JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REGULARIZATION TECHNIQUES BASED AIR QUALITY INDEX PREDICTION

[1]Nampally Shiva Kumar, [2]Eslavath Manoj Kumar, [3]Miyapuram Bhanu Sree, [4]Dr. V Saravana Kumar

[1,2,3]IV Year Students, Dept. of IT, Sreenidhi Institute of Science and Technology(A), Hyderabad, India

[4] Associate Professor, Dept. of IT, Sreenidhi Institute of Science and Technology(A), Hyderabad, India

Abstract: The quality of air index is most important for human lives. Air quality index measure impact of air pollution impact on personal health over less period of time. Air quality of index information provide negative effects on health of human. Gradually increase pollution in Indian cities day to day. Number of mathematical concepts create several ways to determine the air quality index. Several studies found that connection between air pollution impact on health of people. Machine learning methods are play crucial roles to forecast air quality index and analysis. The air quality index prediction assists in climate control. Regularization techniques can be improved air quality index prediction for finding optimal solution. Statistical analysis shows that, using four models and calculating MSE values. Among these models Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) and Least-Angle Regression (LARS) generate same MSE values 2942.93. Elastic Net generate MSE value is 2941.

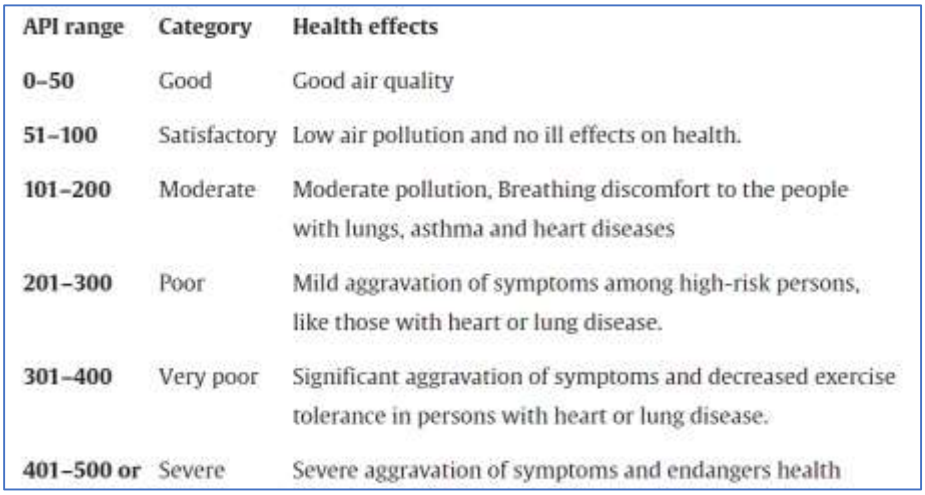
Keywords: regularization techniques, air quality index, prediction.

I. INTRODUCTION

Humans survive only because of air. So, air importance is major role in human life. In the same time air quality also most important for human health. Air quality monitoring is most significant for research [1]. Due to air pollution, so many health issues raised from physiological disorders and respiratory system fail. From scientific research, air pollution is a single crucial environmental task. Taxic gases released by industries, such type of things also increased air pollution [2]. Public health is suffering hugely as a outcome, air become hazardous substances. Each and every day check the air pollution because sometimes AQI has been declined suddenly [3]. The numerical index measures the air quality levels. Mostly 12 parameters consider for air quality measures. They are NO_2, SO_2, CO, O_3, NH_3, PM_{10}, Benzene etc [4]. But some applications they only consider six parameters for air quality checking. All these parameters consider based on situation, area, time, aim, different variables, data accessibility and monitoring frequency [5]. High air quality index indicates the serious negative impact on public health. In real time air quality monitor using the AQI techniques [6]. Different weather stations also captured air quality daily its own backup. This background data also analysed and monitored for air quality index and suggestion taken from them [7].

Machine learning techniques are used to predict the air quality index, and the optimal accuracy will be considered and also determine comparison. The following table 1 describe the API range [8].

Table 1: AQI ranges **[9]**



API range	Category	Health effects
0–50	Good	Good air quality
51–100	Satisfactory	Low air pollution and no ill effects on health.
101–200	Moderate	Moderate pollution, Breathing discomfort to the people with lungs, asthma and heart diseases
201–300	Poor	Mild aggravation of symptoms among high-risk persons, like those with heart or lung disease.
301–400	Very poor	Significant aggravation of symptoms and decreased exercise tolerance in persons with heart or lung disease.
401–500 or	Severe	Severe aggravation of symptoms and endangers health

The following paper continue with second section proposed system and architecture. Section three discuss the result and analysis. Section four comparative study of regression techniques. Final section concludes the paper.

