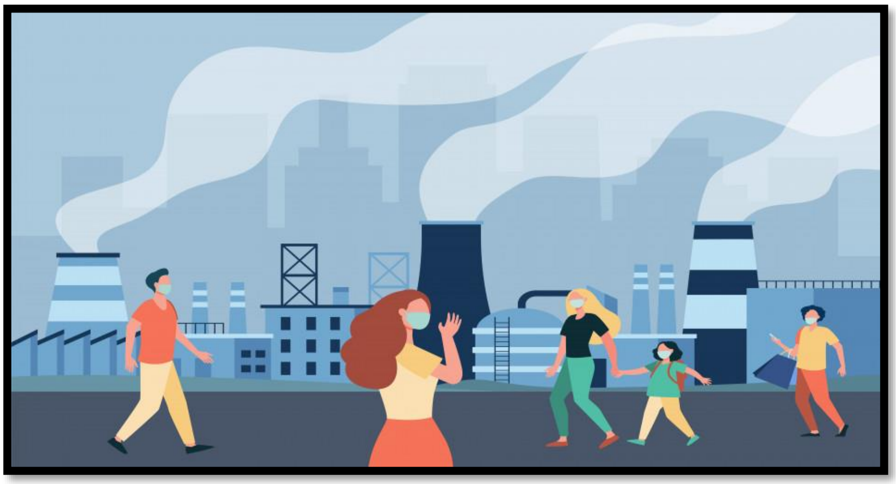
Team Name: **MachineLearners**

Problem Statement: Real-Time Air Quality Monitoring & Weather Forecasting System

Team Members :

1. Sharmistha Das (Team Leader)
2. Agniban Saha
3. Abhirup Basak
4. Deep Sarkar



1. INTRODUCTION :

Climate change is one amongst the foremost significant challenges for humankind. Climate is influenced by natural phenomena like the sun, volcanism, greenhouse gases, pollution caused by industry and vehicles. Climate successively affects air quality by increasing ventilation rates like wind speed, circulation, dry deposition, natural emissions, and background concentrations. This work has been entirely focused on air quality, pollutants, meteorological effects, weather and its prediction.

Air is one amongst the essential things in a human's life, and its quality impacts every living being on the world. Raising awareness among citizens is the key to comprehensive support of these efforts. Therefore, web-based application can provide a valuable means to provide information and warn citizens. as an example, the user could even be alerted about the air quality level, and suggestions on a way to react and improve air quality are often disseminated quickly by a mobile. Today's technology is meant to evolve continuously overtime to assist humans to measure a sustainable and healthy life. With this, computing is that the enabling paradigm that's already utilized in various domains like general information systems, military systems, medicine, and manufacturing and production. In environmental research, machine learning is also accustomed better understand air quality and to predict near-future air quality.

1.1 OVERVIEW :

- ✓ The solution applies machine learning techniques to a web-based application that informs the current and predicts the near future air quality and weather forecasting. Prediction is based on weather forecasts and the experimentation of the concept is done by using the classifier Random Forest.
- ✓ To show the performance of our proposed idea, we trained the dataset and forecasted it by fitting models with optical hyper-parameters. The forecasted air-quality index values were then compared with observed values.
- ✓ The final implementation is exemplified for location Bangalore-India.
- ✓ The results of the prediction are observed and compared with the actual air quality index for 16 days. The overall prediction accuracy is 81%, using a Random Forest algorithm. The value of MAE is 10.27, MSE is 275.95, and RMSE is 16.61.

1.2 OBJECTIVE :

This work aims to predict air quality using weather forecasts and machine learning. As part of this work, a web-based mobile/pc application has been developed to provide the current air quality and the predicted near feature air quality status. The need for a mobile information system for people is to raise an immediate warning and awareness about the poor air quality.

To build an air quality information system that provides the needed functionality, the following steps are taken in this solution:

- An appropriate data source for meteorological information and air pollution information needs to be selected and evaluated.
- A mobile solution will be developed to provide ubiquitous and location-based air pollution information.
- A predictive approach to air quality information provisioning has been introduced that allows forecasting air quality in the near future.

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: everyone may experience more serious health effects.
Hazardous	301 to 500	Health warnings of emergency conditions. The entire population is more likely to be affected.

Note: Values above 500 are considered Beyond the AQI. Follow recommendations for the "Hazardous category." Additional information on reducing exposure to extremely high levels of particle pollution is available here.

Fig 1: AQI

2. ***LITERATURE SURVEY:***

Air quality is heavily dependent on the weather and climate. The change in climate affects air quality by perturbing the ventilation rates (wind speed, mixing depth, convection, frontal passages), precipitation scavenging, chemical production, dry deposition, natural emission, and background concentrations. The air quality is measured in-terms of air pollutants and concentration levels of pollutants. The leading causes of air pollution are the complex blend of particles, vapours, gases emitted from natural, anthropogenic sources, and photochemical transformation processes.

Air pollution causes serious health issues for all human beings. The major air pollutants are carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), lead (Pb), particulate matter with a diameter of about 10 micrometres or smaller (PM_{10}), particulate matter with a diameter of about 2.5 micrometre or smaller ($PM_{2.5}$) and sulphur dioxide (SO_2).

