1:INTRODUCTION

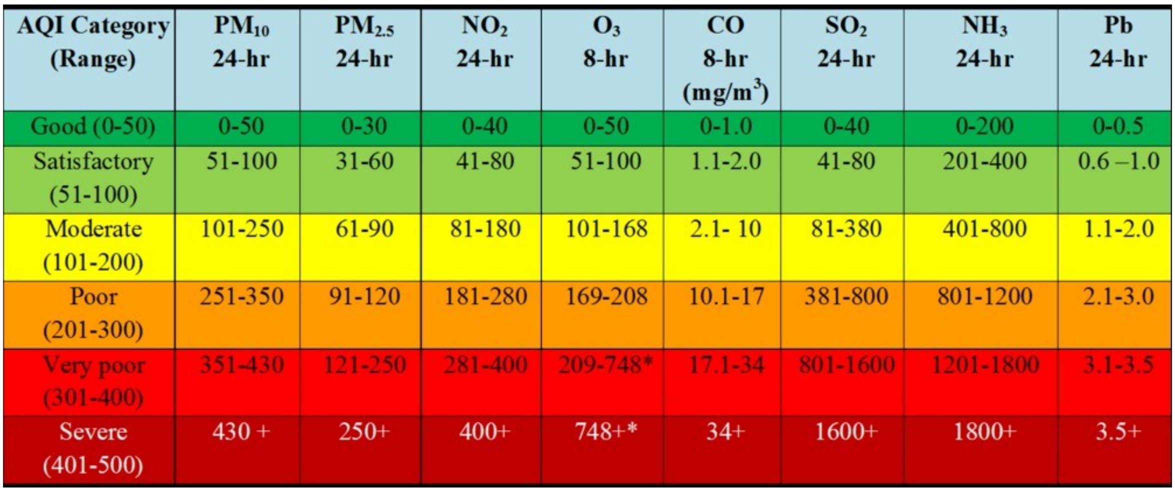
1.1 About the project

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:

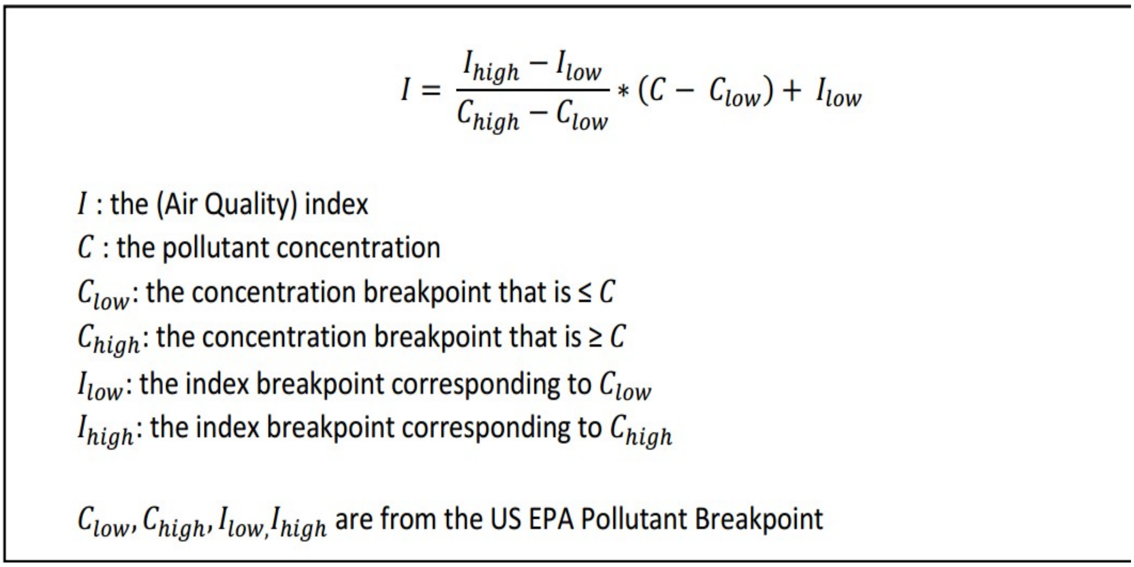
T	Average temperature (c^0)	PP	Total rainfall (mm)
TM	Maximum Temperature (c^0)	VV	Average visibility(km)
Tm	Minimum Temperature(c^0)	V	Average wind speed(km/hr)
SLP	Atmospheric pressure at sea level	VM	Maximum wind speed(km/hr)
H	Average relative humidity		

