AIR QUALITY INDEX PREDICTION OF A SPECIFIC CITY

A PROJECT REPORT

Submitted in the partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER ENGINEERING

Under the Supervision of
Mr. Shamim Ahmad
Assistant Professor

Submitted By
Osama Khan (15BCS0071)
Prabhat Kumar Moore (15BCS0072)

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF ENGINEERING AND TECHNOLOGY
JAMIA MILLIA ISLAMIA, NEW DELHI–110025

CERTIFICATE

This is to certify that the project entitled "**Air Quality Index Prediction of a Specific City**" by Osama khan (15BCS0071) and Prabhat Kumar Moore (15BCS0071) is a record of bonafide work carried out by them. In the Department of Computer Engineering, Jamia Millia Islamia, New Delhi under my supervision and guidance in partial fulfillment of Engineering in Computer Engineering Jamia Millia Islamia in Academic year 2019

prof. Tanveer Ahmad
(Head of the Department)
Department of Computer Engineering
Faculty Of Engineering & Technology
JAMIA MILLIA ISLAMIA
New DELHI

Mr. Shamim Ahmad
(Assistant Professor)
Department of Computer Engineering
Faculty Of Engineering & Technology
JAMIA MILLIA ISLAMIA
New DELHI

ACKNOWLEDGEMENT

We deeply thank Department of Computer Engineering, Jamia Millia Islamia, which provided us the entire infrastructure and help whenever it was needed for this project. We thank the Head of Department, **Prof. Tanveer Ahmad,** for providing all the help that was required for our project work.

We take this opportunity to express our profound gratitude to our mentor Mr. **Shamim Ahmad**, for his exemplary guidance, constant encouragement and monitoring throughout the course of the project. We thank him for his immense support and help.

We would like to extend our gratitude to the faculty members and laboratory coordinators of the department for providing us the infrastructural facilities necessary to sustain the project.

Osama Khan
15BCS0071
Dept. of Computer Engineering
Jamia Millia Islamia
New Delhi

Prabhat Kumar Moore
15BCS0072
Dept. of Computer Engineering
Jamia Millia Islamia
New Delhi

ABSTRACT

Air pollution and its prevention are constant scientific challenges during last decades. However, they still remain huge global problems. Affecting human's respiratory and cardiovascular system, they are cause for increased mortality and increased risk for diseases for the population. Many efforts from both local and state government are done in order to understand and predict air quality index aiming improved public health. This project is one scientific contribution towards this challenge.

In this minor project, we tackle air quality index forecasting by using machine learning approaches to predict the Air Quality Index. Machine learning, as one of the most popular techniques is able to efficiently train a model on big data by using large-scale optimization algorithms. However, the relationships between the concentration of Air pollutant particles and meteorological factors are poorly understood. To shed some light on these connections, This project attempted to apply some machine learning techniques to predict Air quality index category based on a data set consisting of daily meteorological data from May-2015 to Oct-2018 of Delhi, India

1:INTRODUCTION

1.1 About the project

In this minor project, we tackle air quality index forecasting by using machine learning approaches to predict the Air Quality Index. Machine learning, as one of the most popular techniques is able to efficiently train a model on big data by using large-scale optimization algorithms. However, the relationships between the concentration of Air pollutant particles and meteorological factors are poorly understood. To shed some light on these connections, This project attempted to apply some machine learning techniques to predict Air quality index category based on a data set consisting of daily meteorological data from May-2015 to Oct-2018 of Delhi, India

1.2 About Air quality Index

An air quality index (AQI) is a number used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects. Different countries have their own air quality indices, corresponding to different national air quality standards. Fig (1) depicts the air quality indices of India.

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6 –1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748*	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

fig(1)

1.3 Motivation

Air pollution in Delhi has been concerned as a serious problem. Every day measured values of the parameters of air quality are many times above the limit values that are considered safe for human health. Some of the activities to lower the air pollution are undertaken by the government. This project represents our effort on the scientific level to contribute in handling this problem. At the first phase the necessary step is to predict air quality index category in order to help improve the situation. In this project we developed four different classifiers based on different algorithms.

2:Data Collection and Preprocessing

2.1 Data Collection

we collected our meteorological and air pollutant data from two different websites.

2.1.1 Air Pollutant Data Collection

we downloaded our air pollutant data from Central Pollution Control Board (http://cpcb.nic.in/) from(may-2015 to Oct-2018) which is a statutory organization under the Ministry of Environments, Forrest and Climate Change (MoEFC).we selected the PM_{10}, $PM_{2.5}$, NO_2, O_3, CO, SO_2, NH_3, and Pb for air quality index calculation.

2.1.2 Meteorological Data Collection

we downloaded our meteorological data from (https://en.tutiempo.net/climate) from(may-2015 to Oct-2018) the features that has been selected are:

T	Average temperature (c^0)	PP	Total rainfall (mm)
TM	Maximum Temperature (c^0)	VV	Average visibility(km)
Tm	Minimum Temperature(c^0)	V	Average wind speed(km/hr)
SLP	Atmospheric pressure at sea level	VM	Maximum wind speed(km/hr)
H	Average relative humidity		

2.2 Data Preprocessing

We paired the collected meteorological data and air pollutant data on the basis of time to obtain the required data format for applying the machine learning methods. Missing values existed for some variables, which was not tolerable for applying the machine learning methods used in this project. Therefore, we imputed the missing values by using the median strategy.

