Urban Air Quality Prediction With Feature Voting

Ms.Sasi Rekha Sankar
Assistant Professor
Department Of Computational Intelligence
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
CHENNAI, INDIA

Chandra Prakash Singhi
Department Of Computational Intelligence
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
CHENNAI, INDIA
cs6955@srmist.edu.in

Anurag Nayak
Department Of Computational Intelligence
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
CHENNAI, INDIA
an2009@srmist.edu.in

Abstract

In the 21st century, increasing air pollution, urbanization, industrialization, vehicle emissions, and other factors have made breathing in Indian cities difficult. Under the National Clean Air Programme(NCAP) initiated by the Union government, 131 Indian cities were classified as 'non-attainment cities'. These cities fell short of the minimum required air quality standards set by the National Ambient Air Quality Standards(NAAQS). Thus, it is in social interest to produce an accurate prediction of the air quality of an area using machine learning. Given the dynamic nature of weather data, it is often noted that such datasets require extensive data preprocessing and effective feature selection as a pre-requisite for satisfactory training under a machine learning model. It is better to trim the dataset to contain only essential pollutants. However, standalone feature selection models can be unreliable as they can produce contradicting results on the same dataset. Hence, it is imperative to combine multiple feature selection models and devise a generalized, cost-effective approach to eliminate redundant features. This results in better model performance and increased accuracy. Feature Voting can be used to mitigate indecision during the process of Feature Selection. Further predicting the AQI values (hourly) for each city can help to consider the pollutants for each city and their individual causes. This will help policymakers to mitigate urban air pollution more effectively.

Keywords—Machine Learning; Feature Selection; Air Quality; Feature Voting; Random Forest Regression

I Introduction

The efficacy of Feature Voting approach can be demonstrated in the prediction of air quality. The air quality standards are objectively measured using Air Quality Index(AQI). This index takes into account various air pollutants present in the weather data at a particular region, at a specific time or date. It then yields a positive value based on an inherent formula, subject to the pollutants taken into account . The AQI equation has two variants – Indian AQI equation, US-EPA AQI equation. This paper has taken the Indian variant into consideration as issued by the Central Pollution Control Board under the Union Government. The AQI value(0-500) indicates the relationship between human health and air quality in an area. Based on AQI values there are 6 broad categories of safety namely – good(0-50), satisfactory(51-100), moderate(101-200), poor(201-300), very poor(301-400) and hazardous(401-500).

The issue of predicting the air quality of a region can be framed as a machine learning regression problem. By attempting to predict the AQI value of a region(in this case a city), it is possible to gain a measure of air quality in the given timeframe of that region. By analyzing the trend of AQI values of the particular city, the accuracy on the ML model can be tested(R2 score). If the accuracy is above a certain threshold(0.75 R2 score), then we can conclude that the pollutants considered to make the particular prediction are essential to the city and therefore need to mitigated first before addressing other minor pollutants.

II Research Gaps

- Based on the literature review that was done, there is a general classification of feature selection techniques.
- This necessitates a closer examination of the methodologies and a combination of multiple feature selection techniques.
- There is also a lack of discussion regarding the relationship between feature selection and the outcome of the ML operation.
- Urban air quality prediction of numerous Indian cities, pinpointing the specific causes of air pollution is also lacking, hence it is in best interest to perform this exercise.

III Abbreviations

FS - Feature Selection
RF – Random Forest
RFE – Recursive Feature Elimination
VIF – Variance Inflation Factor
FV – Feature Voting

IV Contribution

Since it is probable that multiple FS methods could yield conflicting results which would lead to indecision on the

programmer's side, it is desirable to create a Feature Voting method. The FVS method should be created using multiple FS methods / techniques which individually assess different aspects of a feature in the concerned dataset. This is necessary as repeated evaluation of a single feature using similar metrics by different FS methods is undesirable and a waste of computational resources.