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

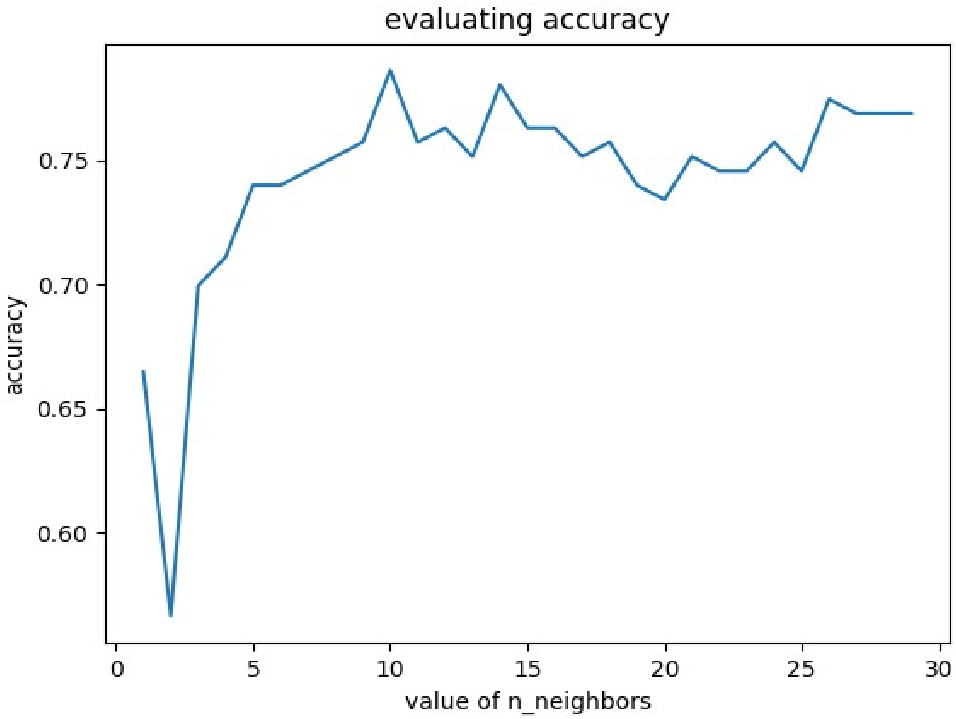
after calculating the air quality index we categorize the air quality index into 4 category for aqi(0-100) category 1, for aqi(101- 200) category 2, for aqi(201-300) category 3 and for aqi(>= 301) category 4

