Presentation on:

Weather & Air Quality Index Prediction

Faculty: AZK

Md. Rashad Tanjim | ID: 1620952042

Introduction

What is the problem?

Weather plays a crucial role in our daily life. With the help of ML, we can predict the daily temperature, Air pollution Index and etc. The prediction can help us to know the every days weather updates.

For this we need a huge amount of data to train an algorithm for predicting.

Background Study

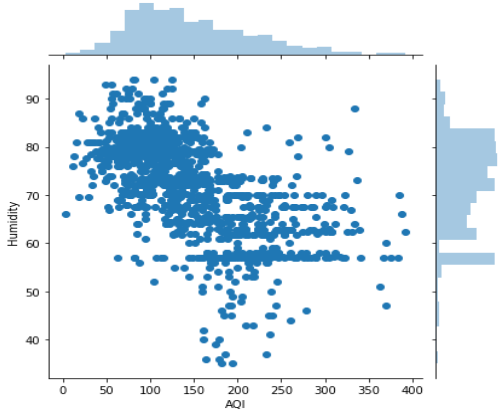
It is important to know what has been done in the current field of work, to get an overall picture where the field currently stands. My motivation is to built a model which can perfectly classify the AQI, Temperature and other weather parameters.

- P. Raj, “Prediction and optimization of air pollution-a review paper,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, pp. 3896–3904, 05 2019.
- A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, “An online air pollution forecasting system using neural networks,” *Environment international*, vol. 34, pp. 592–8, 08 2008.
- G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, “Air quality prediction: Big data and machine learning approaches,” *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.

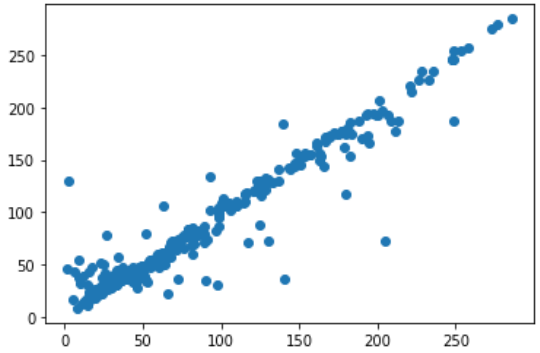
Datasets

For this project, I have used two datasets.

- ***Weather Dataset & Air pollution Dataset***
  - I have collected the weather dataset for Dhaka city from 2018 to 2020 from the website https://en.tutiempo.net/.
  - It had a total of 1300+ instances and 19 attributes.



Joint plot between AQI vs Humidity



Scatter Plot between y_test and prediction

Datasets (Cont.)