2.1 ***EXISTING APPROACHES:***

Serial No.	Title	Author	Description	Year
1.	Principal Component Regression (PCR) and Multiple Linear Regression Techniques	Anikender Kumar, PramilaGoyal	This present study that forecasts the daily AQI value for the city Delhi, India using previous record of AQI and meteorological parameters.	2011
2.	Data analytics Time series Regression forecasting	Nidhi Sharma	It predicted the future trends of various pollutants as Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Suspended Particulate Matter (PM), Ozone (O3), Carbon Monoxide (CO) and Benzene.	2018
3.	ZeroR algorithm in WEKA tool	Mohamed Shakir and N.Rakesh	The study come up with the results that shows that the concentration levels of air pollutants increase during the working days and especially during the peak hours of the day and decrease during week-ends or holidays	2018
4.	The AirQ software (proposed by WHO)	YusefOmidiKhaniabadi	This study is to discover the relation or association between health impacts such as mortality rate of cardiovascular diseases and the air pollutants	2016
5.	Artificial Neural Network (ANN) and Multiple Linear Regression (MLR)	ArchontoulaChaloulakou	The study is to forecast the PM10 concentration over two year time period for the city Athens, Greece.	2003

2.2 PROPOSED METHODOLOGY:

1. **1.** The air quality and meteorological dataset has been scrapped from https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india
2. **2.** The above-mentioned site consists of past and current records of air quality as well meteorological data for each and every state in India, in our test case we have taken the dataset of Bangalore, we can change the state as per our requirement criteria.
3. **3.** The collected data is then modified, as we have calculated the daily hours Air Quality data into daily average data of 2013 to 2017 for target value $PM_{2.5}$.
4. **4.** The contents of input dataset are listed below:

I. Particulate Matter2.5 (PM2.5)

II. Particulate Matter10 (PM10)

III. Nitric Oxide (NO)

IV. Nitrogen Dioxide(No2)

V. Nitrogen Oxides (Nox)

VI. Ammonia (NH3)

VII. Carbon Monoxide (CO)

VIII. Sulphur Dioxide (SO2)

IX. Toluene(Tol)

City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Bengaluru	21-03-2015	48.59		3.47	27	18.04	28.02	3.89	1.94	52.96	21.33	196.72		91	Satisfactory
Bengaluru	22-03-2015	47.38		2.84	22.39	15.33	23.19	11.29	2.05	71.85	14.97	138.2		120	Moderate
Bengaluru	23-03-2015	65.65		3.1	26.35	17.45	27.76	9.95	6.3	72.84	9.88	100.77		154	Moderate
Bengaluru	24-03-2015	60.47		5.39	29.87	20.88	35.1	1.46	6.07	64.12	5.9	61.48		119	Moderate
Bengaluru	25-03-2015	62.56		3.16	23.57	16.39	27.13	10.05	4.98	82.34	4.53	39.99		232	Poor
Bengaluru	26-03-2015	60.73		4.16	32.07	21.03	34.71	1.44	1.67	67.21	1.36	13.28		132	Moderate
Bengaluru	27-03-2015	42.1		4.01	25.55	17.37	30.96	2.85	2.46	51.93	0.14	3.71		123	Moderate
Bengaluru	28-03-2015	40		2.93	17.99	11.36	28.74	8.07	2.5	44.65	0.14	8.39		152	Moderate
Bengaluru	29-03-2015	32.16		2.38	12.9	8.3	21.5	8.53	2.08	56.31	0.15	8.53		143	Moderate
Bengaluru	30-03-2015	30.78		2.68	16.23	14.63	26.96	1.46	2.63	49.52	0.14	2.07		80	Satisfactory
Bengaluru	31-03-2015	38.99		2.6	10.97	12.7	27.52	1.79	2.79	47.67	0.14	2.39		90	Satisfactory
Bengaluru	01-04-2015	36.42		2.68	8.83	9.13	31.37	9.6	3.17	48.25	0.14	8.16		151	Moderate
Bengaluru	02-04-2015	38.15		4.24	10.29	9.99	43.09	4.06	7.65	53.26	0.14	5.29		123	Moderate
Bengaluru	03-04-2015	35.11		2.63	5.99	6.82	26.04	5.34	2.32	57.75	0.14	7.4		178	Moderate
Bengaluru	04-04-2015	36.68		3.63	12.89	10.88	40.71	1.47	3.29	76.36	0.13	1.61		101	Moderate
Bengaluru	05-04-2015	36.92		2.2	9	7.6	26.22	9.58	4.07	65.62	0.14	12.02		160	Moderate
Bengaluru	06-04-2015	31.78		2.48	8.14	8.06	22.34	0.55	7.22	54.31	0.18	0.28		71	Satisfactory
Bengaluru	07-04-2015	37.25		2.86	7.14	11.25	17.78	2.3	5.88	60.42	0.2	2.3		76	Satisfactory
Bengaluru	08-04-2015	29.8		2.59	9.04	7.52	24.2	6.84	1.84	49.55	0.12	20.68		153	Moderate
Bengaluru	09-04-2015	25.06		3.21	8.72	7.2	29.83	5.82	2.6	42.63	0.13	5.92		80	Satisfactory
Bengaluru	10-04-2015	24.4		6.87	7.46	8.7	23.17	5.19	2.03	44.47	8.64	8.48		129	Moderate
Bengaluru	11-04-2015	32.08		6.6	7.09	8.82	30.95	1.7	3.22	53.91	0.29	1.75		107	Moderate
Bengaluru	12-04-2015	19.72		1.81	4.97	5.51	34.85	2.65	5.16	44.53	0.17	2.65		122	Moderate
Bengaluru	13-04-2015	17.03		2.17	8.53	9.95	34.3	2.23	2.75	39.03		2.62		85	Satisfactory
Bengaluru	14-04-2015	19.73		2.26	6.1	6.13	27.48	1.4	4.66	57.7	0.18	1.71		62	Satisfactory
Bengaluru	15-04-2015	22.52		6.71	4.86	14.02	15.15	5.45	4.9	77.05	0.21	5.37		177	Moderate
Bengaluru	16-04-2015	21.36		2.86	7.98	12.69	16.61	0.97	29.11	96.88	0.16	0.54		152	Moderate
Bengaluru	17-04-2015	24.35		3.06	6.64	7.42	38.54	12.93	2.58	97.2		12.95		210	Poor
Bengaluru	18-04-2015	18.33		2.4	5.98	6.47	21.02	3.84	3.99	101.63		4.51		170	Moderate
Bengaluru	19-04-2015	17.66		1.85	6.58	5.89	21.93	0.48	2.49	81.34		0.48		135	Moderate
Bengaluru	20-04-2015	27.46		2.53	7.91	7.1	24.78	4.58	5.53	37.22		5.17		118	Moderate
Bengaluru	21-04-2015	27.94		3.8	6.45	9.94	17.34	6.14	4.43	34.74	0.23	6.14		127	Moderate
Bengaluru	22-04-2015	23.81		4.01	6.21	13.57	5.42		26.42	32.79	0.22			47	Good
Bengaluru	23-04-2015	23.11		4.16	6.84	14.07	18.77	3.65	12.35	15.88		12.16		44	Good
Bengaluru	24-04-2015	15.12		4.35	14.96	12.52	35.77	1.13	5.12	109.66		1.27		67	Satisfactory
Bengaluru	25-04-2015	19.59		4.35	8.42	12.28	17.93	3.1	26.92	157.2		2.17		44	Good
Bengaluru	26-04-2015	69.85		8.6	14.35	19.55	15.8	1.22	15.91	123.32		0.77		82	Satisfactory
Bengaluru	27-04-2015	24.62		3.77	7.74	11.71	13.92	1.26	18.89	65.38		1.36		156	Moderate
Bengaluru	28-04-2015	32.92		7.12	12.95	18.7	11.62	2.31	14.79	57.59	43.78	3.7		96	Satisfactory
Bengaluru	29-04-2015	44.53		7.24	17.71	22.68	11.4	1.34	8.34	27.11	47.26	15.45		72	Satisfactory

