2022

HACETTEPE
University

VBM 655
İstatistiksel Veri Analizi

BUILDING AN ML MODEL TO PREDICT THE FUTURE VALUES OF SUSTAINABILITY INDEX FROM THE AIR POLLUTION MEASUREMENT DATASET

OGÜN ŞERIF ONARGAN
N22137400

1. Introduction

Sustainability Index is an indicator to measure air pollution. It is calculated by using values of some major air pollutants. It doesn't have standardized calculation method. Each approach offers different methods for different purposes. For example, a research which aims to measure air pollution after El Nino disaster considers CO, O3, NO2, SO2 and PM10 with some transformations that calculating index value for each pollutant and maximum index is taken as SI index. [1]. On the other hand, some environmental authorities, such as World Health Organization (WHO), EU Air Quality Directive, EU Common Air Quality Index (CAQI) considers subset of Ozone, PM, NO2, SO2, CO and organic compounds. The calculation groups each pollutant and inferred a category of SI [2]. Likewise, calculated sustainability index can be put forward as both numerical and categorical value. Calculation method of target value (SI) is not given as expected, so one of main goal of the project is to find which features are put into the calculation bin.

In addition, after any burning process, varied ratio of gases are released. For example, quantity of oxygen feeding in a process determine ratio of N1, N2, N3 and N4. This causes correlation between each independent and indirectly dependent features. If an exact formula is applied to SI calculation, unaccounted features may hide behind high correlation. The second main aim is eliminating unrelated features.

2. Explanatory Data Analysis

a. Dataset

	Id	City	Date	WP1	WP2	N1	N2	N3	N4	Carbon	Ozone	Sulpher	OC1	IN1	OC2	IN2	SI
0	1	LosAngeles	02-12-2015	NaN	NaN	1.99	12.60	20.99	NaN	0.26	37.78	4.99	2.32	1.80	0.77	17815.96	112.0
1	2	Bakersfield	03-06-2018	37.57	103.65	11.61	27.60	35.45	NaN	1.18	55.73	15.08	NaN	NaN	0.49	18370.96	107.0
2	3	Shreveport	18-11-2019	75.41	137.06	3.97	16.67	11.05	NaN	11.59	23.68	74.52	14.68	2.52	17.18	1379.96	173.0
3	4	Camden	06-08-2017	73.86	NaN	8.47	25.58	NaN	NaN	0.60	40.06	7.05	0.00	NaN	0.00	14763.96	392.0
4	5	Mexico	06-02-2020	23.58	43.8	2.18	8.04	7.55	1.36	0.58	32.57	8.30	2.78	0.11	0.05	7346.96	61.0

Table (1): train Dataset

	Id	City	WP1	WP2	N1	N2	N3	N4	Carbon	Ozone	Sulpher	OC1	IN1	OC2	IN2
0	16001	9	133.46	88.68	27.62	27.72	28.67	15.31	2.06	33.94	48.01	19.55	10.89	3.53	19134.96
1	16002	13	43.15	96.04	29.11	38.54	43.48	56.60	0.42	32.53	8.30	6.07	2.53	1.29	4981.96
2	16003	3	135.36	88.68	5.25	3.51	22.24	15.31	0.76	17.01	4.86	1.94	1.11	0.03	16643.96
3	16004	19	45.53	88.68	9.51	20.24	22.24	15.31	0.89	27.68	8.76	2.79	1.11	0.90	19117.96
4	16005	16	36.60	85.25	8.84	25.79	34.19	55.99	0.84	31.31	10.31	14.60	2.26	1.24	5306.96

Table (2): test Dataset

Id	It is the series number of rows
City	City names
Date	Date of the values recorded
WP1	It stands for Weather_parameter-1
WP2	It stands for Weather_parameter-2
N1 ,N2,N3,N4	These are the compounds of Nitrogen like No2,No,etc.
Carbon	level of Carbon contaminant
Sulphur	level of Sulphur contaminant
Ozone	level of Ozone
IN1,IN2	level of Inert gases.
OC1, OC2	level of Organic compounds
SI	Sustainability Index the higher its value the higher is the damage to health