3:Machine Learning Approach for prediction

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

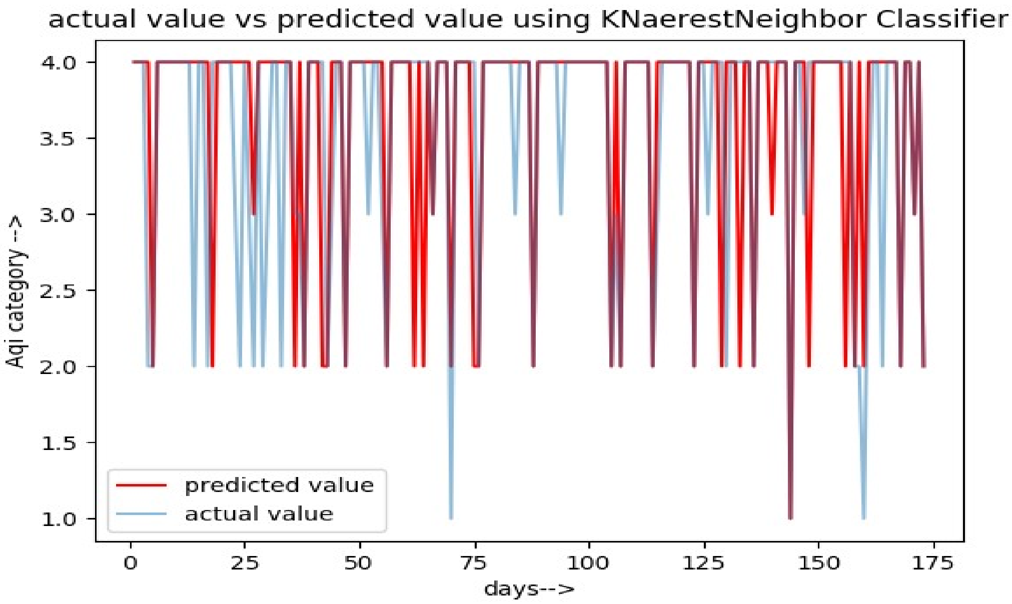
3.1.1 Experiments with k-NN algorithm

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)

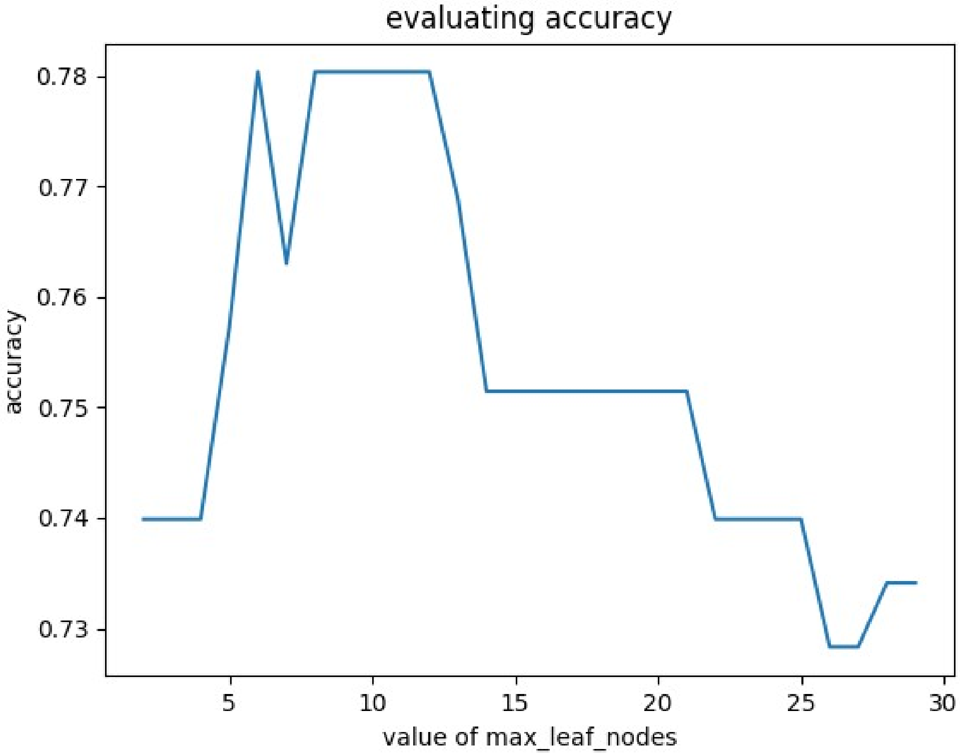
Analyzing the experiments, it can be seen that the best accuracy is for n_neighbors=10 (Figure 3.1).



fig(3.1.2)