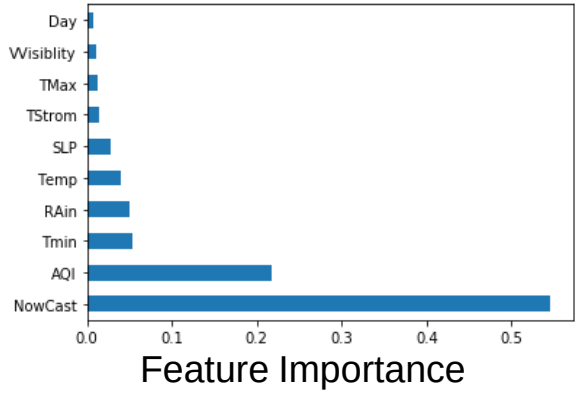
B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Month	Day	Temp	TMax	Tmin	SLP	Humidity	PPresure	VVisiblity	VavgWinc	VMWindS	VGspeed	RAin	SNow	TStrom	FoG	NowCast	AQI	Raw	AQI_CATE	Temp_Cat	AQI_CATEGORY	
1	1	21.1	27.6	14.8	1018.8	66	2.79	1.8	0.6	3.5	0	1	0	0	11	77.74	388	78.47	1	ok	HAZARDOUS	
1	2	21.5	27	14.2	1018	64	2.79	2.6	2.4	9.4	0	1	0	0	11	77.74	312	78.47	1	ok	HAZARDOUS	
1	3	20.9	25.8	14.8	1016.5	63	2.79	2.6	2.2	9.4	0	1	0	0	11	77.74	305	78.47	1	ok	HAZARDOUS	
1	4	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	307	78.47	1	cold	HAZARDOUS	
1	5	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	294	78.47	2	cold	VERY UNHEALTHY	
1	6	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	252	78.47	2	cold	VERY UNHEALTHY	
1	7	21.6	26.6	14.6	1016.6	57	2.79	2.6	0.7	3.5	0	1	0	0	11	77.74	253	78.47	2	ok	VERY UNHEALTHY	
1	8	21.2	26.8	14.4	1014.5	68	2.79	1.9	1.1	3.5	0	1	0	0	11	77.74	254	78.47	2	ok	VERY UNHEALTHY	
1	9	21.7	25	18.4	1013.2	80	2.79	1	0.9	3.5	0	1	0	0	11	77.74	304	78.47	1	ok	HAZARDOUS	
1	10	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	325	78.47	1	cold	HAZARDOUS	
1	11	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	274	78.47	2	cold	VERY UNHEALTHY	
1	12	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	242	78.47	2	cold	VERY UNHEALTHY	
1	13	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	307	78.47	1	cold	HAZARDOUS	
1	14	20.3	25.2	14.2	1015	57	2.79	1.6	0	5.1	0	1	0	0	11	77.74	304	78.47	1	ok	HAZARDOUS	
1	15	21.2	26	14	1012.6	57	2.79	1.9	0.7	3.5	0	1	0	0	11	77.74	375	78.47	1	ok	HAZARDOUS	
1	16	19.2	26.6	14.9	1012.2	69	2.79	1.4	1.1	3.5	0	1	0	0	11	77.74	307	78.47	1	cold	HAZARDOUS	
1	17	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	331	78.47	1	cold	HAZARDOUS	
1	18	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	306	78.47	1	cold	HAZARDOUS	
1	19	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	290	78.47	2	cold	VERY UNHEALTHY	
1	20	19.1	22.8	16.5	1014.6	74	2.79	1.9	4.6	5.4	0	1	0	0	11	77.74	103	78.47	3	cold	SENSITIVE	
1	21	18.9	22.7	15.6	1015.9	74	2.79	1.9	3.5	7.2	0	1	0	0	11	77.74	102	78.47	3	cold	SENSITIVE	
1	22	17.3	20	14.8	1017.5	75	2.79	1.9	8.9	18.3	0	1	0	0	11	77.74	143	78.47	3	cold	SENSITIVE	
1	23	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	131	78.47	3	cold	SENSITIVE	
1	24	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	182	78.47	4	cold	UNHEALTHY	
1	25	19.9	24.8	14.3	1015.2	62.4	2.79	2	1.7	5.1	0	1	0	0	11	77.74	306	78.47	1	cold	HAZARDOUS	
1	26	16.2	22.1	10.2	1015.1	58	2.79	1.9	1.1	3.5	0	1	0	0	11	77.74	301	78.47	1	cold	HAZARDOUS	