II. PROPOSED SYSTEM AND ARCHITECTURE

Machine learning methods are play crucial roles to forecast air quality index and analysis. The air quality index prediction assists in climate control. Regularization techniques can be improved air quality index prediction for finding optimal solution. Statistical analysis shows that, using four models and calculating MSE values [10]. Among these models Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) and Least-Angle Regression (LARS) and Elastic Net.

The main focus of our research is to calculate MSE values with the help of regularization techniques of machine learning for air quality index. The following figure 1 shows the proposed architecture with different phases of Ahmadabad city AQI.

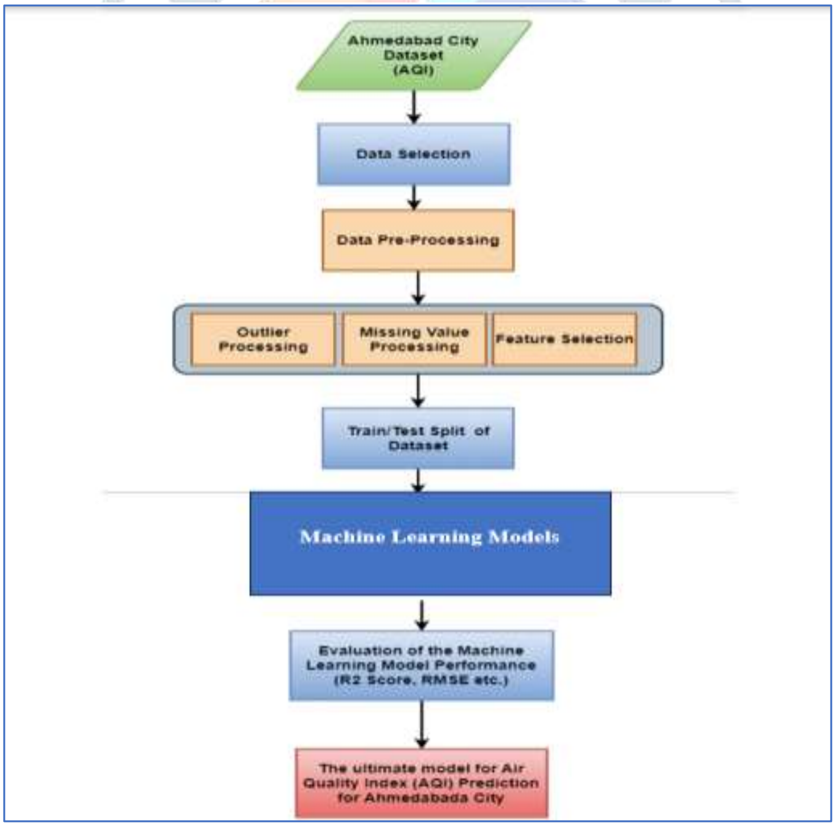

Ahmedabad City Dataset (AQI)
Data Selection
Data Pre-Processing
Outlier Processing
Missing Value Processing
Feature Selection
Train/Test Split of Dataset
Machine Learning Models
Evaluation of the Machine Learning Model Performance (R2 Score, RMSE etc.)
The ultimate model for Air Quality Index (AQI) Prediction for Ahmedabada City

Figure 1: The overall architecture for AQI Prediction [11]

The architecture shows the primary stage of dataset of AQ, city of Ahmadabad. After that data section from raw data. Data selection is completed and then preprocess the data to remove abnormal values. The preprocessing connecting with three types of data processing. They are outliers, missing values and then feature selection. Train and test split of dataset and then run the machine learning model. Evaluate the machine learning model with different error metrics. Finally display the prediction of air quality index of Ahmadabad city.