FS methods are broadly divided into Supervised and Unsupervised FS methods. Supervised FS techniques are further classified into 3 categories namely(Filter, Wrapper and Embedded). The FVS should be constituted of at least 3 FS methods each falling in a different FS classification to prevent repeated evaluation of a single feature with multiple FS methods, and to achieve a clear majority for a single feature. The 3 FS methods taken into account in this research paper are – Pearson Correlation, VIF, RFE. The results of each FS method implemented is stored and presented in the form a Pandas dataframe with values of 0(meaning not voted) and 1(meaning voted). A single feature can have a total of 0 to 3 votes within the dataframe. The programmer can see which FS method voted for which feature with the help of the dataframe. Based on this the programmer can decide which features to keep and which to eliminate.

Then the votes of every city is tallied and recorded into another dataframe which has the total vote tally of every pollutant measured with respect to every city. Thus, an order of preference is formed for every city, with which deletion of unnecessary columns can be performed. After the deletion of redundant features, AQI prediction is performed for every city in consideration, and the accuracy of each machine learning prediction is recorded. If the accuracy(R2 score) is equal to or greater than 0.75, it is labelled as a successful prediction.

V DIAGRAMS

ARCHITECTURE DIAGRAM

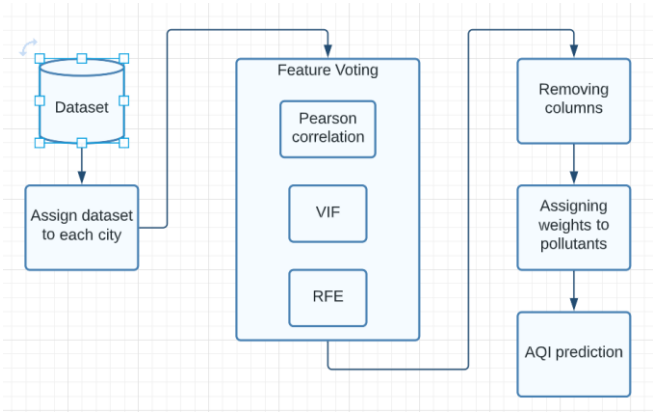

Fig.1. System architecture of the whole methodology

ACTIVITY DIAGRAM

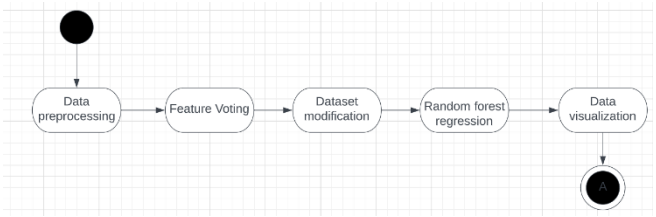

Fig.2. Activity diagram of the whole methodology

VI Methodology

First, the original dataset is divided into smaller subsets corresponding to 26 cities in the dataset. To perform FV, we use Pearson correlation, VIF and RFE. Each of the subset undergoes FV and the individual votes for every pollutant are totaled for every city in the form of a dataframe. This total vote tally of every pollutant is used to create another dataframe where city and the total vote tallies for every pollutant with respect to that city is displayed. After this, we perform deletion of columns to remove redundant pollutants for every city and then predict the AQI for every city using RF.

1) Dataset

The dataset initially has 26 cities and has 12 pollutants. After removing the last column(AQI Bucket) and handling missing data, we then divide this dataset into 26 subsets , corresponding to each city.