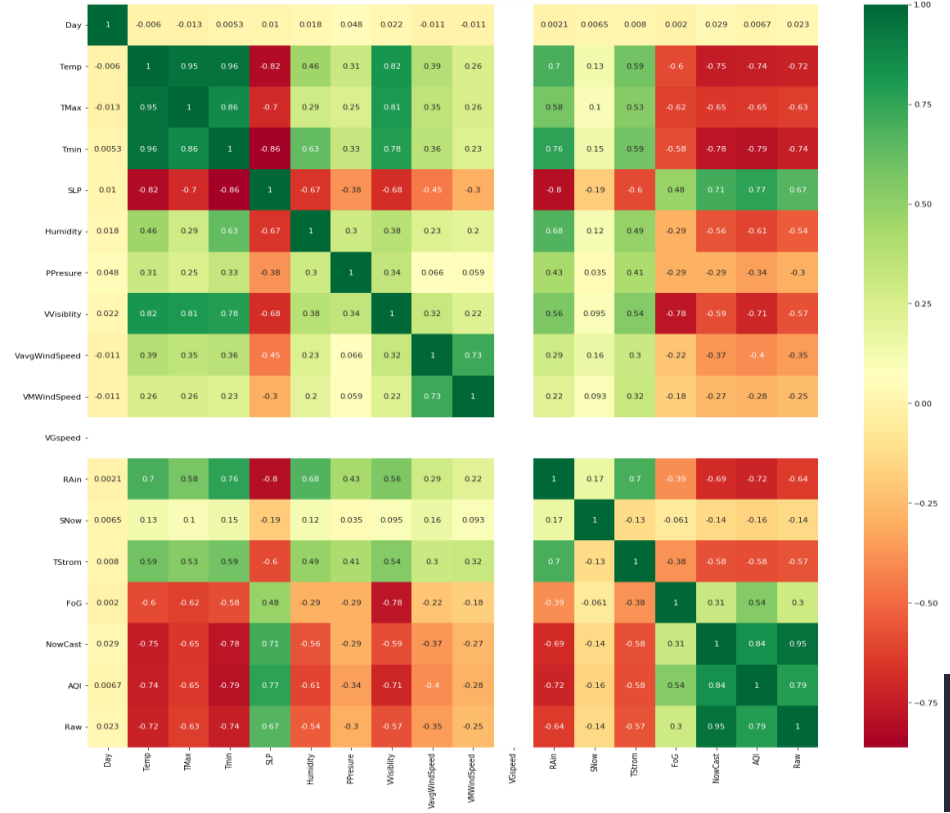
Dataset Screenshot

Dataset Preprocessing

The raw dataset contained many missing as well as repeated values. I have preprocessed the data for correct analysis. The steps followed were:

- **Replacement**
- **Normalization of the Attributes**
- **Feature Importance**
- **Correlation Matrix**





Model Selection- First thing I tried, first result I got

- Resampling method is used (cross validation) as model selector.
- The table describes the root mean square error of different models for selection

	precision	recall	f1-score	support
GOOD	0.00	0.00	0.00	10
HAZARDOUS	0.00	0.00	0.00	13
MODERATE GOOD	0.66	0.42	0.51	100
SENSITIVE	0.44	0.83	0.58	109
UNHEALTHY	0.87	0.53	0.66	78
VERY UNHEALTHY	0.89	0.82	0.85	71
accuracy			0.61	381
macro avg	0.48	0.43	0.43	381
weighted avg	0.64	0.61	0.59	381

CONFUSION MATRIX OF SVR

Model	RMSE
Logistics Regression	16.70
Xgboost Regressor	16.81
Random Forest Regressor	17.30
SVR	16.97

DIFFERENT Model Evaluation for model selection

Methodology– Motivation to use another method

This table shows the different metrics that were considered for evaluating the models for Decision Tree.

Rules	Class level
NowCast <= 62.47 and NowCast <= 34.37 and Raw > 133.87	Good
NowCast > 62.47 and NowCast > 140.8 and NowCast <= 248.7	Very Unhealthy
NowCast > 62.47 and NowCast <= 140.8 and Vvisiblity <= 2.9 and NowCast <= 82.4 and VMWindSpeed > 2.8	Sensitive
NowCast > 62.47 and NowCast <= 140.8 and Vvisiblity >2.9 and Tmin <= 26.1 and Raw <= 72.79 and Tmax <= 30.2	Unhealthy
NowCast > 62.47 and NowCast <= 140.8 and Vvisiblity > 2.9 and Tmin > 26.1 and SLP > 1004.7 and PPressure > 305.64	Moderate Good
NowCast > 62.47 and NowCast > 140.8 and NowCast > 248.73	Hazardous

Methodology – Hyper-parameter Tuning, Improvement

- I have used Resampling method as Cross Validation techniques.
- I have imported (from sklearn.model_selection import RandomizedSearchCV)
- Use the Random grid to search for best hyperparameters
- Manually changing Number of trees in random forest
- Number of features to consider at every split
- Maximum number of levels in tree
- Random search of parameters, using 3 fold cross validation
- Search across 100 different combinations
- Best params = {'n_estimators': 300, 'min_samples_split': 5, 'min_samples_leaf': 5, 'max_features': 'auto', 'max_depth': 15}