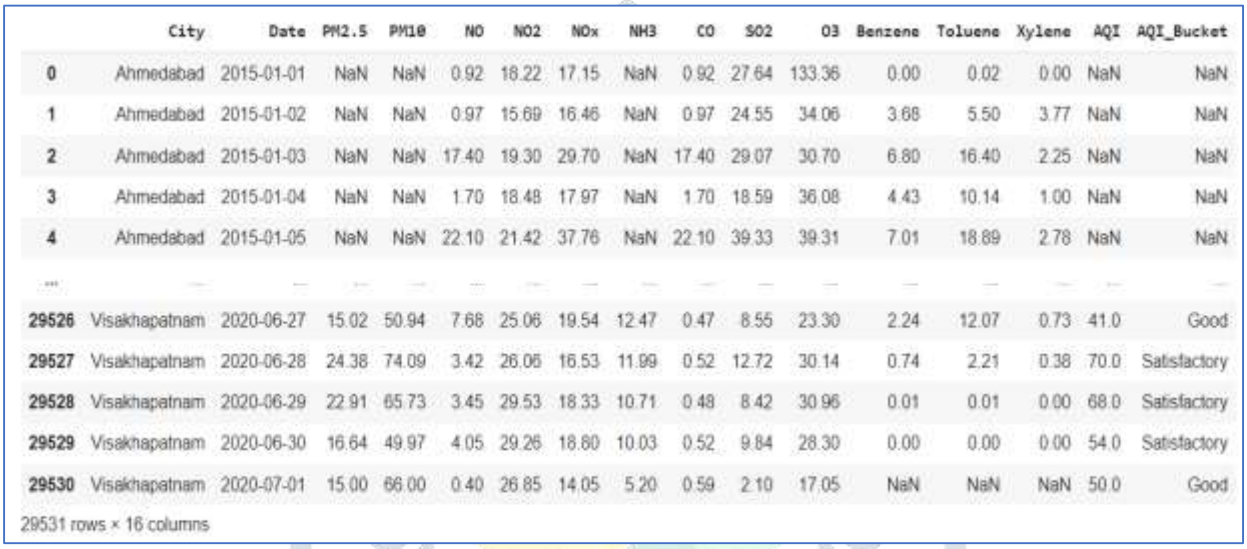
III. RESULTS AND ANALYSIS

Machine learning methods are play crucial roles to forecast air quality index and analysis. The air quality index prediction assists in climate control. Regularization techniques can be improved air quality index prediction for finding optimal solution. Statistical analysis shows that, using four models and calculating MSE values. Among these models Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) and Least-Angle Regression (LARS) and Elastic Net.

3.1 Loading Dataset

Upload widget is only available when the cell has been executed in the current browser session. Python programming environment implementation on Jupyter notebook.

Table 2: Dataset



	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29526	Visakhapatnam	2020-06-27	15.02	50.94	7.68	25.06	19.54	12.47	0.47	8.55	23.30	2.24	12.07	0.73	41.0	Good
29527	Visakhapatnam	2020-06-28	24.38	74.09	3.42	26.06	16.53	11.99	0.52	12.72	30.14	0.74	2.21	0.38	70.0	Satisfactory
29528	Visakhapatnam	2020-06-29	22.91	65.73	3.45	29.53	18.33	10.71	0.48	8.42	30.96	0.01	0.01	0.00	68.0	Satisfactory
29529	Visakhapatnam	2020-06-30	16.64	49.97	4.05	29.26	18.80	10.03	0.52	9.84	28.30	0.00	0.00	0.00	54.0	Satisfactory
29530	Visakhapatnam	2020-07-01	15.00	66.00	0.40	26.85	14.05	5.20	0.59	2.10	17.05	NaN	NaN	NaN	50.0	Good

29531 rows × 16 columns

3.2 Data Preprocessing

Data pre-processing is a crucial phase for getting optimal result because it removes abnormal, missing values from raw data. It is very useful for feature selection.

3.2.1 Missing value handling

Raw data set include with errors that means missing values and abnormal values. Using different types of statistical techniques remove missing values otherwise fill with average value of nearby cells and ignore the values.

3.2.2 Outlier processing

Outlier data means abnormal values. They also create stir in data processing. It is mandatory to remove from data getting accurate result.

3.2.3 Feature selection

Feature selection done after completing of pre processing of raw data. It also plays crucial role for finest attribute selection. Then only get optimal accuracy.

After completion of scaling data, now select the finest features from existing features. It also supports for higher accuracy for prediction.

A heatmap represents the values for a key variable of attention across two axis variables as a grid of coloured squares. The axis variables are alienated into a bar chart or histogram.