2.2.1 Air quality index calculation

the pollutant data that we have collected from (http://cpcb.nic.in/) doesn't contain the air quality index so we have calculated the air quality index of each day from (may-2015 to Oct- 2018) from the pollutant data. To calculate the air quality index we used the following formula:

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} * (C - C_{low}) + I_{low}$$

I : the (Air Quality) index
C : the pollutant concentration
C_{low}: the concentration breakpoint that is ≤ C
C_{high}: the concentration breakpoint that is ≥ C
I_{low}: the index breakpoint corresponding to C_{low}
I_{high}: the index breakpoint corresponding to C_{high}

$C_{low}, C_{high}, I_{low}, I_{high}$ are from the US EPA Pollutant Breakpoint

after calculating the air quality index we categorize the air quality index into 4 category for aqi(0-100) category 1, for aqi(101- 200) category 2, for aqi(201-300) category 3 and for aqi(>= 301) category 4

3:Machine Learning Approach for prediction

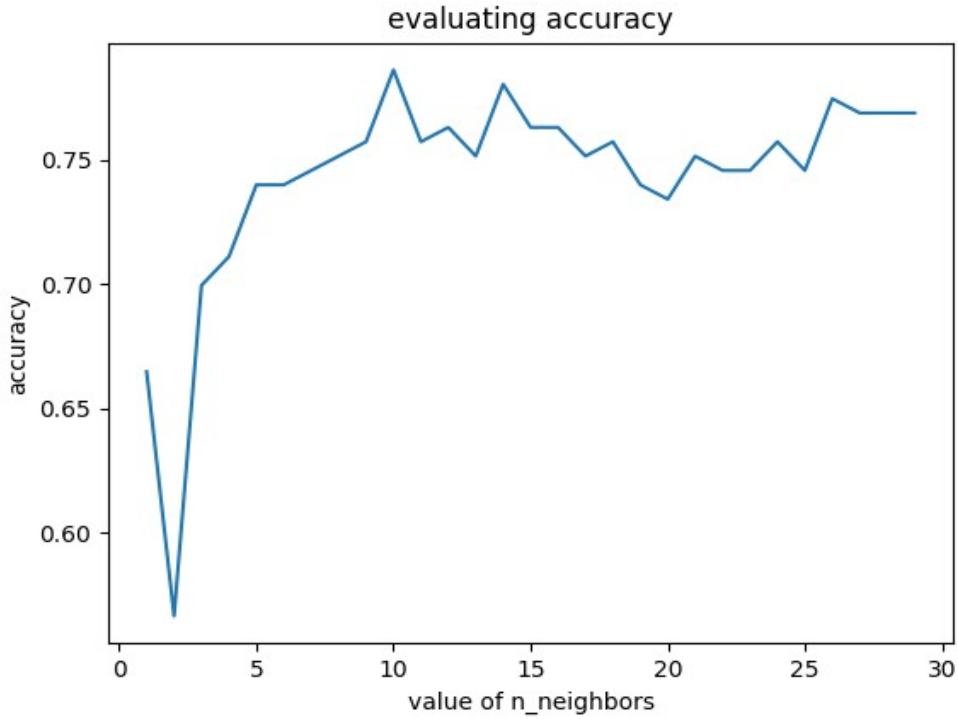
In this section, we describe the proposed approaches for predicting the category of air quality index.

3.1 K-NEAREST NEIGHBOR ALGORITHM MODELLING

K-NN is Supervised Learning classifier. K-NN are non-parametric techniques which are used for classification and regression. If used for classification, result is object classification depending on the result of the nearest neighbors voting, and the new object is dedicated to the class with more votes. On the other side, if used for regression, then the result is value that is dedicated to the object as an average of the values of its neighbors .In K-NN algorithm the training phase is not performed. Unlike the other algorithms, K-NN algorithm does not make any presumptions for data distributions, and, also, does not bring general conclusions. Because there is not training phase, it must keep all the data for training and search over them for neighbors. This process requires more time and more resources for testing phase. There are many distance metrics in K-NN. The most used distance metrics is Euclidean distance if continuous variables are considered. Other metrics that are often used are: city-block, hamming distance, cosine and correlation. The best choice for k for neighbors depends on data. In general, higher values of k are decreasing the effect of noise on classification, but they, also, limit differences between classes.