Fig 1: Arrangement of Dataset

5. **5.** The processed data is then divided into 2 groups training set and testing set and then fed into the **Random Forest Regressor.** The training test is then used to develop the architecture of the program, while the testing set is used to test the accuracy of the developed prototype.

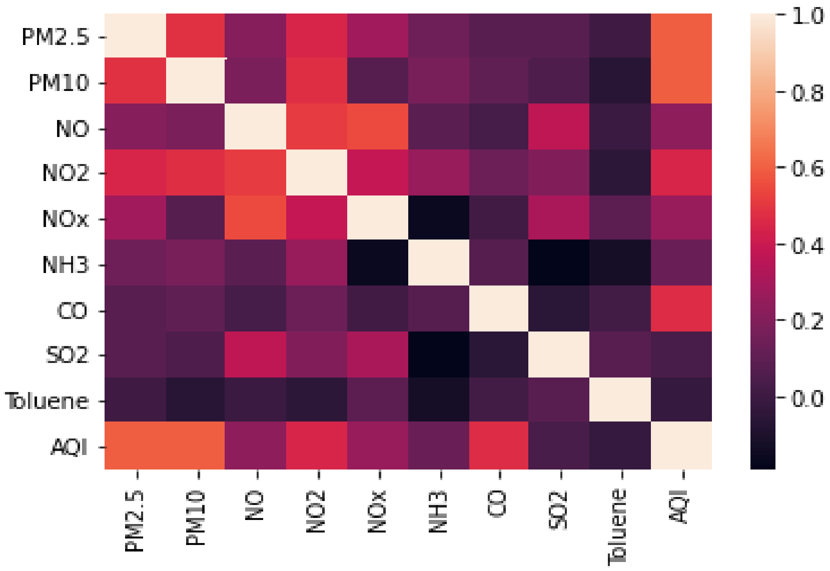


Fig 2: Correlation Matrix with heatmap

6. To show the performance of our proposed idea, we trained the weather and air quality data and forecasted it by fitting models with optical hyper-parameters. The forecasted AQI values were then compared with observed AQI values.

7. The performance is calculated in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

8. Then we integrate our developed learnings to a Web-based application so that the user could also be alerted about the air quality index, and suggestions on how to react and improve air quality can be disseminated quickly by a mobile or computer.