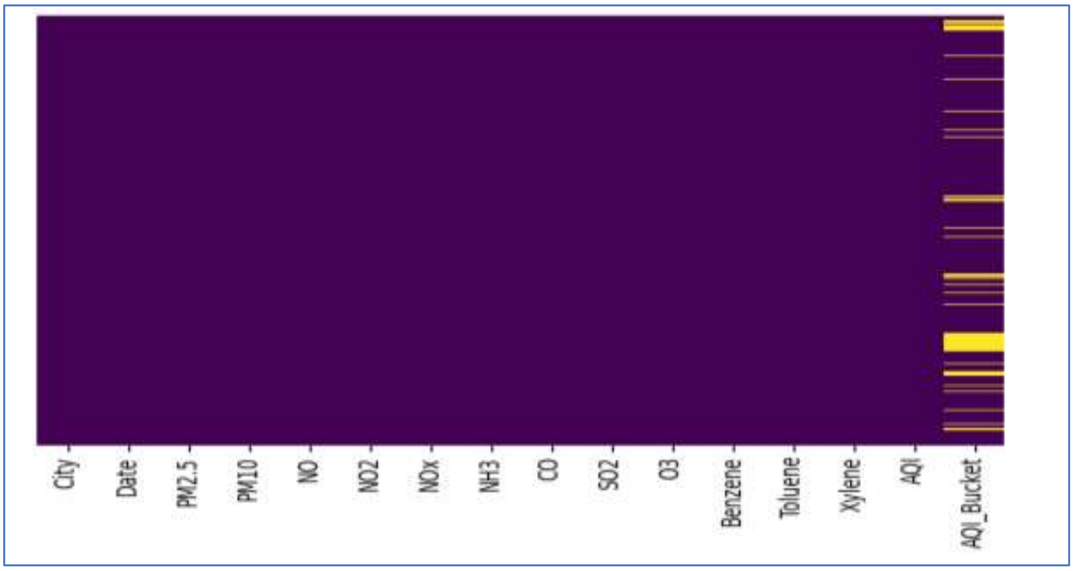


Figure 2: Heatmap

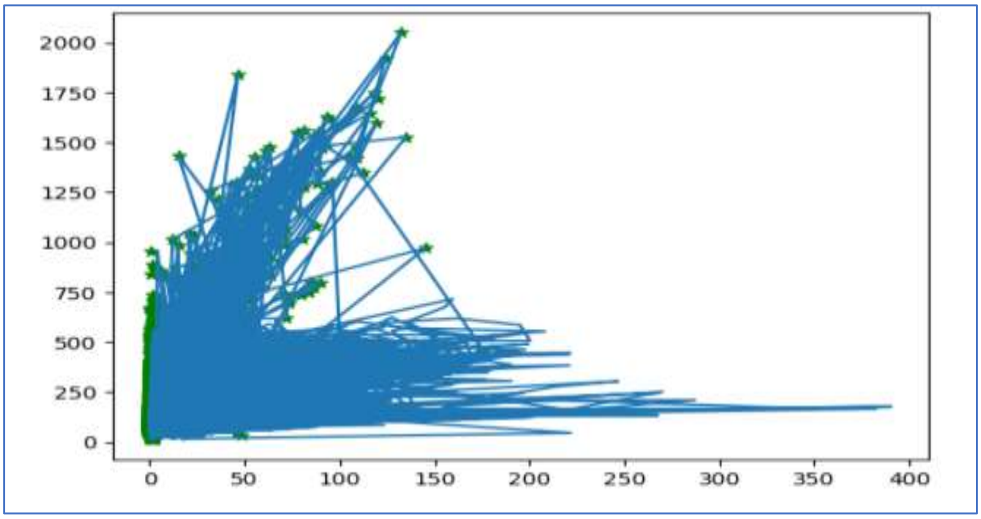
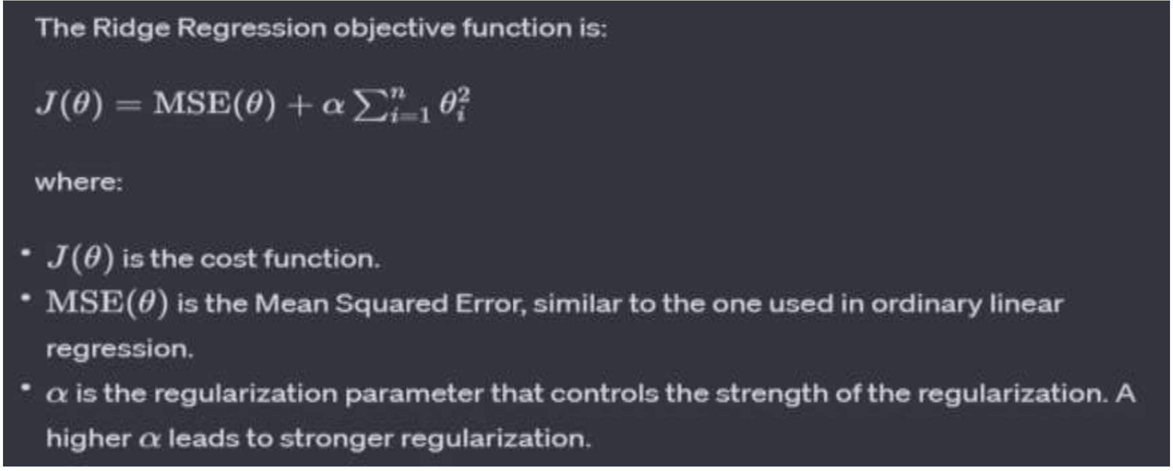


Figure 3: AQI representation map based on heatmap

Figure 3 represents the heat map-based graph with values of variables.

3. 3 Ridge Regression

Ridge Regression or L2 regularization, is a linear regression method that familiarizes regularization to the linear regression model. Regularization is a technique used to avoid overfitting by addition a penalty term to the linear regression cost function [12, 13].