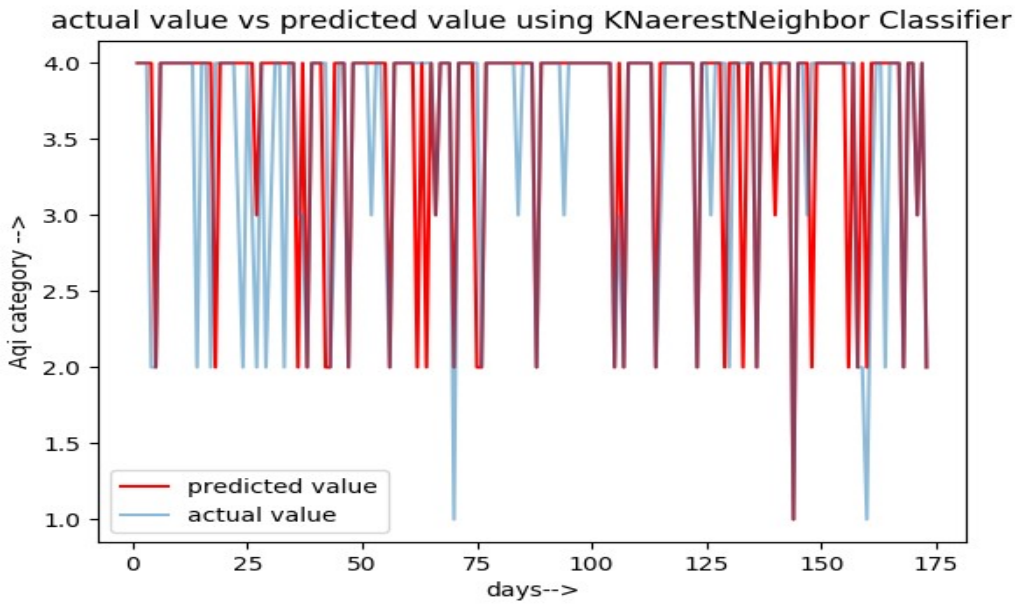
3.1.1 Experiments with k-NN algorithm

The experiments were done with few combinations in order to get the highest accuracy taking into account different value of nearest neighbors (n_neighbors). Values of n_neighbors were in interval from 1 to 30. Project has 4 classes for classification. Testing was performed with several types of metrics: Euclidean, correlation and cosine.



Fig(3.1.1)

Analyzing the experiments, it can be seen that the best accuracy is for n_neighbors=10 (Figure 3.1).



fig(3.1.2)

3.2 DECISION TREE MODELLING

Decision tree as a supervised learning algorithm is used as a model for statistic prediction, data mining and machine learning. There are two types of decision tree algorithms. classification tree (predicted outcome is the class which contains the data) and regression tree (predicted result is real number). Classification and regression tree with one name are called CART (classification and regression tree).

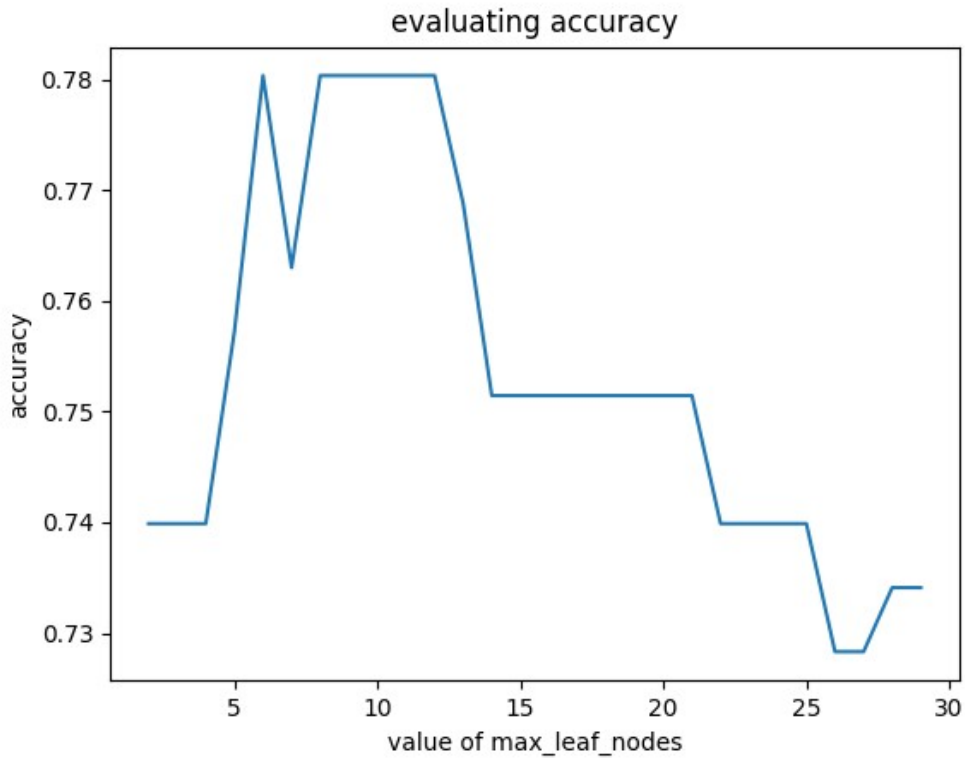
Classification trees are tree models where target variable contains set of discrete values. In this structures, leaves are presenting class signs and branches are conjunctions of properties which lead to signs of the class. Regression trees are decision trees where target variable gets continuous values (typically real numbers). Decision trees are used to make decisions and to visually or explicitly present them. The aim is to create a model which will predict target value based on previously learnt input data. Decision tree algorithm is constructed as a steam where each internal node represents test for an attribute, each branch represents value for tested feature and each leaf is classification for an instance. Highest node of the tree is the root. Hierarchy of rules has to be implemented. Rules of internal nodes (roots) test values of some feature of the data. Each internal node fits with one of the input variables. Each leaf is value of the target variables, according to input values, presented on path from the root to the leaves. Training data has to be used for tree construction. Afterwards, depending on the tree model, output values are predicted. Information with highest value are located on the top of the tree.

