 1983-1990.
- K. I. Hoi, K. V. Y., 2010. Is a Complex Neural Network Based Air Quality Prediction Model Better Than a Simple. *AIP Conference Proceedings*, pp. 764-769.

Kaminski, W. S. J. a. S. J., 2008. Application of Artificial Neural Networks (ANNs) to Predict Air Quality Classes in Big Cities.. *19th International Conference on Systems Engineering*., pp. 135-140.

Mahmoudzadeh, S. O. Z. Y. M. a. B., 2012. Carbon Monoxide Prediction Using Artificial Neural Network and Imperialist Competitive Algorithm. *Journal of Basic and Applied Sciences*, pp. 735-744.

Masih, A., 2019. "Application of Random Forest Algorithm to Predict the Atmospheric Concentration of NO₂". *Institute of Electrical and Electronics Engineers International Conference, Yekaterinburg Russia*.

Nadjet Djebbri, M. R., 2017. "Artificial Neural Networks Based Air Pollution Monitoring in Industrial Sites". *Institute of Electrical and Electronics Engineers International Conference, Antalya Turkey*.

Ping Wei Soh, J. W. C. J. W. H., 2018. "Adaptive Deep Learning-Based Air Quality Prediction Model using the most Relevant Spatial-Temporal Relations". *Institute of Electrical and Electronics Engineers*, Volume 6.

Pratyush Singh, L. N. T. C. S. L., 2019. DeepAir: Air Quality Prediction using Deep Neural. *IEEE Region 10 Conference (TENCON)*, pp. 871-873.

Sahin, A. B. C. a. U., 2011. Application of cellular neural network (CNN) to the prediction of missing air Pollutant data. *Journal of Atmospheric Research*, pp. 314-326.

Sankar Ganesh, S. H. M. S. R., 2017. "Forecasting Air Quality Index using Regression Models". *Institute of Electrical and Electronics Engineers International Conference, Tirunelveli India*, pp. 248-254.

Soni A, S. S., 2012. Application of Neuro-Fuzzy in Prediction of Air Pollution in Urban Areas. *IOSR International Journal of Engineering*., pp. 1182-1187.

Temesegan Walelign Ayele, R. M., 2018. "Air Pollution Monitoring and Prediction using IOT". *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT) Coimbatore, India*, pp. 1741-1745.

Xiang Li, L. P. Y. H. J. S. & T. C., 2016. Deep learning architecture for air quality predictions. *Environ Sci Pollut Res*, p. 10.

Xijuan Song, J. H., 2019. Air Quality Prediction based on LSTM-Kalman. *IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pp. 695-699.

Zhongang Qi, T. W. G. S. W. H. X. L., 2010. Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, p. 14.

Ziyue Guan, R. O. S., 2018. "Prediction of Air Pollution through Machine Learning Approaches on the Cloud". *Institute of Electrical and Electronics Engineers International Conference, Zurich Switzerland*, pp. 51-60.