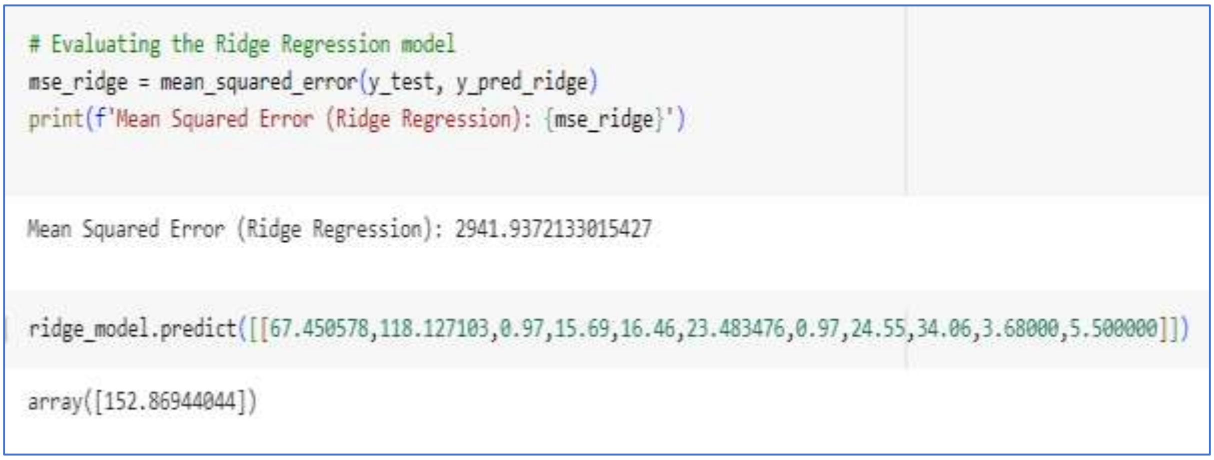


The Ridge Regression objective function is:

$$J(\theta) = \mathrm{MSE}(\theta) + \alpha \sum_{i=1}^{n} \theta_i^2$$

where:

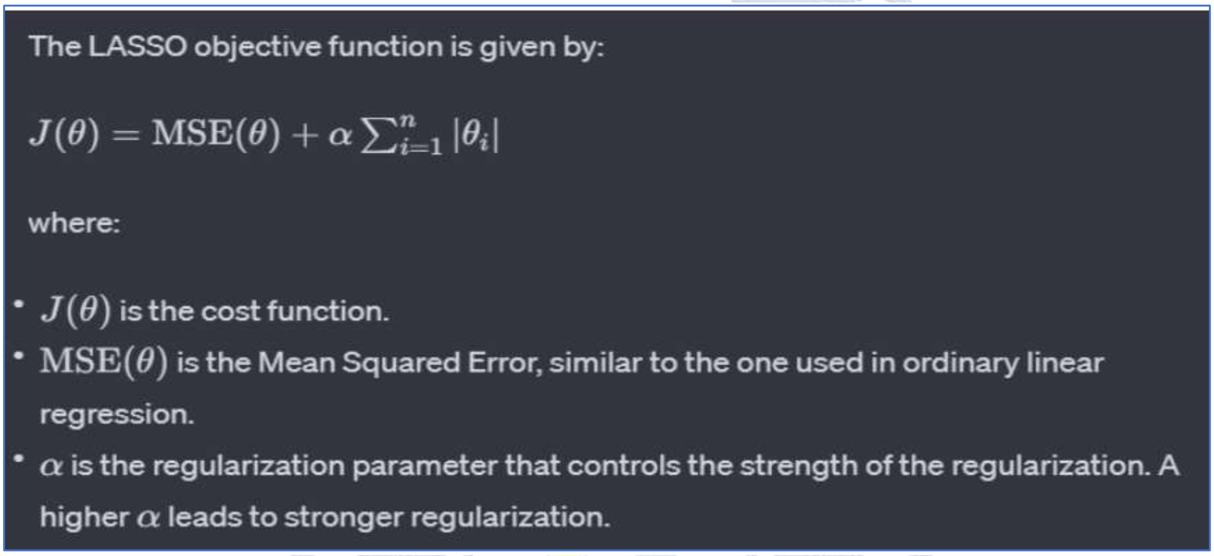
- $J(\theta)$ is the cost function.
- $\mathrm{MSE}(\theta)$ is the Mean Squared Error, similar to the one used in ordinary linear regression.
- α is the regularization parameter that controls the strength of the regularization. A higher α leads to stronger regularization.



```
# Evaluating the Ridge Regression model
mse_ridge = mean_squared_error(y_test, y_pred_ridge)
print(f'Mean Squared Error (Ridge Regression): {mse_ridge}')

Mean Squared Error (Ridge Regression): 2941.9372133015427

ridge_model.predict([[67.450578,118.127103,0.97,15.69,16.46,23.483476,0.97,24.55,34.06,3.68000,5.500000]])

array([152.86944044])
```

3.4 Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO is another form of linear regression model. LASSO regularization (L1 regularization) enhances a penalty term to the linear regression cost function based on the absolute values of the coefficients.