The process of learning the tree can be performed by splitting the source data presented in subsets based on test characteristics values. This splitting should be repeated on every of the subgroups and it is called recursive partitioning. At the moment when subset in the node have the

similar value of the target variable and when the splitting does not increase value of the predictions recursion is finished.

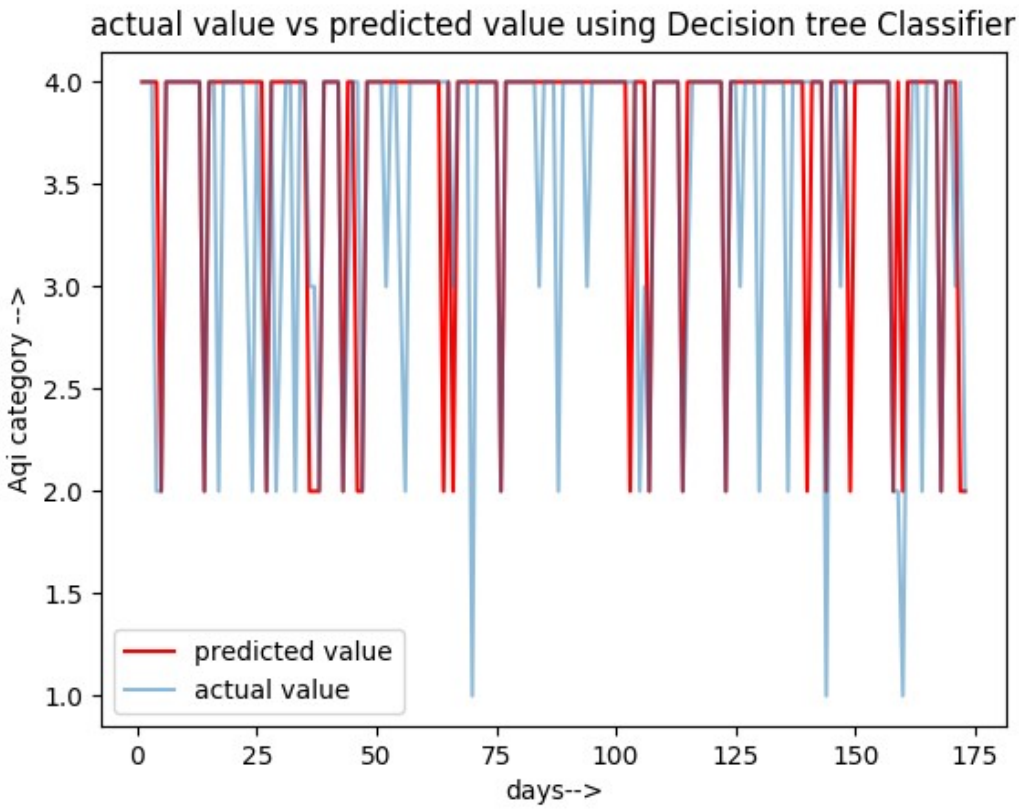
3.2 Experiments with Decision Tree algorithm

In order to perform experiments, DT was constructed with assumption that all input functions have final discrete domains, there is one target function classification of air pollution data (three classes). Every internal node in the DT contains input feature. Every leaf of the tree contains class.



Fig(3.2.1)

Analyzing the experiments, it can be seen that the best accuracy is for max_leaf_node=6 (Figure 3.1).which is 78.03468208092485%


actual value vs predicted value using Decision tree Classifier
Aqi category -->
days-->
predicted value
actual value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
25
50
75
100
125
150
175

fig(3.2.2)

3.3 Support Vector Machine

Kernel methods are a class of machine learning techniques that has become an increasingly popular tool for learning tasks such as pattern recognition, classification or novelty detection. This popularity is mainly as a result of the success of the support vector machines (SVM), probably the most popular kernel method, and to the fact that kernel machines can be used in many applications as they provide a bridge from linearity to non-linearity. The third model that we choose for comparison in our project is support vector machines. Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Some of the advantages of SVM are : effective in high dimensional spaces; different kernel functions can be used for various decision function; kernel functions can be combined to achieve more complex type of planes, although SVM have poor performance when number of features is grater then number of samples and SVM do not provide probability estimates which is the reason why cross-field validation is used.