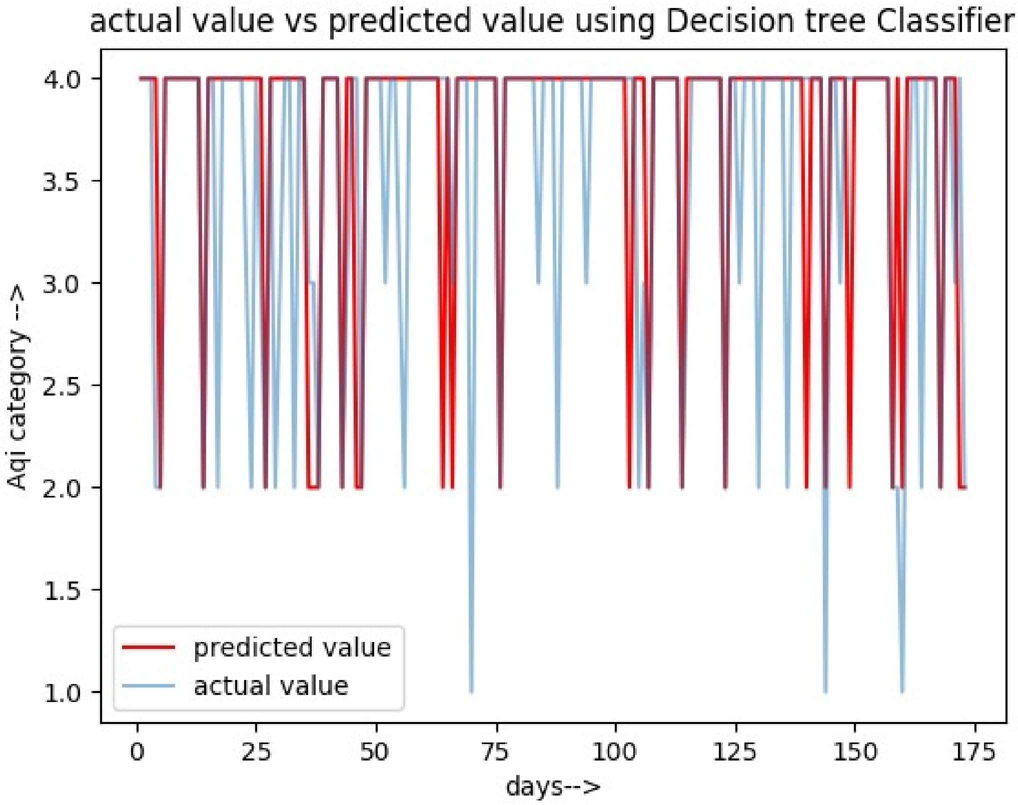
3.2 Experiments with Decision Tree algorithm

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 -->
4.0
3.5
3.0
2.5
2.0
1.5
1.0
predicted value
actual value
0
25
50
75
100
125
150
175
days-->

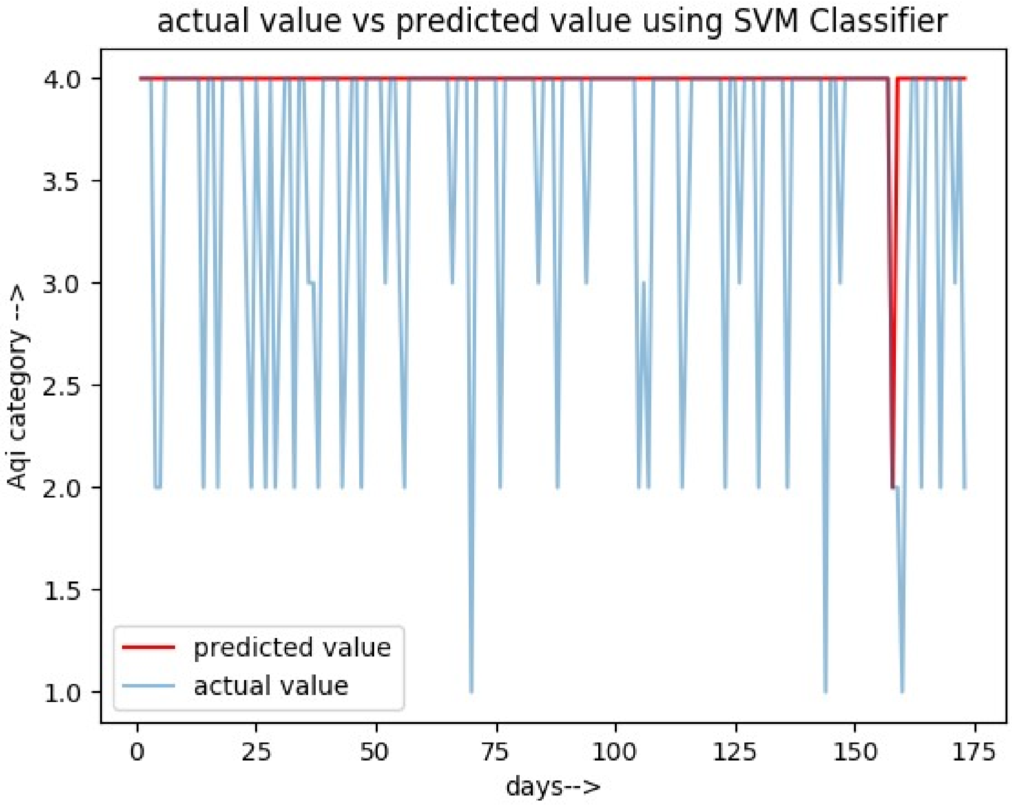
fig(3.2.2)