Table (3): Features and their Descriptions

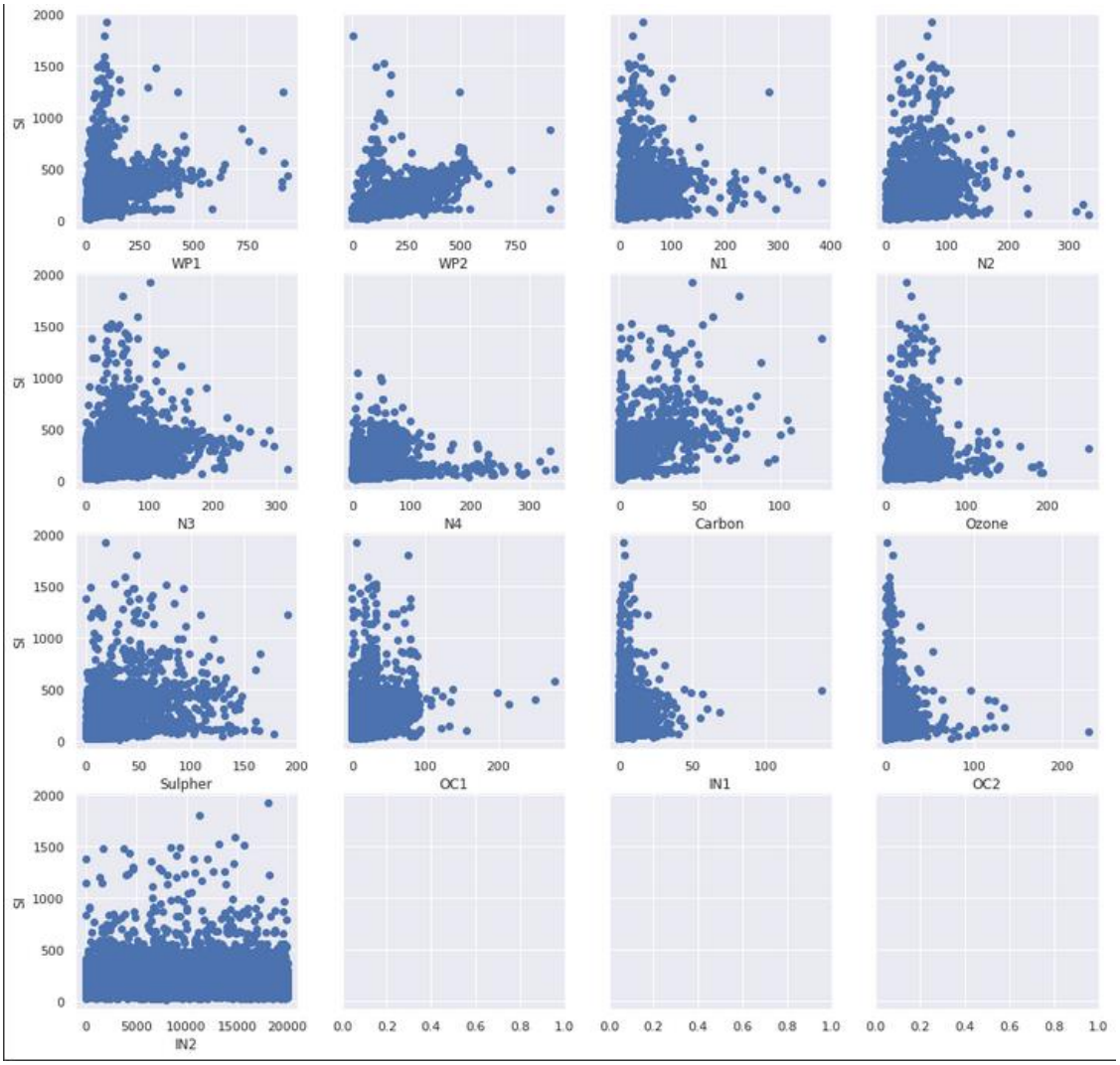
- Train dataset has 16.000 rows and test dataset has 4.000 rows. They should be investigated together, so the datasets are concatenated.
- Test dataset hasn't got any missing values.
- Date column does not exist in test dataset. It is removed.
- Date and WP2 values are not in correct type, so the column must be corrected.
- There are some missing values at almost each column. Especially, if IN1 is not useful, it should be removed.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20000 entries, 0 to 3999
Data columns (total 18 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Id          20000 non-null  int64
 1   City        20000 non-null  object
 2   Date        15998 non-null  object
 3   WP1         18488 non-null  float64
 4   WP2         14431 non-null  float64
 5   N1          19128 non-null  float64
 6   N2          19139 non-null  float64
 7   N3          18347 non-null  float64
 8   N4          15211 non-null  float64
 9   Carbon      19396 non-null  float64
 10  Ozone       19044 non-null  float64
 11  Sulpher     18969 non-null  float64
 12  OC1         16050 non-null  float64
 13  IN1         10181 non-null  float64
 14  OC2         17365 non-null  float64
 15  IN2         20000 non-null  float64
 16  SI          16000 non-null  float64
 17  train_test  20000 non-null  int64
dtypes: float64(14), int64(2), object(2)
memory usage: 2.9+ MB
```