SVM model is a representation of the examples as points in space, mapped in such a way that the examples of the separate categories are divided by a major vector (hyper plane) which is as wide as possible. Left and right from that major vector, supports vectors at the same distances from major vector are positioned. New examples are then mapped into the same space and predicted to belong to a category based on that on which side of the vector they fall. So, the result depends on the position of the major vector. This is called linear support vector machine (LSVM) [23]. Except performing

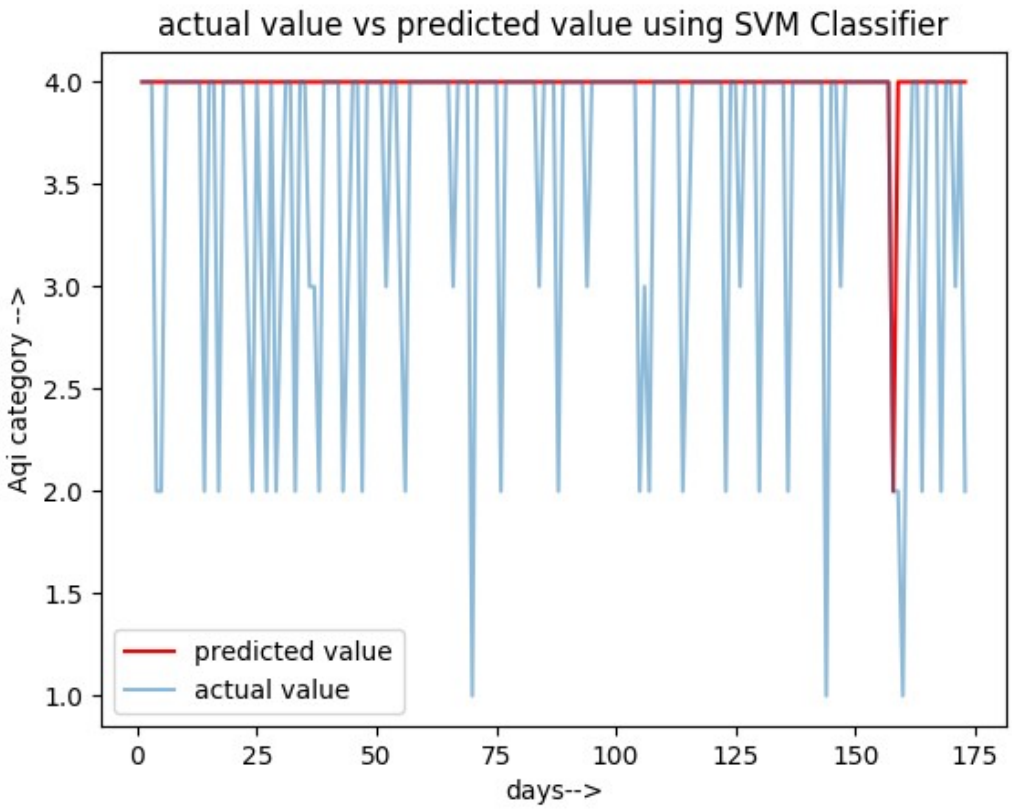
linear classification, SVMs can, also, efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Using kernel function, two vectors can be applied and every point is mapped into high dimensional space by a transformation. The idea is to transform non-linear space into linear space. There are several popular kernel types that can be used to transform data into high dimensional feature space: polynomial kernel, radial basic function kernel, sigmoid kernel.

3.3 Experiment with Support Vector Machine

In SVM algorithm different kernel functions were tried to get highest accuracy result. Experiments lead to a conclusion that maximum accuracy of SVM was when rbf kernel function was used. Table 2 shows the results for SVM with different kernel functions.

Table 2:Accuracy of SVM with different kernel function

Kernel function	accuracy
linear	73.98843930635837%
rbf	74.56647398843931%
sigmoid	73.98843930635837%


actual value vs predicted value using SVM Classifier
Aqi category -->
days-->
predicted value
actual value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
25
50
75
100
125
150
175

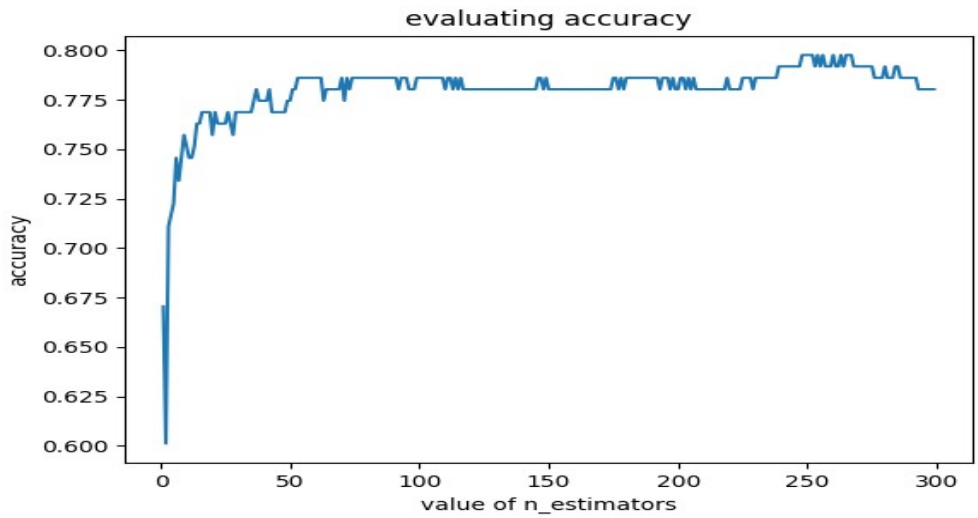
Fig(3.3.1)