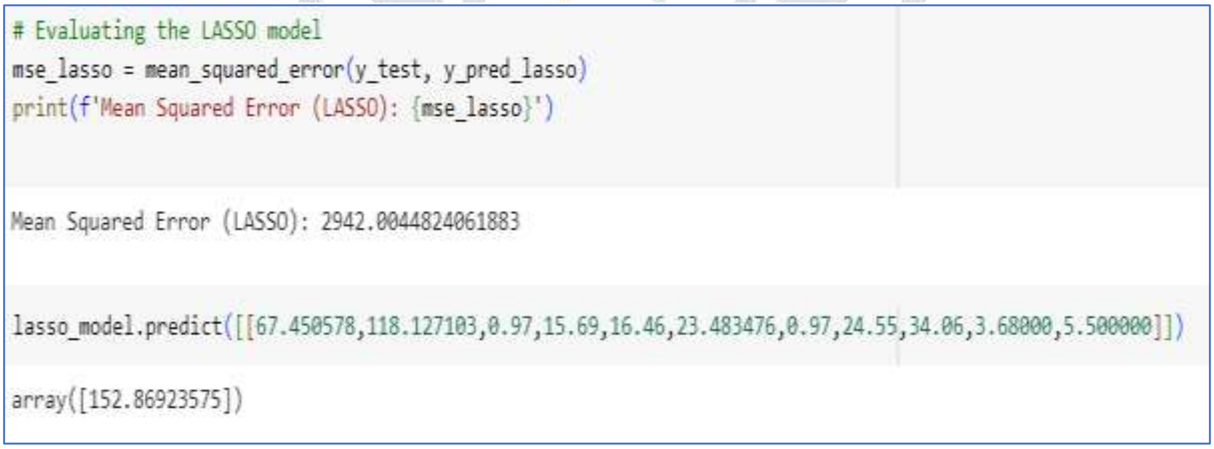


The LASSO objective function is given by:

$$J(\theta) = \mathrm{MSE}(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

where:

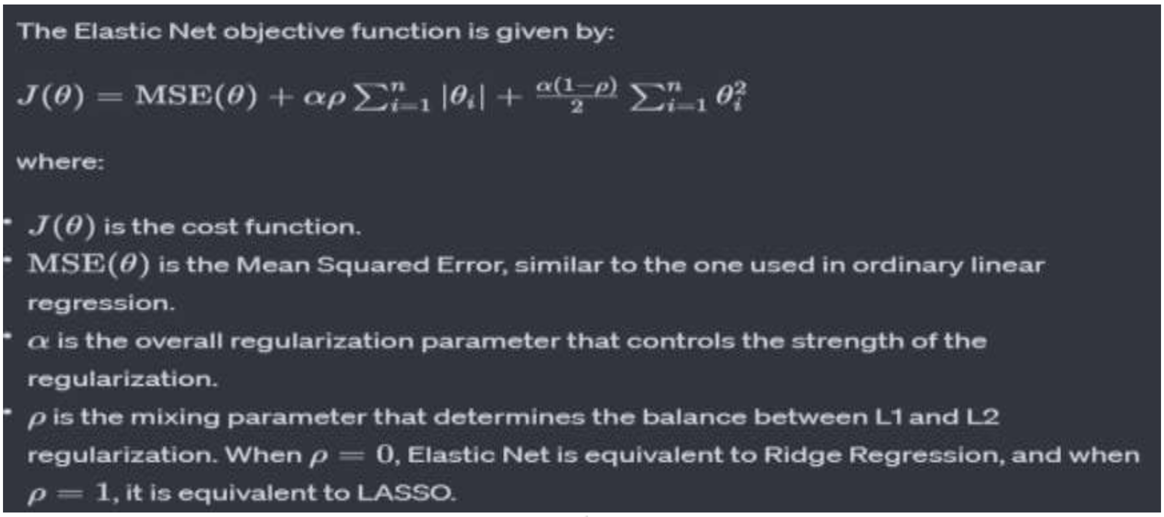
- $J(\theta)$ is the cost function.
- $\mathrm{MSE}(\theta)$ is the Mean Squared Error, similar to the one used in ordinary linear regression.
- α is the regularization parameter that controls the strength of the regularization. A higher α leads to stronger regularization.



```
# Evaluating the LASSO model
mse_lasso = mean_squared_error(y_test, y_pred_lasso)
print(f'Mean Squared Error (LASSO): {mse_lasso}')

Mean Squared Error (LASSO): 2942.0044824061883

lasso_model.predict([[67.450578,118.127103,0.97,15.69,16.46,23.483476,0.97,24.55,34.06,3.68000,5.500000]])

array([152.86923575])
```

5.5 Elastic Net

Elastic Net is a linear regression method that associations both L1 regularization (LASSO) and L2 regularization (Ridge) by adding both additional terms to the linear regression cost function. It is particularly useful when dealing with multicollinearity in the dataset, where multiple features are highly correlated.