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%

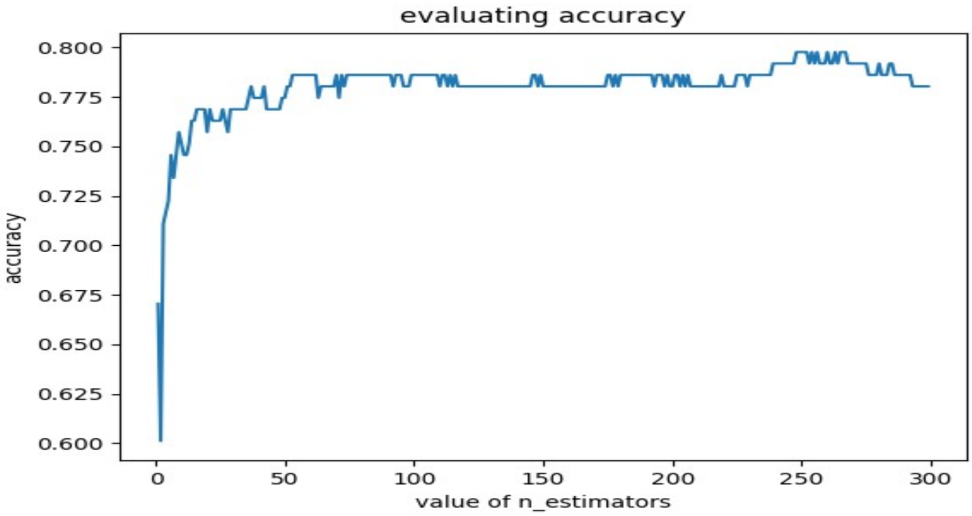

actual value vs predicted value using SVM Classifier
Aqi category -->
4.0
3.5
3.0
2.5
2.0
1.5
1.0
predicted value
actual value
0
25
50
75
100
125
150
175
days-->

Fig(

3.3.1)

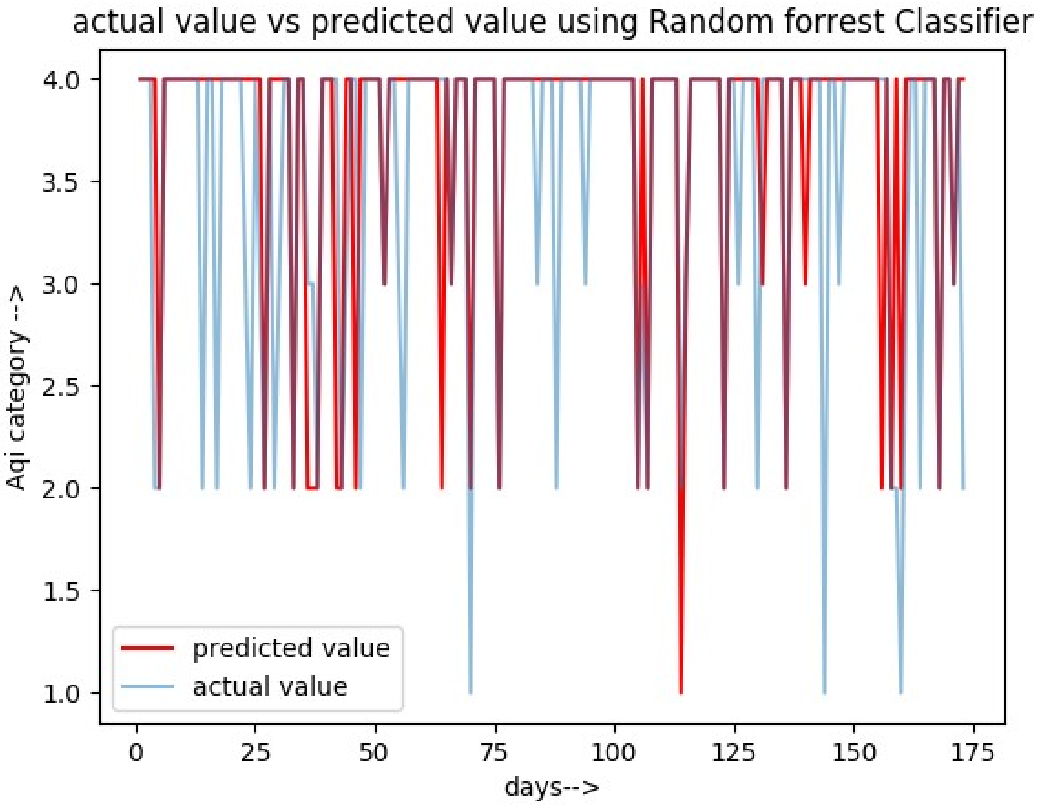
3.4.1 Experiment Random Forrest Classifier