3.4 Random Forrest Classifier

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

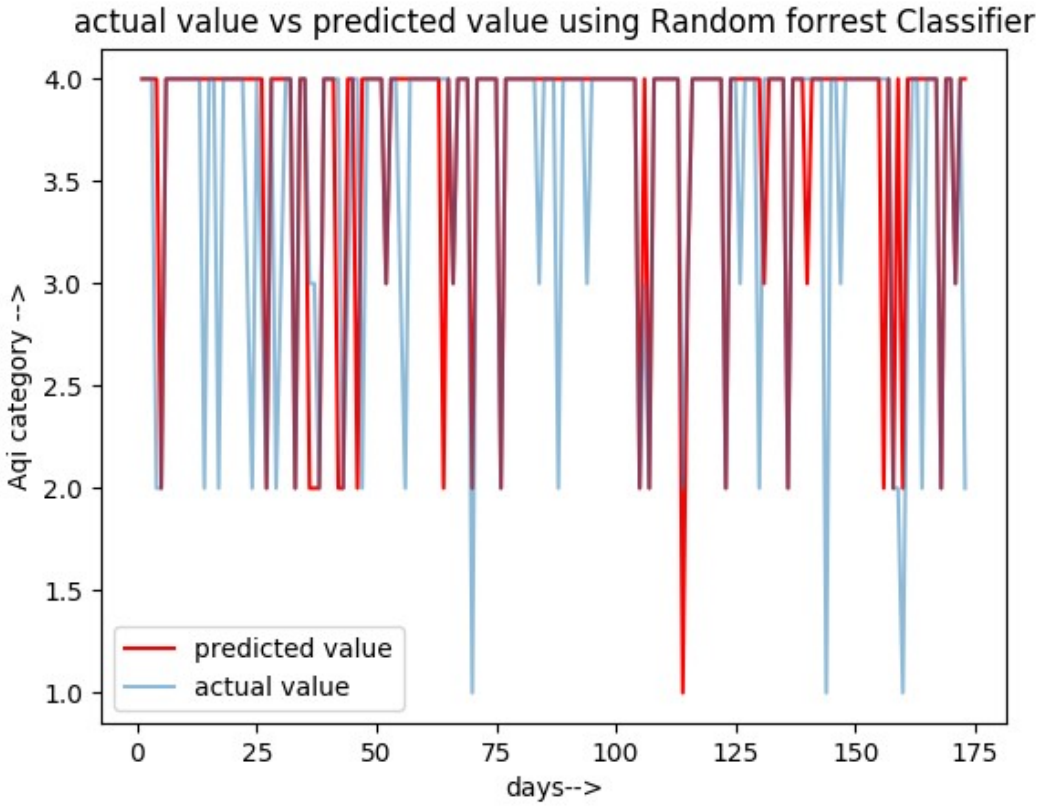
3.4.1 Experiment Random Forrest Classifier

The experiments were done with few combinations in order to get the highest accuracy taking into account different value of n_estimators. Values of n_estimators were in interval from 1 to 300. Project has 4 classes for classification.