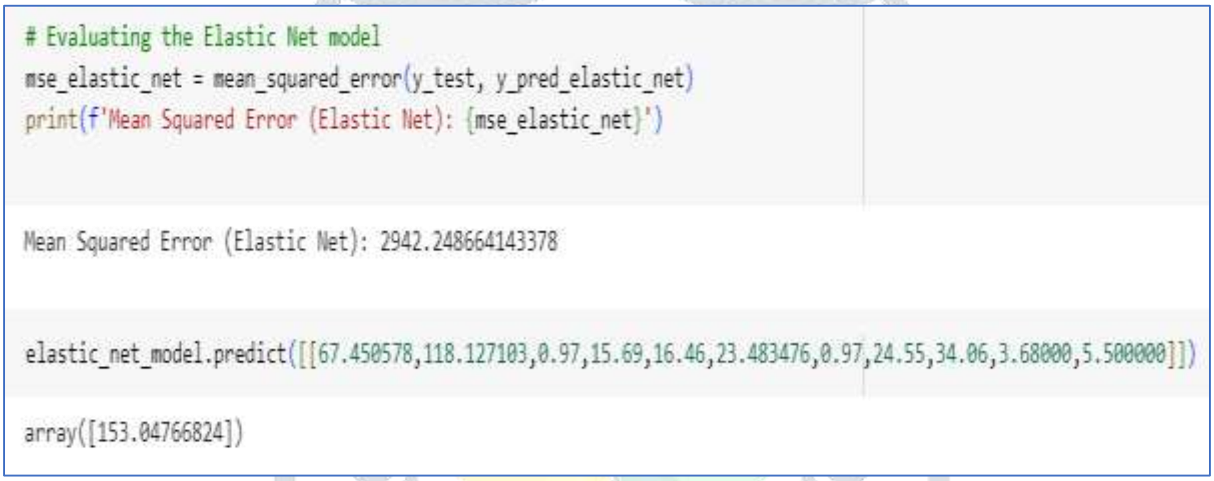


The Elastic Net objective function is given by:

$$J(\theta) = \text{MSE}(\theta) + \alpha\rho \sum_{i=1}^{n} |\theta_i| + \frac{\alpha(1-\rho)}{2} \sum_{i=1}^{n} \theta_i^2$$

where:

- $J(\theta)$ is the cost function.
- $\text{MSE}(\theta)$ is the Mean Squared Error, similar to the one used in ordinary linear regression.
- α is the overall regularization parameter that controls the strength of the regularization.
- ρ is the mixing parameter that determines the balance between L1 and L2 regularization. When $\rho = 0$, Elastic Net is equivalent to Ridge Regression, and when $\rho = 1$, it is equivalent to LASSO.



```
# Evaluating the Elastic Net model
mse_elastic_net = mean_squared_error(y_test, y_pred_elastic_net)
print(f'Mean Squared Error (Elastic Net): {mse_elastic_net}')
```

```
Mean Squared Error (Elastic Net): 2942.248664143378
```

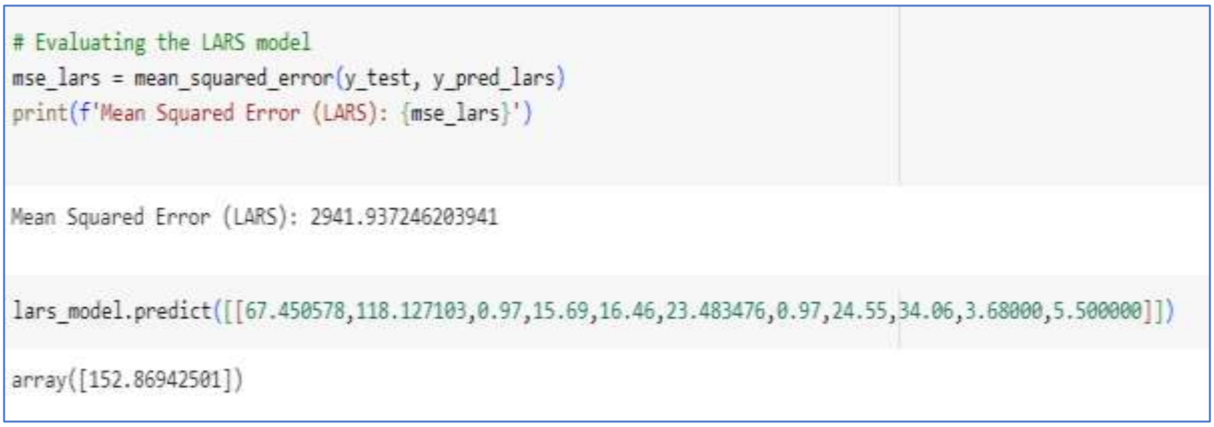
```
elastic_net_model.predict([[67.450578,118.127103,0.97,15.69,16.46,23.483476,0.97,24.55,34.06,3.68000,5.500000]])
```

```
array([153.04766824])
```

3.6 Least-Angle Regression (LARS)

Least-Angle Regression (LARS) is a regression method designed for high-dimensional data. It is a forward stepwise regression method that efficiently computes the entire solution path for the LASSO problem.