```
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Id       16000 non-null  int64
 1   City     16000 non-null  object
 2   Date     15998 non-null  object
 3   WP1      14488 non-null  float64
 4   WP2      10432 non-null  object
 5   N1       15128 non-null  float64
 6   N2       15139 non-null  float64
 7   N3       14347 non-null  float64
 8   N4       11211 non-null  float64
 9   Carbon   15396 non-null  float64
 10  Ozone    15044 non-null  float64
 11  Sulpher  14969 non-null  float64
 12  OC1      12050 non-null  float64
 13  IN1      6181 non-null   float64
 14  OC2      13365 non-null  float64
 15  IN2      16000 non-null  float64
 16  SI       16000 non-null  float64
dtypes: float64(13), int64(1), object(3)
memory usage: 2.1+ MB
```

Table (4): (Left): Concatenated DataFrame Info, (Right): Training Dataset Info

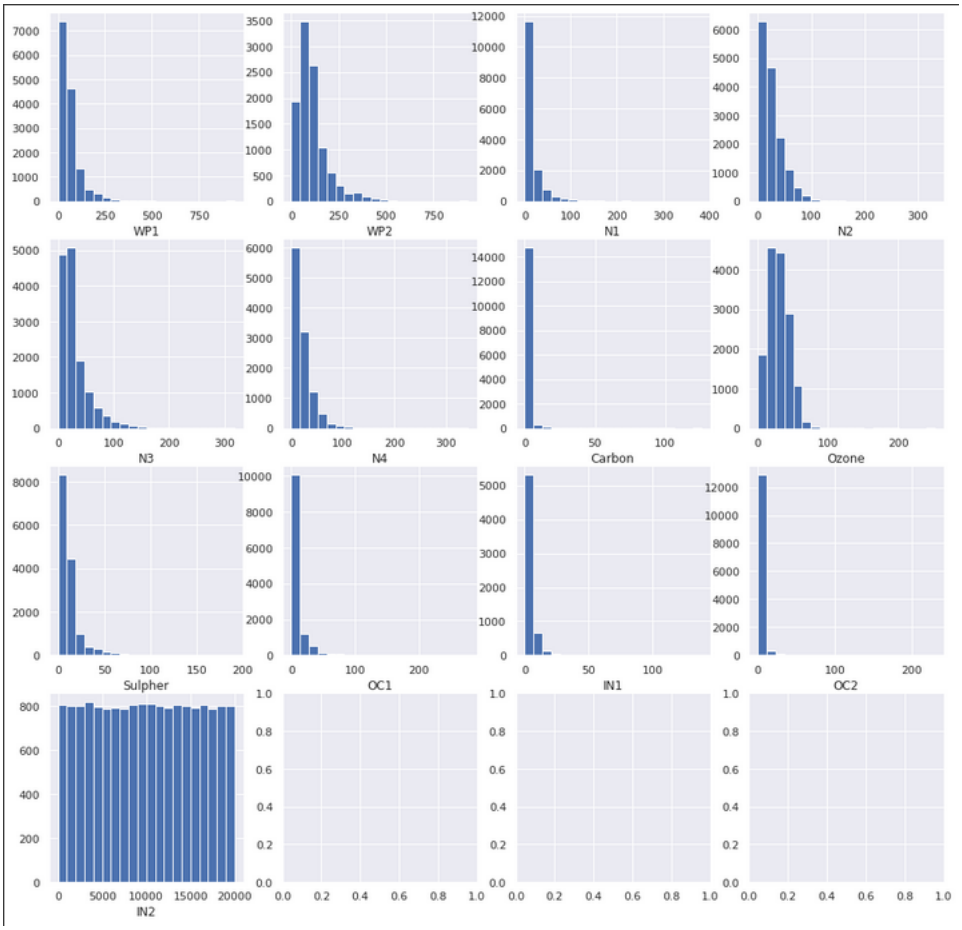
b. Numerical Columns: ScatterPlots



Graph (1): Scatter plot of raw data

- IN2 looks like uniformly distributed.
- WP1, WP2 and Carbon looks like strongly correlated.
- Bunching generally occurs at low values. All columns, except IN2, probably have right skewed distribution. Therefore, outlier detection should be done after data transformation.
- In Graph (1), some outliers are observed clearly.

c. Numerical Columns: Histograms



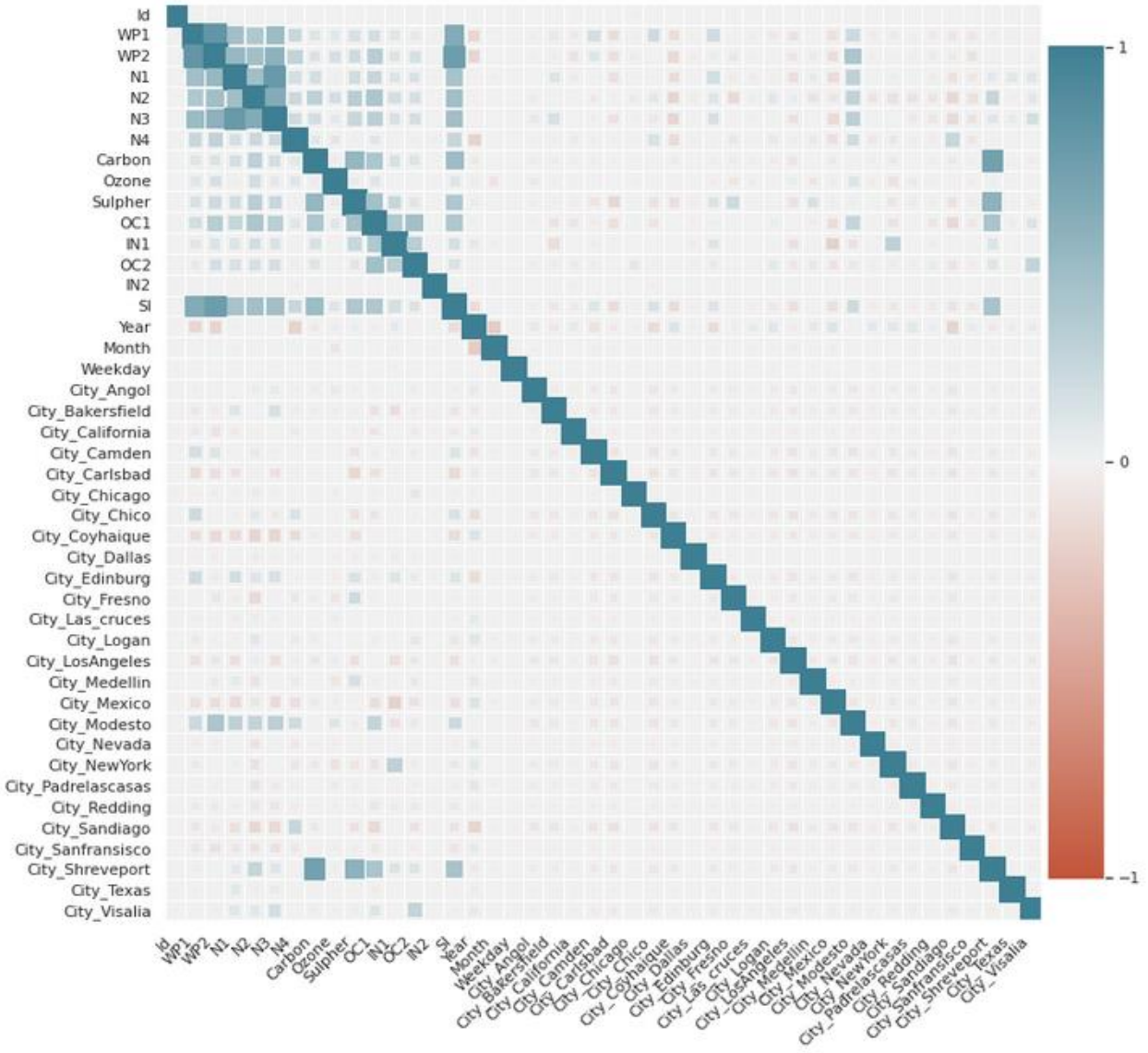
Graph (2): Histogram of raw data

- As it is observed in previous graph, all columns, except Ozone and IN2, have right skewed distribution.
- IN2 has uniform distribution, so it doesn't provide any information about SI. It is deleted.
- Ozone has nearly normal distribution.
- Box-Cox transformation will be applied to make distributions normal.

d. Columns: City and Date

- The dataset contains Id, City, Date, air content measurements and Sustainability index (SI) values.
- To analyze City, OneHotEncoder is applied.