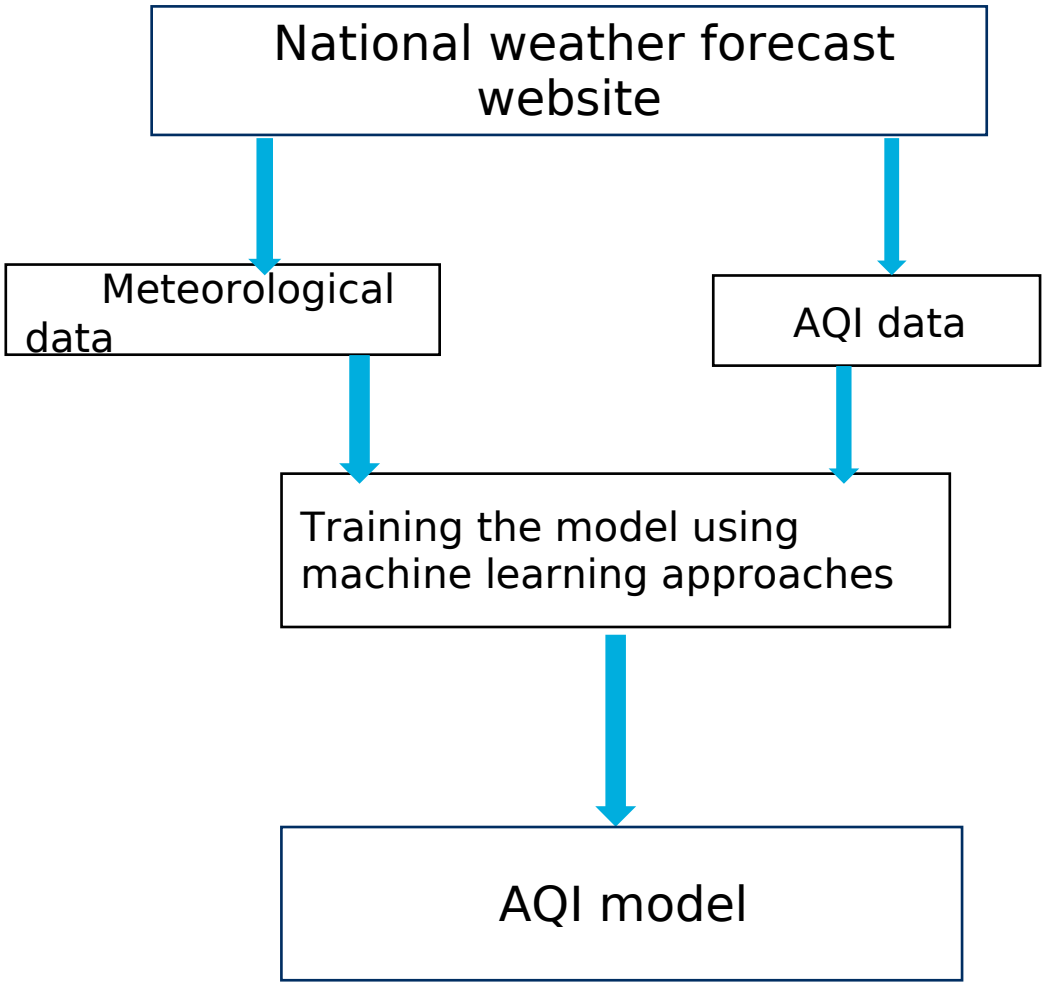


fig(3.4.1)

from fig(3.4.1) it can be seen that the best accuracy is 79.7687861271% for n_estimators = 266.

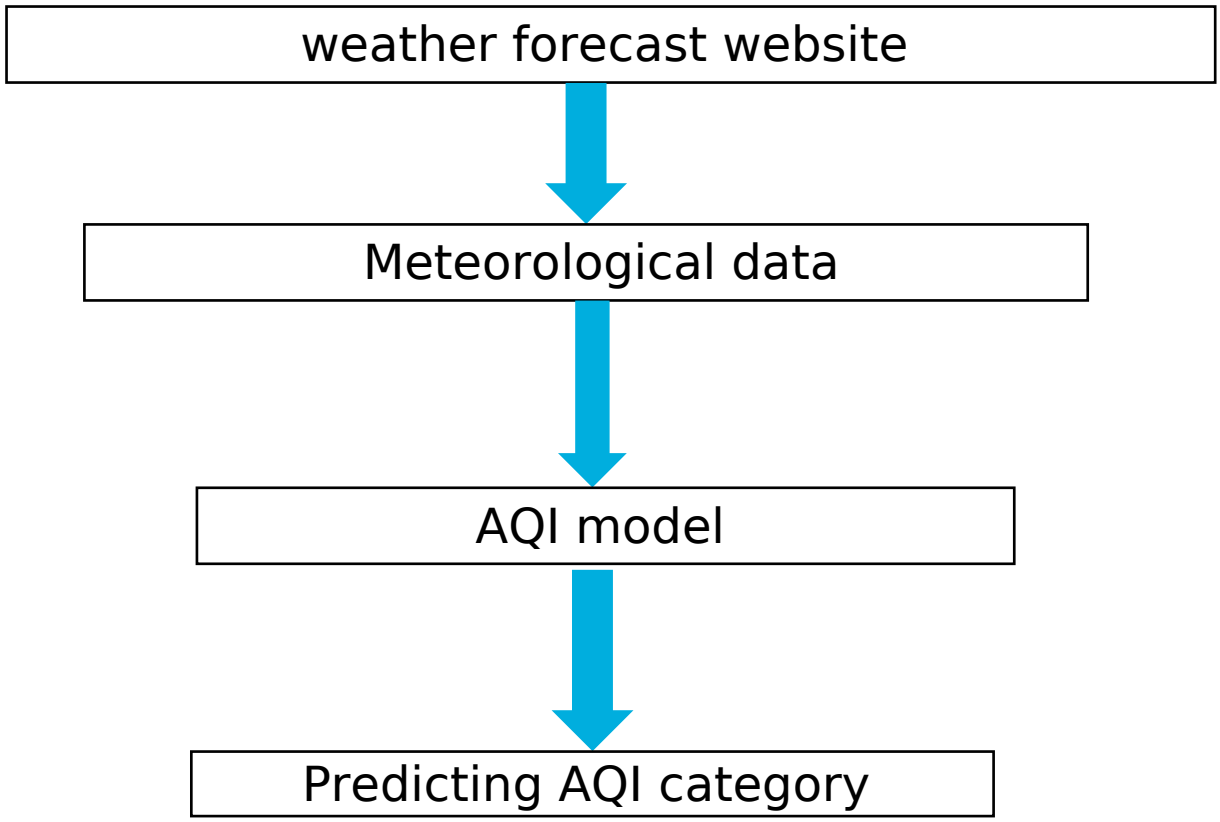


3.5 flow-diagram for training our model



fig(3.5)