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	67.450578	118.127103	0.92	18.22	17.15	23.483476	0.92	27.64	133.36	0.00	0.02	0.00	166.463581	NaN
1	Ahmedabad	2015-01-02	67.450578	118.127103	0.97	15.69	16.46	23.483476	0.97	24.55	34.06	3.68	5.50	3.77	166.463581	NaN
2	Ahmedabad	2015-01-03	67.450578	118.127103	17.40	19.30	29.70	23.483476	17.40	29.07	30.70	6.80	16.40	2.25	166.463581	NaN
3	Ahmedabad	2015-01-04	67.450578	118.127103	1.70	18.48	17.97	23.483476	1.70	18.59	36.08	4.43	10.14	1.00	166.463581	NaN
4	Ahmedabad	2015-01-05	67.450578	118.127103	22.10	21.42	37.76	23.483476	22.10	39.33	39.31	7.01	18.89	2.78	166.463581	NaN

Fig.3. Initial dataset

2) Pearson Correlation

Pearson Correlation measures the extent till which 2 features are correlated to each other. This method yields a value between the range of -1 to 1, which can be converted to absolute numbers for better evaluation. After conversion to absolute value, the Pearson Correlation value is between 0 and 1. The closer it is to 1, the more correlation present between the pair of features under supervision. A value of less than or equal to 0.30 is considered bad and indicates low level of correlation between the 2 features.

3) Variance Inflation Factor

In numerous regression analysis, multicollinearity is quantified by the Variance Inflation Factor (VIF). It measures how much collinearity among predictor factors inflates the variance of the estimated regression coefficients. The variance inflation factor (VIF) measures the extent to which correlations among independent factors amplify the variance of a given regression coefficient. Strong correlation between

the independent variables, as indicated by a high VIF, can produce unstable and unreliable approximations of the regression coefficients. If the VIF is 1, then the independent variables are uncorrelated, and if it is higher than 1, then multicollinearity is possible. Values of VIF greater than 5 or 10 are typically deemed high and may necessitate additional research or corrective action to address the issue of multicollinearity.

4) Recursive Feature Elimination

It is Wrapper FS method which divides the full dataset into smaller subsets of features. The programmer provides an end machine learning model for which RFE begins scoring each subset and then ultimately deciding the best subset of features based on the individual performance of the subsets. However, it is to be noted that RFE decides based on feature importance and not simply on model performance. It provides a Numpy array of feature rankings which is used to tally votes.

5) Feature Voting

By combining the above 3 FS methods, it is prudent to present the findings in a presentable manner, for the programmer to understand and decide on which features he wants to eliminate. Granting each feature, a vote(1 for vote given, 0 for vote not given) every time a FS method selects them is a simple method to create a metric , which conveys which feature is redundant. If the FVS yields a value of less than 2, then the feature is redundant. If the FV vote tally for a feature is more than threshold then it is to be retained. The end result is a dataframe of dimension 3xN (N being the number of features in the initial dataset). When implementing the FV, it is important that the input dataset only contains relevant features (as a result of effective feature extraction) as the FV is a purely statistical exercise.

The threshold value to determine if a pollutant is to be retained or not, is specific to individual cities and is determined as follows:

Step-1: Find the maximum votes achieved by a pollutant for the city and assign the value to the variable max

Step-2: threshold = max-1

Step-3: If the pollutant's votes are more than threshold , retain the feature, else perform deletion.

6) Random Forest Regression

In the realm of machine learning, the regression job is best handled by random forest regression. It's an improvement upon the widely-used random forest method for classifying data. By sampling both the training data and the predictor factors at random, many decision trees can be built in a random forest regression model. Recursively dividing the data into subsets according to the predictor variables, and then giving a predicted value to each subset based on the average (for regression) of the response variable values in that subset, constitutes the construction of each decision tree. The average of the predictions made by each decision tree in the random forest is used to determine the predicted value for a new measurement. This method improves the model's precision while decreasing the effects of overfitting.