    - City Column has different labeling method in training and test dataset. Student's t-Test is applied to the column with the aim of matching numbers and city names. Although City column provides effective information about target variable, Student's t-Test does not provide valid outcomes. Then, it is removed.
- To analyze Date, date is separated into smaller parts (Weekdays Name, Month, Years)

e. Correlation Table



Graph (3): Correlation Matrix of train Dataset

- Some of variables have strong correlation with not only target variable, but also other independent variables. Correlation between independent variables will support handling missing values to impute these values.
- Year, Month and Weekday do not correlate with any variable. Deletion does not affect the model.

3. Handling Missing Values

During modeling phase of report, an exact SI calculation formula will be considered. Therefore, filling missing values with a number like mean and median or imputing with a function like IterativeImputer() and KNNImputer() is not best practice for the model, because each method will add an error to calculation. Removing missing values is best practice for the strategy.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7949 entries, 4 to 3999
Data columns (total 18 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Id          7949 non-null   int64
 1   City        7949 non-null   object
 2   Date        3949 non-null   object
 3   WP1         7949 non-null   float64
 4   WP2         7948 non-null   float64
 5   N1          7949 non-null   float64
 6   N2          7949 non-null   float64
 7   N3          7949 non-null   float64
 8   N4          7949 non-null   float64
 9   Carbon      7949 non-null   float64
 10  Ozone       7949 non-null   float64
 11  Sulpher     7949 non-null   float64
 12  OC1         7949 non-null   float64
 13  IN1         7949 non-null   float64
 14  OC2         7949 non-null   float64
 15  IN2         7949 non-null   float64
 16  SI          3949 non-null   float64
 17  train_test  7949 non-null   int64
dtypes: float64(14), int64(2), object(2)
memory usage: 1.4+ MB
```