3. ***THEORITICAL ANALYSIS:***

The real time air quality data has been fitted to the Random Forest Algorithm based model. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

3.1 BLOCK DIAGRAM

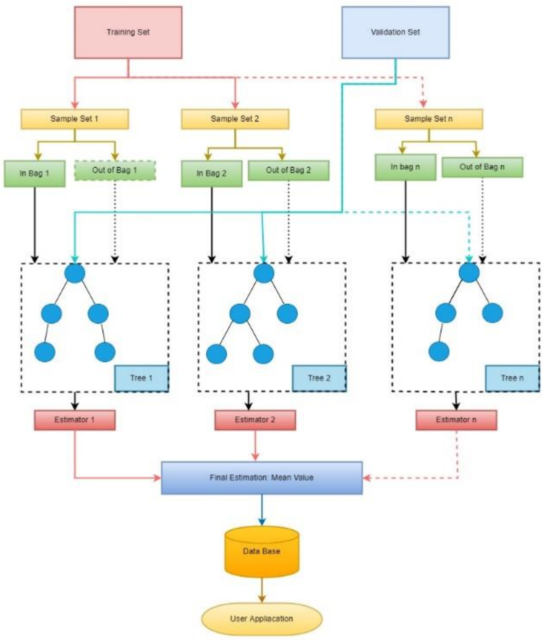


Fig 2: Data Flow Diagram of Proposed Model

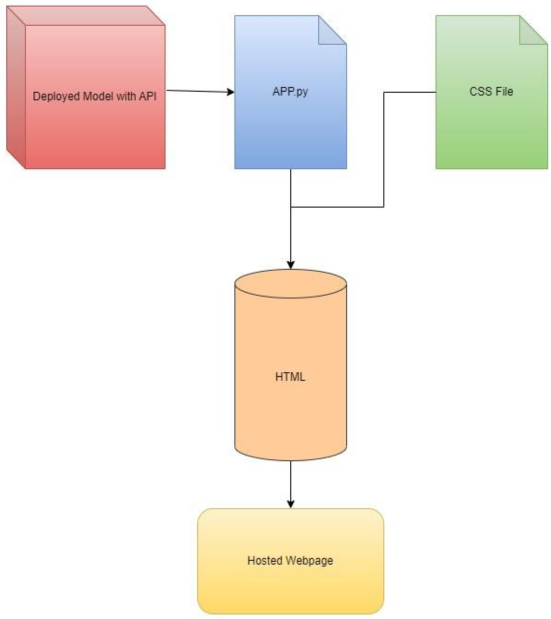


Fig 3: Block Diagram of User Interface

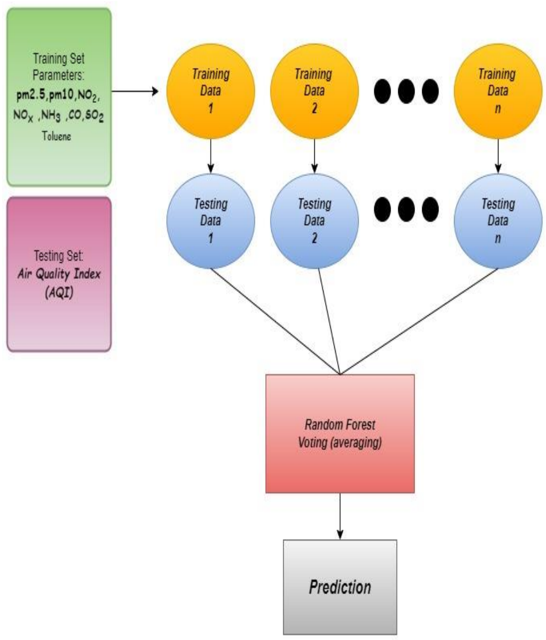


Fig 2: Block Diagram of Proposed Random Forest Algorithm with specific parameters

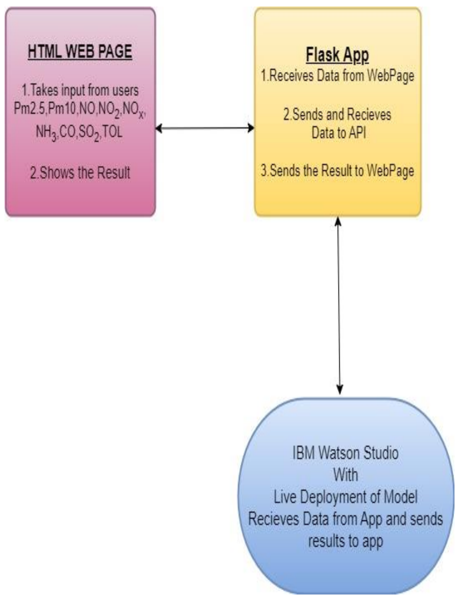


Fig 3: Flow Diagram in UI / UX

3.1 ***SOFTWARE DESIGNING:***

The main components required for designing the proposed software are:

Basis of Classification	Proposed ML Model	User Interface
Programming Language	• Python	• HTML • CSS • Python
Framework	• IBM Watson Studio	• Flask
Imported Libraries	• Pandas • Numpy • Matplot • Seaborn • Sklearn • Joblib • Json	• Flask • Jsonify • Render_template • Requests
Algorithms	• Random Forest Regression	NA
Host	• IBM Watson Studio	• Local Host
Feasibility	NA	• Should have Python installed. • Should have a Browser.
Outcomes	• Predicted Values	• Webpage