The algorithm starts with an empty set of selected features and iteratively adds the most correlated (or equi-correlated) feature to the active set. It continues until it reaches the least squares solution, effectively creating a piecewise linear path of coefficient values for each feature.



```
# Evaluating the LARS model
mse_lars = mean_squared_error(y_test, y_pred_lars)
print(f'Mean Squared Error (LARS): {mse_lars}')
```

```
Mean Squared Error (LARS): 2941.937246203941
```

```
lars_model.predict([[67.450578,118.127103,0.97,15.69,16.46,23.483476,0.97,24.55,34.06,3.68000,5.500000]])
```

```
array([152.86942501])
```

IV. COMPARATIVE STUDY

In our research concentrate on Mean Squared Error. Generally, focus on less MSE value for good accuracy. Here we have to practice four models for getting good values. If you compare four models of MSE in the following table 2.

Table 2: Comparative MSE values

S. No.	Name of the Regressor	MSE Value
1	Ridge Regression	2941.93
2	Least Absolute Shrinkage and Selection Operator (LASSO)	2942.00
3	Elastic Net	2941.24
4	Least-Angle Regression (LARS)	2941.93

The above table 2 given information about four models MSE values. Among these models Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) and Least-Angle Regression (LARS) generate same MSE values 2942.93. Elastic Net generate MSE value is 2941.

Our dataset practice on these models generates almost same MSE values. So, if you can use any model for calculation of Air quality Index. There is no difference on this dataset.

V. CONCLUSION

Air quality index measure impact of air pollution impact on personal health over less period of time. Air quality of index information provide negative effects on health of human. Gradually increase pollution in Indian cities day to day. Number of mathematical concepts create several ways to determine the air quality index. Several studies found that connection between air pollution impact on health of people. Machine learning methods are play crucial roles to forecast air quality index and analysis. The air quality index prediction assists in climate control. Regularization techniques can be improved air quality index prediction for finding optimal solution. Statistical analysis shows that, using four models and calculating MSE values. Among these models Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) and Least-Angle Regression (LARS) generate same MSE values 2942.93. Elastic Net generate MSE value is 2941.

REFERENCES

1. H. Liu, Q. Li, D. Yu, and Y. Gu, “Air quality index and air pollutant concentration prediction based on machine learning algorithms,” *Applied Sciences*, vol. 9, p. 4069, 2019.
2. M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L.Vanneschi, “A machine learning approach to predict air quality in California,” *Complexity*, vol. 2020, Article ID 8049504, 23 pages, 2020.
3. G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, “Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models,” *Journal of Engineering Research*, vol. 9, 2021.
4. S. V. Kottur and S. S. Mantha, “An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data,” *Int. J. Adv. Res.* Comput. Commun. *Eng*, vol. 4, pp. 146–152, 2015.
5. S. Halsana, “Air quality prediction model using supervised machine learning algorithms,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, pp. 190–201, 2020.
6. A. G. Soundari, J. Gnana, and A. C. Akshaya, “Indian air quality prediction and analysis using machine learning,” *International Journal of Applied Engineering Research*, vol. 14, p. 11, 2019.
7. C. R. Aditya, C. R. Deshmukh, N. D K, P. Gandhi, and V. astu, “Detection and prediction of air pollution using machine learning models,” *International Journal of Engineering Trends and Technology*, vol. 59, no. 4, pp. 204–207, 2018.
8. J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, “Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters,” Journal of Electrical and Computer Engineering, vol. 2017, Article ID 5106045, 14 pages, 2017.
9. P. Bhalgat, S. Pitale, and S. Bhoite, “Air quality prediction using machine learning algorithms,” International Journal of Computer Applications Technology and Research, vol. 8, pp. 367–370, 2019.
10. M. Bansal, “Air quality index prediction of Delhi using LSTM,” *Int. J. Emerg. Trends Technol. Comput. Sci*, vol. 8, pp. 59–68, 2019.
11. Mr. Sujan Reddy, Ms. Renu Sri and Subhani Shaik,” Sentimental Analysis using Logistic Regression”, International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.
12. P. Pranathi, V. Revathi, P. Varshitha, Subhani Shaik and Sunil Bhutada,” Logistic Regression Based Cyber Harassment Identification”, Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 76-85, June-2023.
13. Nagaraju Devarakonda, Shaik Subhani, Shaik Althaf Hussain Basha ,” Outliers Detection in Regression Analysis Using Partial Least Square Approach”, ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-13-15, Dec -2013 in Visakhapatnam, India, Vol. II, pp 125–135, 2014.