After the deletion, 3949 rows are remained. Test dataset has no missing value, because the missing values in test dataset has already filled with mean values.

Table (5): Concatenated Dataset Info after Removing Missing Values

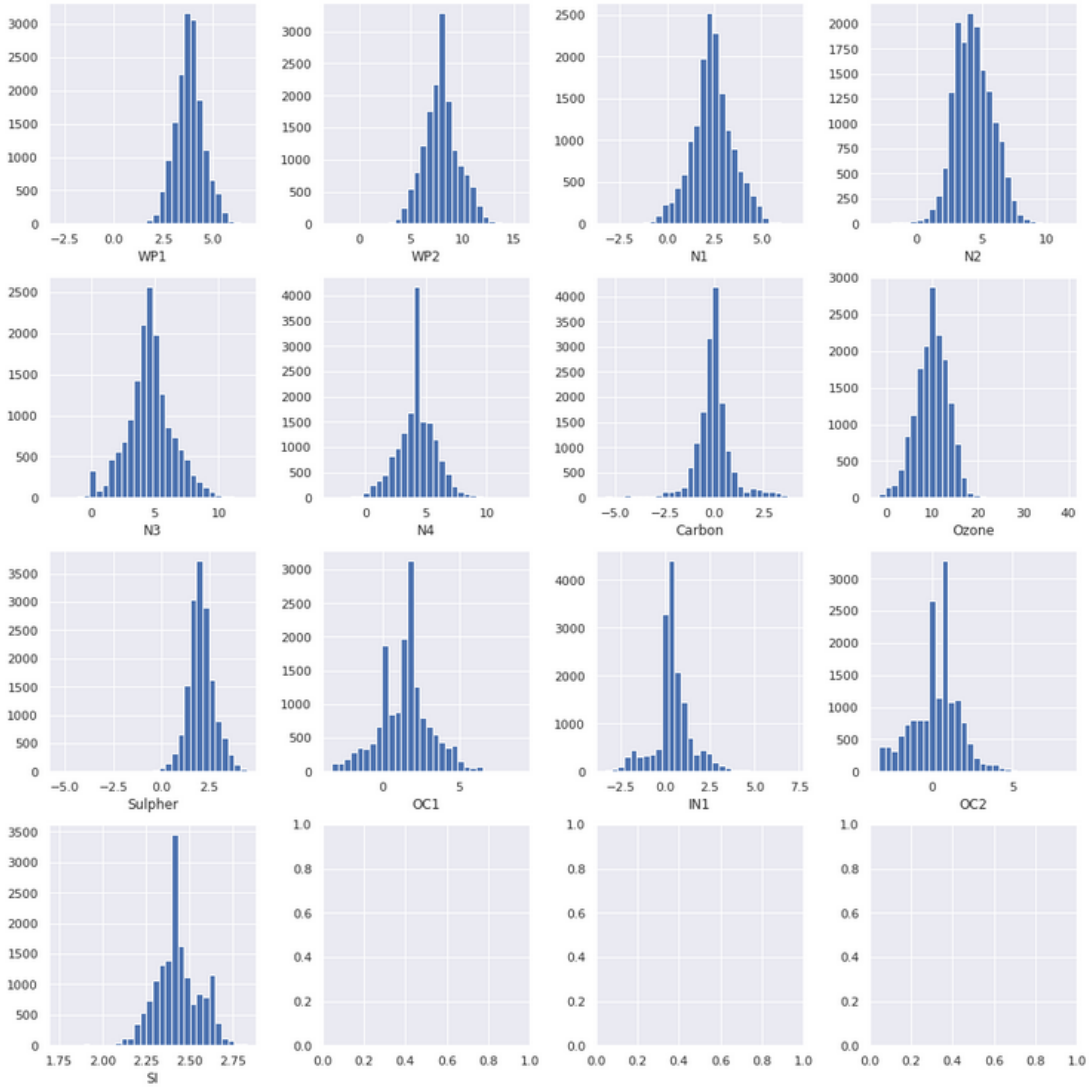
4. Data Transformation

BoxCox Transformation is used to normalize all numerical features of dataset. Values under zero which is pre-request of BoxCox transformation is satisfied by replacing non-positive values with 1. The value is chosen, because all variables' mean values are much higher than 1 and the value is ineffective for each transformation and lambda value of BoxCox transformation.