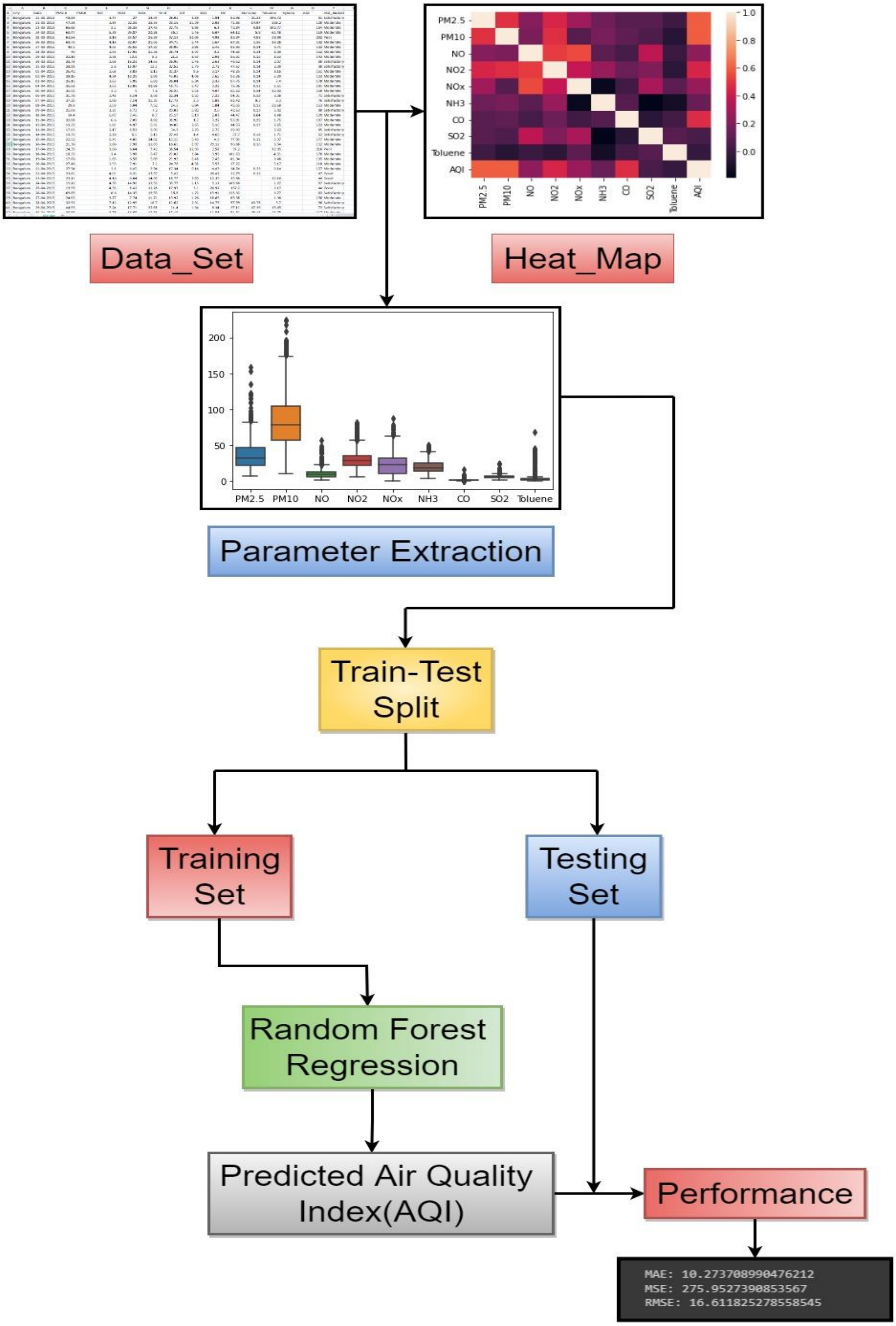
4. ***EXPERIMENTAL INVESTIGATIONS:***

Forest Regression & Classification, which is the best algorithm for regression for our project compared to other pre-existing algorithms. Our model is already providing an accuracy of 81% which can be further improved by incorporating more variations-based data which is perfect for Random Forest as it can handle large datasets with high dimensionality. Our model has a slightly higher training time compared to other developed machine learning based algorithms which have lower accuracy. One of benefits of Random Forest which exists me most is, the power of handle large data sets with higher dimensionality. It can handle thousands of input variables and identity most significant variables so it is considered as one of the dimensionality reduction method. Further, the model outputs importance of variable, which can be a very handy feature. It has an effective method for estimating missing data and maintains accuracy when large proportion of the data are missing. It has methods for balancing errors in data sets where classes are imbalanced.

```
MAE: 10.273708990476212
MSE: 275.9527390853567
RMSE: 16.611825278558545
```