3.6 flow-diagram for testing our model



4. Related work

Many previous works have been proposed to apply machine learning algorithms to air quality predictions. Some researchers have aimed to predict targets into discretized levels. Kalapanidas et al. elaborated effects on air pollution only from meteorological features such as temperature, wind, precipitation, solar radiation, and humidity and classified air pollution into different levels (low, med, high, and alarm) by using a lazy learning approach, the case-based reasoning (CBR) system. Athanasiadis et al. employed the σ-fuzzy lattice neurocomputing classifier to predict and categorize O 3 concentrations into three levels (low, mid, and high) on the basis of meteorological features and other pollutants such as SO2, NO, NO2, and so on. Kurt and Oktay modeled geographic connections into a neural network model and predicted daily concentration levels of SO2, CO, and PM10 3 days in advance. However, the process of converting regression tasks to classification tasks is problematic, as it ignores the magnitude of the numeric data and consequently is inaccurate.

Other researchers have worked on predicting concentrations of pollutants. Corani worked on training neural network models to predict hourly O3 and PM10 concentrations on the basis of data from the previous day. Mainly compared were the performances of feed-forward neural networks (FFNNs) and pruned neural networks (PNNs). Further efforts have been made on FFNNs: Fu et al. applied a rolling mechanism and gray model to improve traditional FFNN models. Jiang et al. explored multiple models (physical and chemical model, regression model, and multiple layer perceptron) on the air pollutant prediction task, and their results show that statistical models are competitive with the

classical physical and chemical models. Ni, X. Y. et al. compared multiple statistical models on the basis of PM2.5 data around Beijing, and their results implied that linear regression models can in some cases be better than the other models.

Therefore, we decided to use meteorological and pollutant data to perform predictions of daily concentrations on the basis of classifiers. In this project we divided divided our air quality into 4 categories: good/satisfactory, mild, poo, very poor, and we compare four different algorithm of machine learning : K-NN,Random Forrest, SVM, Decision tree.

5. Conclusion

Our project compares four different algorithms for machine learning: k-NN, DT,Random Forrest and SVM. The experiments were conducted using database of air pollution in new Delhi India.

For the k-NN algorithm several combinations were made to obtain the highest accuracy with different values of the nearest neighbors (k). Values were taken in the interval from k=1 to k=21. The most accurate value that we have got is 78.61271676300578% when the value of k is 10.

In SVM algorithm different kernel functions were tried to get highest accuracy result. Experiments lead to a conclusion that maximum accuracy of SVM was 74.566% when rbf kernel function was used.

The decision trees are faster in data processing and easy to understand, the best accuracy is 78.03468208092485% when the max_leaf_node = 6.

In Random Forrest Algorithm different value for n_estimator were used to evaluate the accuracy of the algorithm. The best accurate value that we have got is 79.7687861271% when n_estimators = 266 which is better than all other algorithms that has been used in this project.

6. Future Work

the data set in this project is not large enough. Air quality is a long-term formed problem and it is better to use a large data data covering a variety of years and locations. Furthermore, beside the meteorological factors, traffic and industrial parameters such as power plant emissions also play significant roles in air pollution. This project did use these features

because they are not publicly available in New Delhi. In order to get better prediction results, the data should include more industrial condition features if possible.

6. References

[1].https://en.wikipedia.org/wiki/Air_quality_index

[2].http://cpcb.nic.in/

[3].https://en.tutiempo.net/climate

[4].L. Wang, Y.P. Bai, “Research on Prediction of Air Quality Index Based on NARX and SVM”, Applied Mechanics and Materials (Volumes 602-605), 3580- 3584, 2014

[5].BC. Liu, et al, “Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-TianjinShijiazhuang”, PLOS, 2017

[6].H. Wang, et al., “Air Quality Index Forecast Based on Fuzzy Time Series Models”, Journal of Residuals Science & Technology, Vol. 13, No. 5, 2016

[7].Kostandina Veljanovska1 , Angel Dimoski2, “Air Quality Index Prediction Using Simple Machine Learning Algorithms”,Volume 7, Issue 1, January - February 2018

[8].N.S. Altman, “An introduction to kernel and nearestneihbor nonparametric regression”, The American Statistician, Vol. 46, No. 3, pp. 175-185, 1992

[8].R.J. Samworth, “Optimal weighted nearest neighbor classifiers”, 2012

[9].]D. Coomans; D.L. Massart, “Alternative k-Nearest neighbor rules in supervised pattern recognition: Part 1. k-Nearest neighbor classification by using alternative voting rules”. Analytica Chimica Acta, 1982

[11].S. Canu, “SVM and kernel machines: linear and nonlinear classification”, OBIDAM, Brest, 2014

[7].L. Breiman, et al., “Classification and regression trees”. Montrey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984

[12].C. Cortes, V. Vapnik, “Support-vector network”, Machine Learning, 1995.

[13].Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appl.2010, 37, 7986–7992.

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