Column Name	Lambda (Box-Cox)
WP1	0.015975202183214524
WP2	
N1	0.15427374826450263
N2	0.0039005034036939086
N3	0.23724789388111028
N4	0.1629683231587555
Carbon	0.1668700576976145
Ozone	0.05177816752493404
Sulpher	0.5761107201985739
OC1	0.03488992111180966
OC2	0.1423128801181148
SI	0.26993321618736016

Table (6): Box-Cox Transformation Lambda Values

After the normalization, histograms formed as below,



Graph (4): Histogram of Transformed Concatenated Data

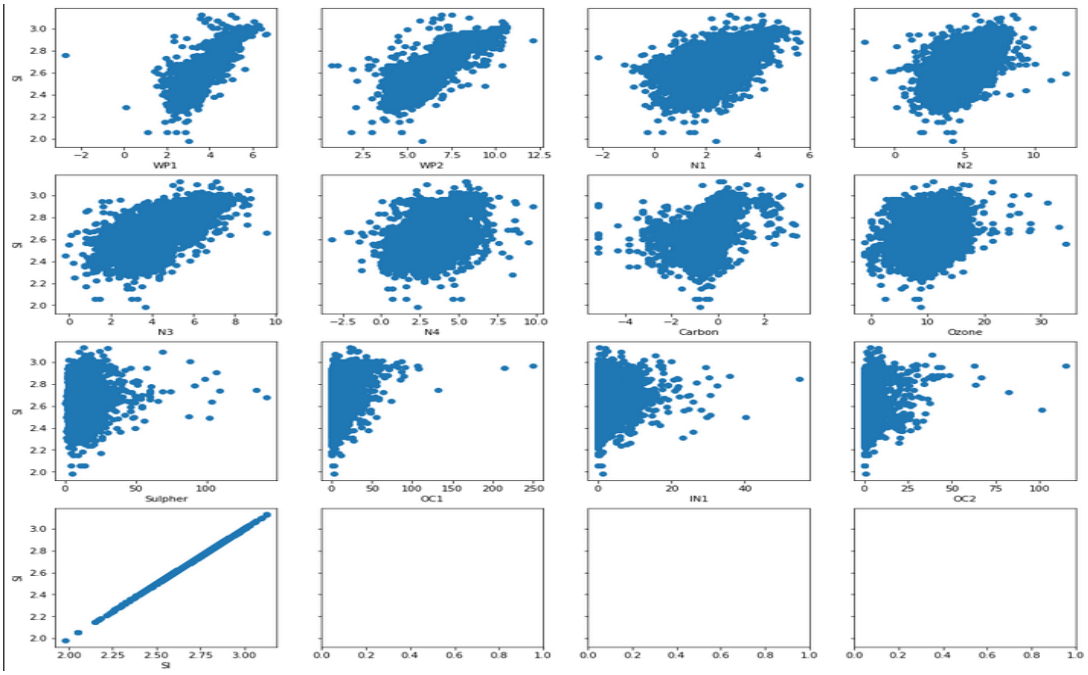
To check normality of features' distributions, Shapiro-Wilk Test is applied. Test's success condition is,
If Shapiro Result>0.85 and P-Value<0.05 then distribution is normal.

	Shapiro Result	**P-Value**
WP1	0.993	3.428e-26
WP2	0.992	2.461e-29
N1	0.994	2.268e-25
N2	0.992	2.191e-28
N3	0.987	4.792e-35
N4	0.982	7.013e-41
Carbon	0.891	0.0
Ozone	0.988	4.251e-34
Sulpher	0.984	3.178e-38
OC1	0.987	8.728e-36
IN1	0.934	0.0
OC2	0.980	1.760e-42
SI	0.987	1.120e-35

Table (7): Shapiro Result and P-values of Transformed Concatenated Dataset

All numerical features succeed Shapiro-Wilk Test after Box-Cox transformation.