Fig 4: Training Outcomes

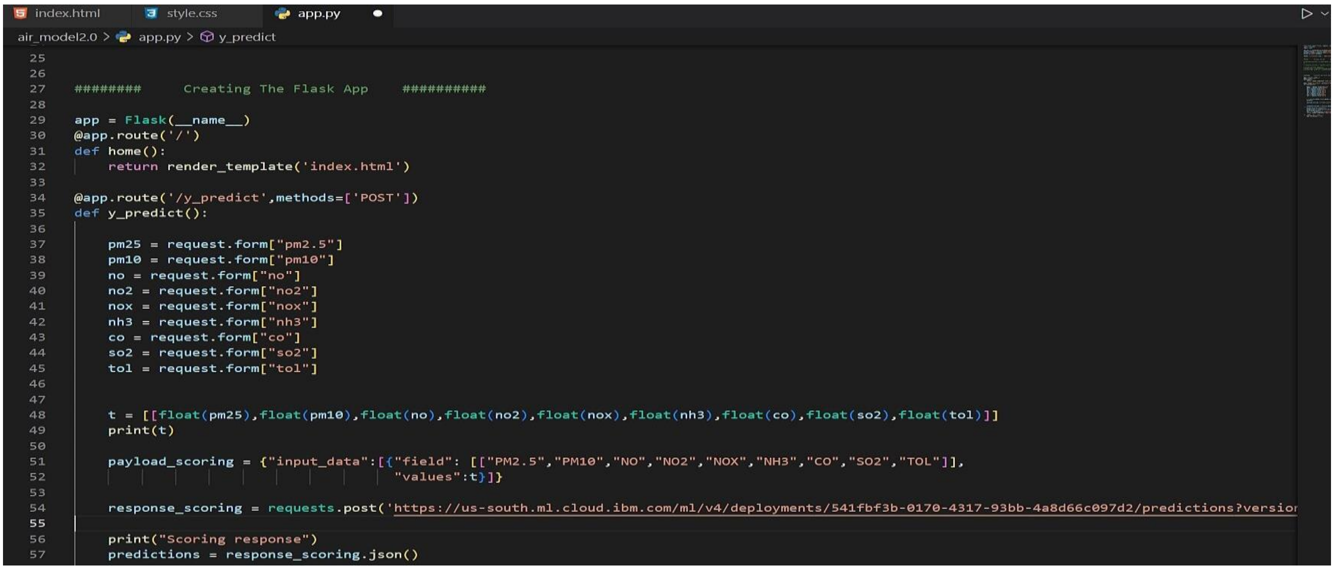
5. ***FLOWCHART***



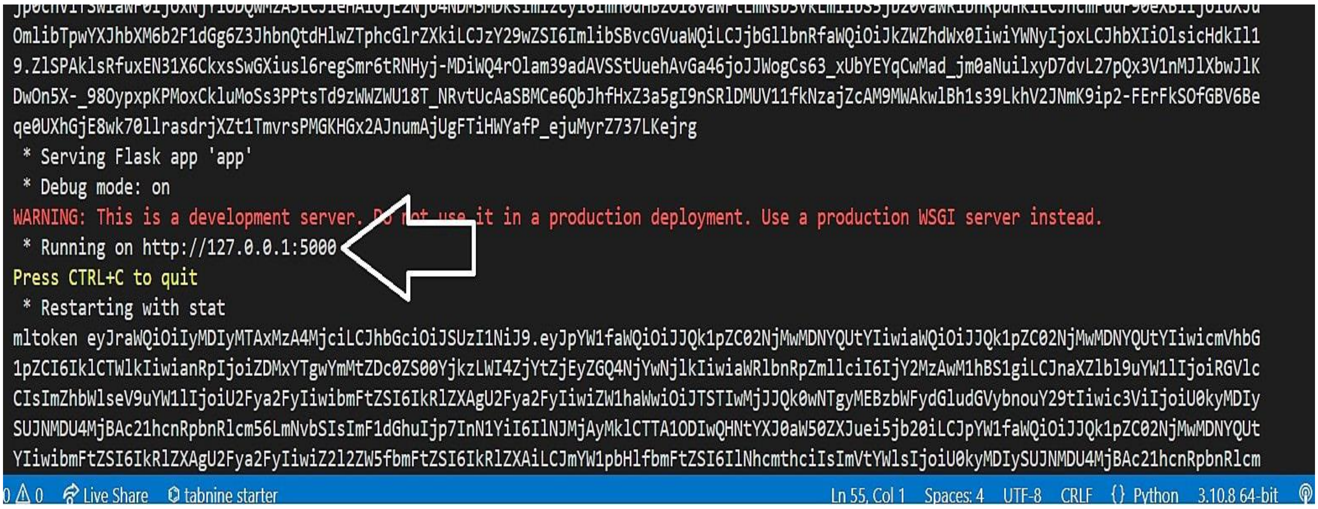
Data_Set
Heat_Map
PM2.5
PM10
NO
NO2
NOx
NH3
CO
SO2
Toluene
AQI
Parameter Extraction
Train-Test Split
Training Set
Testing Set
Random Forest Regression
Predicted Air Quality Index(AQI)
Performance
MAE: 10.273708990476212
MSE: 275.9527390853567
RMSE: 16.611825278558545

6. ***RESULTS:***

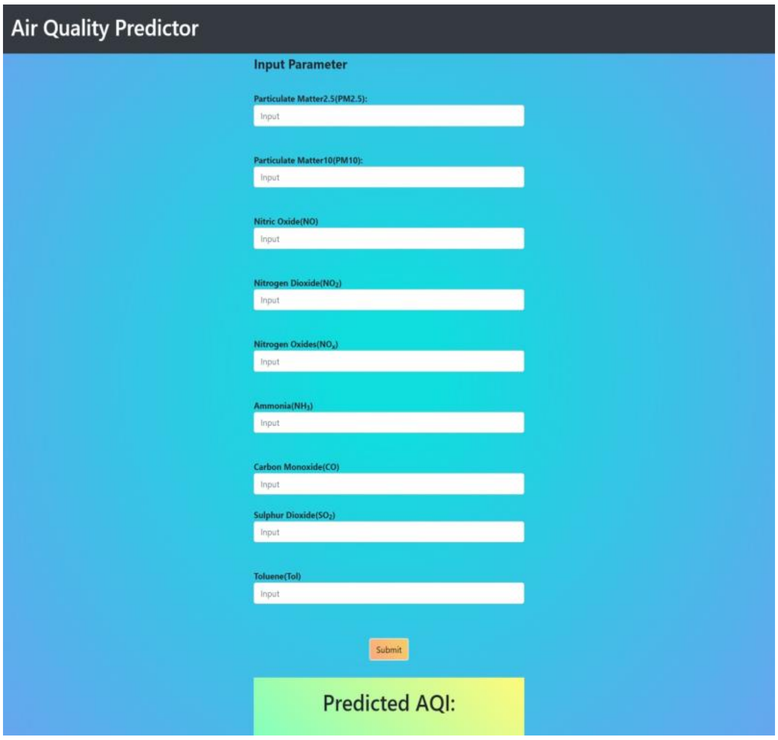
Since our model is capable of predicting the current data with 81% accuracy it will successfully predict the upcoming air quality index of any particular data within a given region. The value of **MAE** is 10.27, **MSE** is 275.95, and **RMSE** is 16.61. Using this model, we can forecast the AQI and alert the respected region of the country also it a progressive learning model it is capable of tracing back to the particular location needed attention provided the time series data of every possible region needed attention.