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.

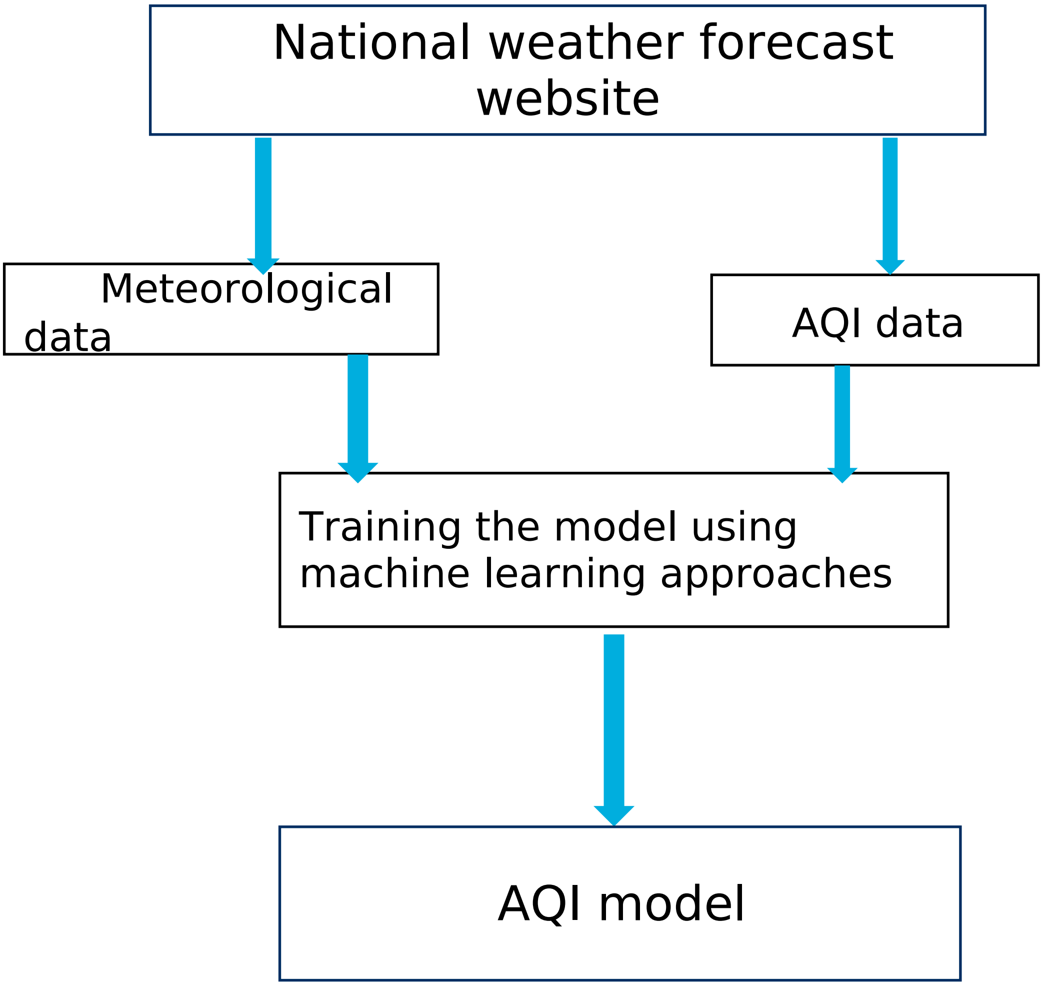


fig(3.4.1)