5. Handling Outliers

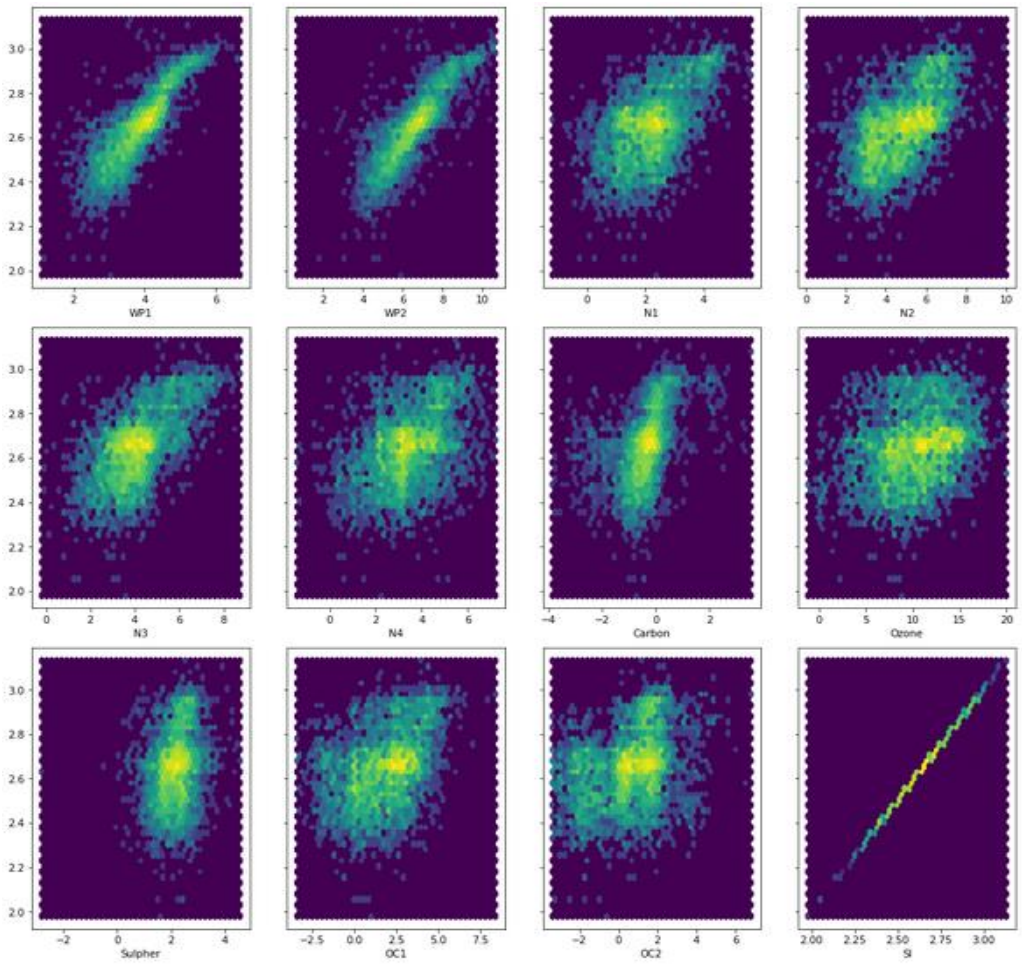


Graph (5): Scatter Plot of Transformed Concatenated Dataset

As it can be seen in the plot, there are some obvious outliers that effect the model. Heuristic way is chosen to remove them, because air pollutant data is collected from sensors and an exact formula is applied to get SI value. Therefore, just the most marginal outliers are deleted. The outlier conditions are mentioned below.

```
train[train["WP1"]<0.2] = np.nan
train[(train["WP2"]<-2) | (train["WP2"]> 12)] = np.nan
train[train["N1"]<-2] = np.nan
train[(train["N2"]<0) | ((train["N2"]>8) & (train["SI"]<2.6))] = np.nan
train[train["N3"]>9] = np.nan
train[(train["N4"]>7.5) | (train["N4"]<-1.5)] = np.nan
train[train["Carbon"]<-4] = np.nan
train[train["Ozone"]>20] = np.nan
train[train["Sulpher"]>60] = np.nan
train[train["OC1"]>100] = np.nan
train[train["IN1"]>35] = np.nan
train[train["OC2"]>50] = np.nan
```

Target column's outliers are deleted according to 75,25 percentiles.



Graph (6): Hexbin Plot of Transformed Concatenated Data

6. Model Development

As mentioned in domain knowledge phase, the dataset has features that relations between independent variables and target variables is not clear, because of correlations between independent variables and uncertainty of Sustainability formulation. To investigate the useful features, Elastic Net and XGBRegressor is used to compare models. Also, Sequential Feature Selection is used to investigate effects of fully removing useless features.

a. Model Selection

Columns that selected by SFS,

```
Index(['WP1', 'WP2', 'N1', 'Carbon', 'Ozone', 'OC1'], dtype='object')
```