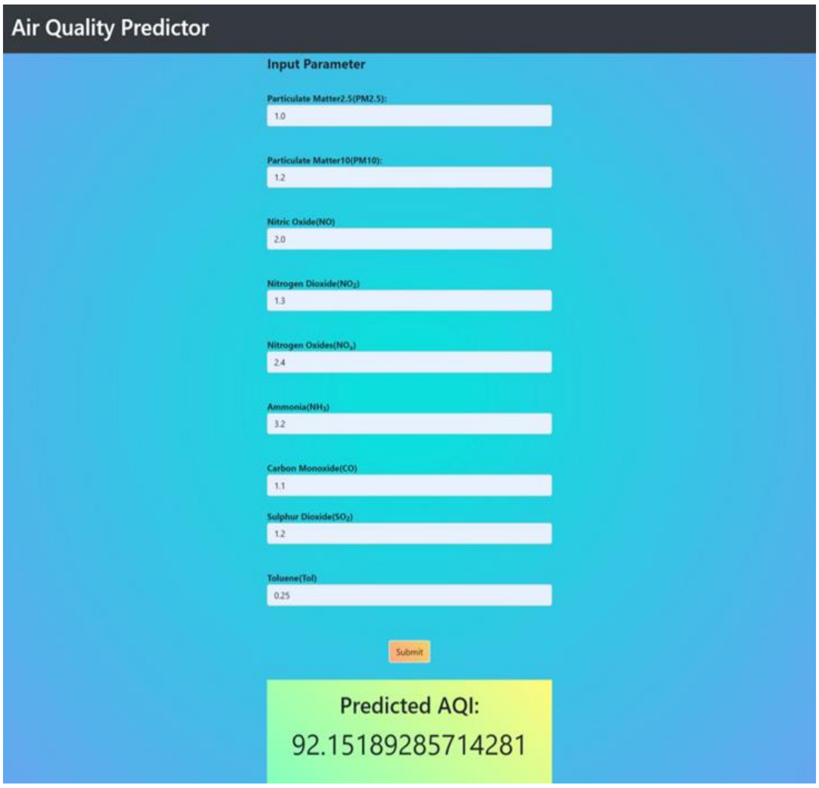
STEP 1: Run the Program



STEP 2: Copy the Link Pointed by the arrow



STEP 3: Paste the link in any Web Browser to get to the User Interface (UI)



STEP 4: Input the values for respective parameters and click on SUBMIT Option to get the prediction

7. ADVANTAGES & DISADVANTAGES:

Advantages	Disadvantages
➢ The proposed Methodology uses the algorithm of Random Forest Regression & Classification, which is the best algorithm for compared to other pre-existing algorithms.	➢ Our model has a slightly higher training time compared to other developed machine learning based algorithms which have lower accuracy.
➢ **RMSE**:16.62; **MAE**:10.27; **MSE**: 275.95 From these values we can say the model is performing efficiently.	➢ Random forest can feel like a black box approach for a statistical-modelers we have very little control on what the model does.
➢ The model can handle large datasets efficiently.	➢ When using a random forest, more resources are required for computation.
➢ It has an effective method for estimating missing data and maintains accuracy when large proportion of the data are missing.	➢ It doesn't predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy.

8. COMPARISON WITH OTHER ALGORITHMS:

For the contrast tests, Naïve Bayes, Logistic Regression, Single Decision Tree and ANN are chosen. Here we use Weka as the tool to conduct all the comparison tests. After realizing different algorithms, tests are carried out. Table 1. shows the results of the test cases in which Y means correct predictions and N means incorrect predictions. The precision is calculated by the formula Y/(Y + N) where Y is the number of correct predictions and N is the number of incorrect predictions. This figure shows RAQ (Random Forest Regression) performs steadily, even when the data size is relatively small. Other algorithms are less accurate at all time.

Table 1. Precision table of different algorithms.

Algorithm	Precision	Y	N
Naïve Bayes	52.1	1408	1493
Logistic	66.2	1790	911
Decision Tree	77.4	2092	609
ANN	71.8	1940	761
RAQ	81.5	2203	498

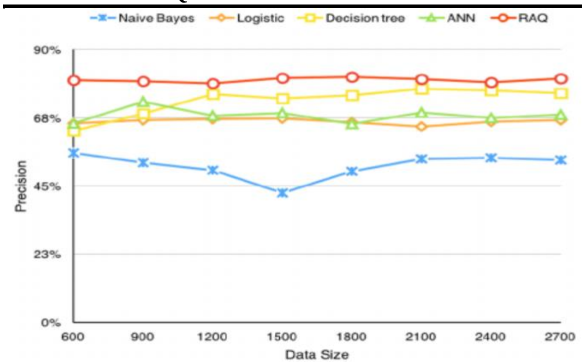


Fig 5: Precision changes according to data size.