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

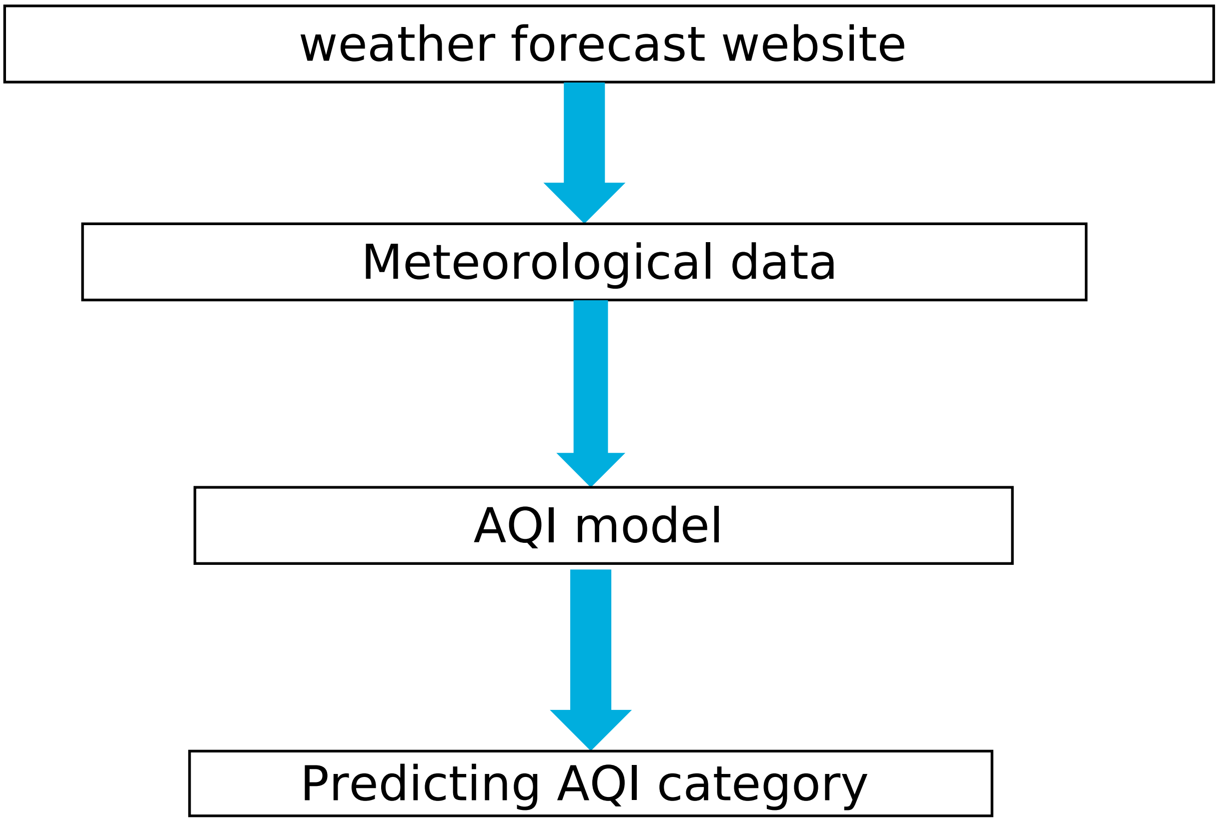

actual value vs predicted value using Random forrest Classifier
Aqi category -->
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
25
50
75
100
125
150
175
days-->
predicted value
actual value

3.5 flow-diagram for training our model


National weather forecast website
Meteorological data
AQI data
Training the model using machine learning approaches
AQI model

fig(3.

3.6 flow-diagram for testing our model


weather forecast website
Meteorological data
AQI model
Predicting AQI category