Elastic Net Regression Formula,

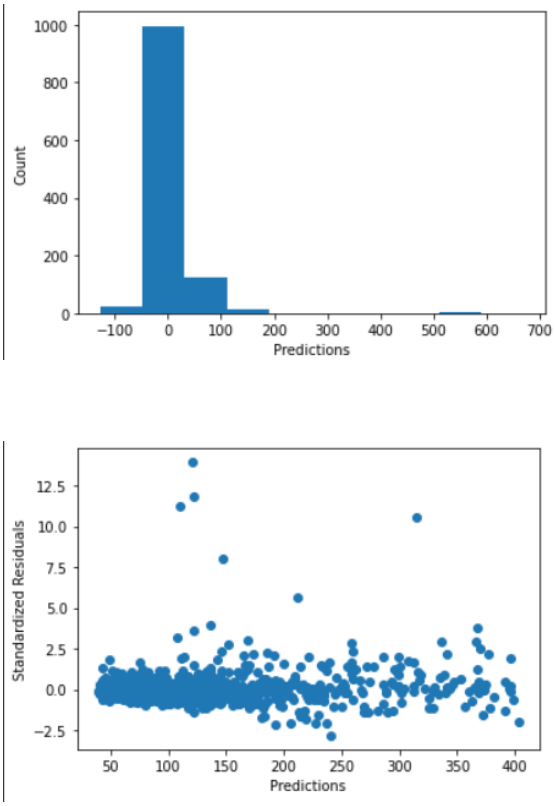
$SI = 2.647$ + **(6.23e-2 x WP1)** + **(5.77e-2 x WP2) + (1.21e-2 x N1)** + (0.05e-2 x N2) + (0.48e-2 x N3) + (0.61e-2 x N4) + **(0.92e-2 x Carbon)** + **(0.89e-2 x Ozone)** + (0.31e-2 x Sulpher) + **(0.45e-2 x OC1)** + (-0.07e-2 x OC2)

	Best Hyper parameters	**RMSE (invBoxCox)**	**R2**
Elastic Net	'alpha': 0.01, 'l1_ratio': 0.0	29.09	0.75
Elastic Net + SFS	'alpha': 0.01, 'l1_ratio': 0.0	29.50	0.75
XGBRegressor	'max_depth': 6, 'learning_rate': 0.0941548854691756, 'n_estimators': 753, 'min_child_weight': 10, 'gamma': 0.0533596801197716, 'subsample': 0.410840191638505, 'colsample_bytree': 0.784814607701741, 'reg_alpha': 0.13227013596142556, 'reg_lambda': 0.33658782573955054	27.24	0.78
XGBRegressor + SFS	'max_depth': 10, 'learning_rate': 0.01126420636455537, 'n_estimators': 644, 'min_child_weight': 9, 'gamma': 0.033745571154095905, 'subsample': 0.6888748757231115, 'colsample_bytree': 0.8548546808182247, 'reg_alpha': 0.1855749804376162, 'reg_lambda': 0.7195896261105815	27.90	0.78

Table (8): Machine Learning Models' Results

According to R2 and RMSE scores, it can be claimed that L1 and L2 regularization did their job and get better performance than SFS.

b. Residual Analysis



Histogram of residuals shows us residuals are nearly normal distributed. However, there are some outliers that affect RMSE heavily.

The reasons may be [3],

- The solution to this is almost always to transform your data, typically your response variable.
- It's also possible that your model lacks a variable.

First reason has already done.

For second reason, new feature creation can be done, but number of residual's outliers are 5, so 6they can be count as exceptions.

Except these outliers, the model is valid.

Graph (7): Residual Analysis – Scatter Plot and Histogram of Residuals

The model is applied to test dataset which is provided for submission of Kaggle Competition and result becomes 61.06 (RMSE). It is fairly close to the best score in leaderboard which is 59.52.

7. Conclusion

The main goal of the project is to predict Sustainability Index from air pollutant measurements. During domain knowledge part, information about SI calculation are defined.

During EDA part, training and test datasets comparison, correlation analysis and distribution of raw dataset are done.

With the detections in EDA, defining strategy of missing values and outliers are established, data transformation to get normal distribution is done.

In model development phase, 2 main models (Elastic Net and XGBRegressor) and 1 feature engineering tool (Sequential Feature Selection) are applied and compared with their combinations. After selecting best model which is XGBRegressor, residual analysis is done. According the analysis, most of data distributed normally and just 6 outliers are detected.

At future works, feature engineering can be done more deeply. Also, different type of handling missing value methods may be experienced, because submission test dataset's missing values are filled with means of features, before the task created.

REFERENCES

[1] Payus, C. M., a,b,c, Syazni, M. N., b, & Sentian, J., b (2022). Extended air pollution index (API) as tool of sustainable indicator in the air quality assessment: El-Nino events with climate change driven. *Heliyon*. https://doi.org/10.1016/j.heliyon.2022.e09157

[2] airly (n.d.). *Air Quality Index CAQI and AQI – Methods of Calculation*. Airly.org. https://airly.org/en/air-quality-index-caqi-and-aqi-methods-of-calculation/

[3] Qualtrics (n.d.). *Interpreting Residual Plots to Improve Your Regression*. Qualtrics. https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/