9. *APPLICATIONS :*

- Predictions of air pollution and weather prediction are crucial information for airlines.
- Locations for research facilities which require a clean environment can be easily predicted by tracking air quality of various areas.
- With high pollution rates and all-time high emissions monitoring air quality is a must. If we don't track air quality in metropolitan areas and take necessary steps that might become extremely polluted before we know it.
- Tracking air quality in metropolitan areas, construction sites, near hospitals and studying its growth is essential to raise the standard of living.
- Indoor air pollution (IAP) is a relevant area of concern for most developing countries as it has a direct impact on mortality and morbidity since humans spend 80–90% of their routine time indoors, so indoor air quality (IAQ) leaves a direct impact on overall health and work efficiency.

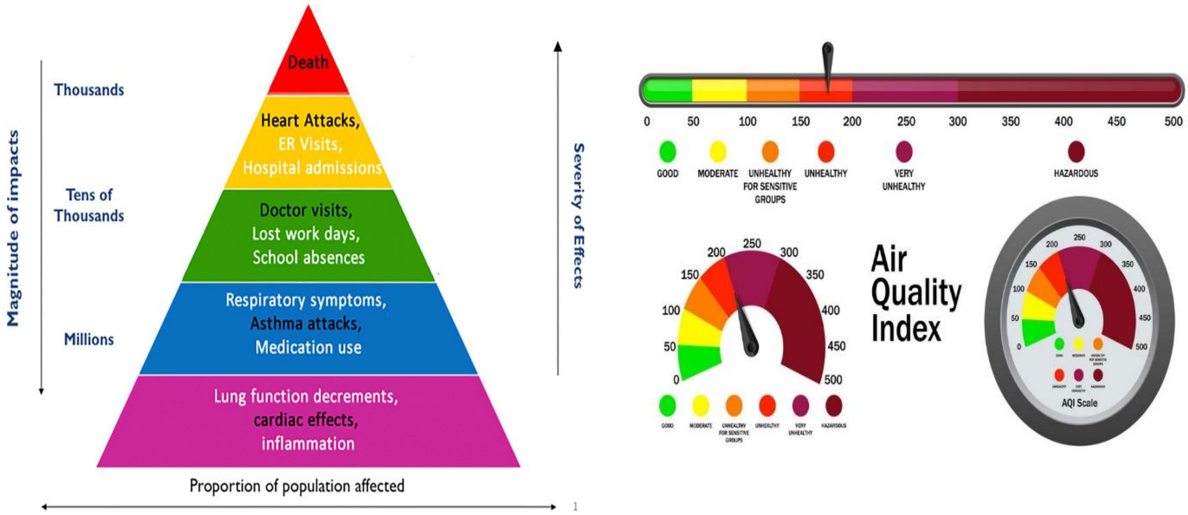


Fig 4: AQI Impacts on health

10. *CONCLUSION :*

Air pollution is a major environmental issue. It can affect the health and life support systems as well. Since clear air is an essential factor life for respiration, it is necessary to prevent the sources of air pollution. Due to air pollution destructive phenomena like acid rain, global warming etc.

The proposed model helps us in predicting air quality index using machine learning based Random Forest regression. This will help us to know the progression of real time air quality index in various places of the world specially Indian Sub-continent. Web-based applications can provide a valuable means to supply information and warn citizens. For instance, the user could also be alerted about the air quality level, and suggestions on how to react and improve air quality can be disseminated quickly by a mobile or a computer. Beside this, our proposed model is cheapest in current scenario. We also have a weather prediction system which will tell us when the air quality mixed with the weather situation of that placer to determine how detrimental it is for human health.

11. FUTURE SCOPE:

- ✓ Predictions of air pollution weather predictions are crucial information for airlines. Locations for research facilities which require a clean environment can be easily predicted by tracking air quality of various areas.
- ✓ With high pollution rates and all-time high emissions monitoring air quality is a must. If we don't track air quality in metropolitan areas and take necessary steps that might become extremely polluted before we know it.
- ✓ Tracking air quality in metropolitan areas, construction sites, near hospitals and studying its growth is essential to raise the standard of living. Our model predicts the pollution data in the near future to warn us about extreme dangers in the future.
- ✓ Real-time air quality monitoring systems are a booming research and investment area, and we have witnessed an increasing interest in topics like low-cost sensor networks, development of wireless communication, crowdsourcing, intelligent prognosis, distributed storage systems, together with a growing public awareness about air quality.

12. BIBLOGRAPHY:

- ➢ Anikender Kumar, PramilaGoyal, "Forecasting of air quality in Delhi using principal component regression technique", Atmospheric Pollution Research, 2 (2011) 436-444.
- ➢ Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models", International Journal of Engineering Trends and Technology (IJETT) - volume 59 Issue 4 – May 2018.
- ➢ YusefOmidiKhaniabadi, GholamrezaGoudarzi, Seyed Mohammad Daryanoosh, Alessandro Borgini, Andrea Tittarelli, Alessandra De Marco, "Exposure to PM10, NO2, and O3 and impacts on human health", Environ SciPollut Res, 2016.
- ➢ https://www.aqi.in/blog/aqi/
- ➢ https://www.kaggle.com/code/mihiranpasindu/air-qulity-prediction-final
- ➢ https://www.kaggle.com/code/kewalchavan/air-quality-index-prediction-using-svm