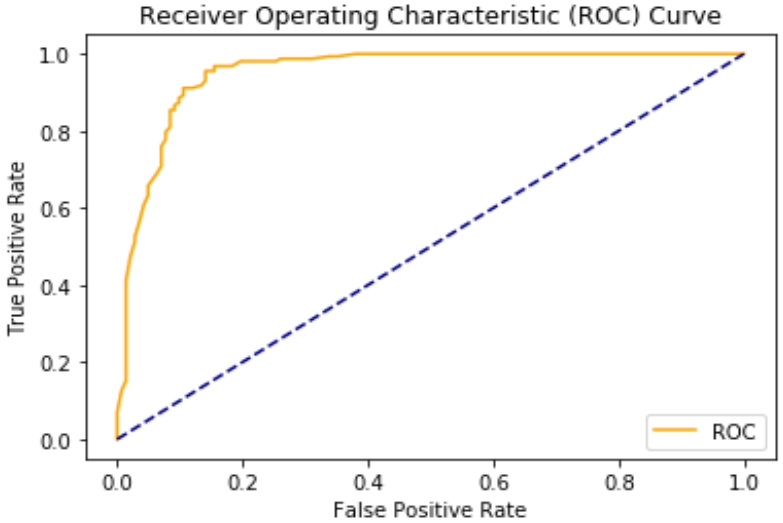
Result Analysis - Improvements using model selection techniques

Among all classifiers the Random Forest gives the best result, so I have selected it for further analysis.

	precision	recall	f1-score	support
GOOD	0.00	0.00	0.00	7
HAZARDOUS	0.00	0.00	0.00	8
MODERATE GOOD	0.86	0.90	0.88	71
SENSITIVE	0.83	0.87	0.85	71
UNHEALTHY	0.71	0.74	0.72	57
VERY UNHEALTHY	0.76	0.88	0.81	40
accuracy			0.80	254
macro avg	0.53	0.56	0.55	254
weighted avg	0.75	0.80	0.77	254

CONFUSION MATRIX OF RANDOM FOREST

Result Analysis (Cont.) – ROC



A receiver operating characteristics curve for Random Forest

Result Analysis (Cont.) – Using Built-in Tools

	TP Rate	FP Rate	Precision	Recall	F meas.	ROC	Class
	0.841	0.056	0.833	0.841	0.837	0.960	A
	0.889	0.048	0.856	0.889	0.872	0.977	B
	0.758	0.022	0.823	0.758	0.789	0.933	C
	0.842	0.028	0.818	0.842	0.830	0.958	D
	0.871	0.019	0.859	0.871	0.865	0.971	E
	0.807	0.023	0.854	0.807	0.830	0.953	F
Weight Avg,	0.841	0.037	0.841	0.841	0.841	0.961	

The model matrix on class level for all categories of AQI.

Here,

A = Hazardous,

B = Very Unhealthy,

C = Moderate,

D = Good,

E = Moderate Unhealthy,

F = Unhealthy

Conclusion, Limitations & Future Work

- I have learnt the different ML model and algorithms by doing this project.
- I have applied different types of machine learning algorithms to predict the condition of weather based on previous air pollution and weather data.
- Random forest classifier gives the best accuracy of 80%.
- Analyzing this attributes I can more accurately predict the Air Quality Index and daily pollution rate as well as Temperature & Humidity.
- More instance of data and attributes might gives a better analysis of this project on highly polluted city like Dhaka.
- I will find patterns and relation how greenery and weather pollution is related to each other, how Covid changes our environment pattern recently and find interesting patterns such as the AQI level during the holidays.

References - Sources used to do this project

- P. Raj, “Prediction and optimization of air pollution-a review paper,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, pp. 3896–3904, 05 2019
- G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, “Air quality prediction: Big data and machine learning approaches,” *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.
- Dhaka Weather Datatset, 2016 to 2019, Tutiempo Network, S.L., 2019. [Online]. Available: https://en.tutiempo.net/climate/ws-419230.html. Accessed: 2019-12-11
- Air Pollution Datatset, 2016 to 2019, National Oceanic and Atmospheric Administration., USA. [Online]. Available: https://www.airnow.gov/index.cfm?action=airnow.global_summary#Bangladesh$Dhaka. Accessed: 2019-12

Thank You!

Any Question?