7) AQI

$$Ip = [IHi - ILo / BPHi - BPLo] (Cp - BPLo) + ILo$$

Where,

Ip = index of pollutant p

Cp = truncated concentration of pollutant p

BPHi = concentration breakpoint i.e. greater than or equal to Cp

BPLo = concentration breakpoint i.e. less than or equal to Cp

IHi = AQI value corresponding to BPHi

ILo = AQI value corresponding to BPLo

VII Result

After conducting FV on all 26 cities, we get the FV output for every city in a fashion as shown below:

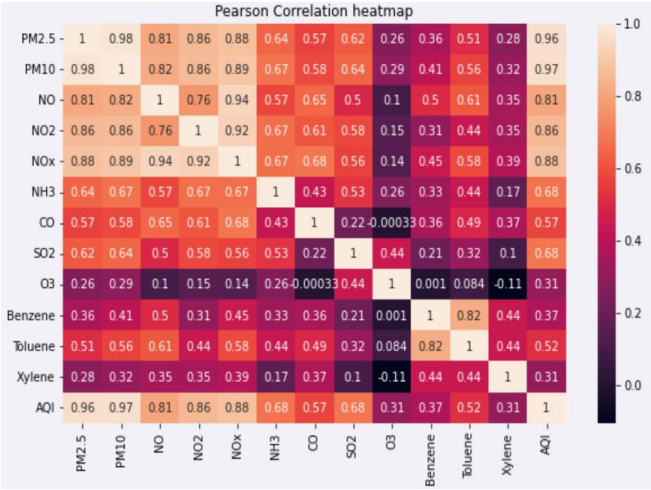

Fig.4. Pearson correlation heatmap of Kolkata

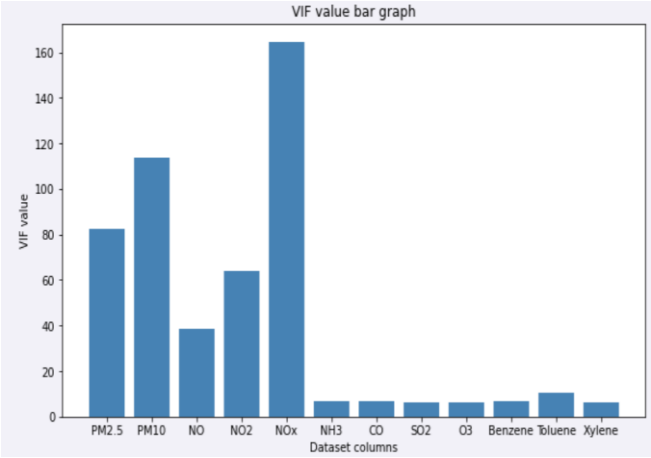

Fig.5. VIF bar graph for Kolkata

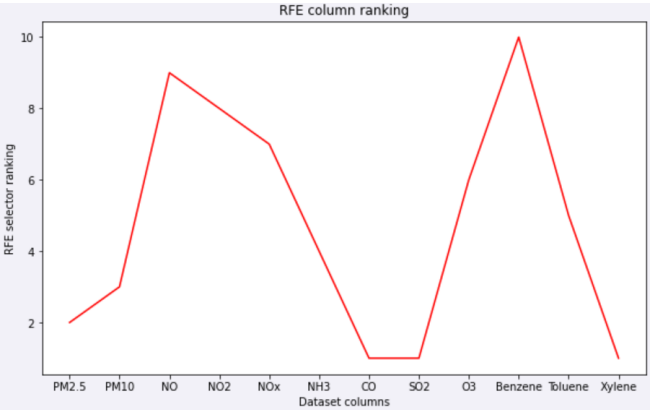

Fig.6. RFE column ranking for Kolkata

	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
Pearson	1	1	1	1	1	1	1	1	0	0	1	0
VIF	0	0	0	0	0	0	0	0	0	0	0	0
RFE	0	0	0	0	0	0	1	1	0	0	0	1

Fig.7. Votes assigned to pollutants for Kolkata

After this, we tally the total votes city wise, for every pollutant and then we create another dataframe storing the votes assigned to every pollutant for every city.

	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
Ahmedabad	2	1	2	1	1	0	2	2	1	1	1	1
Aizawl	2	2	1	2	0	1	2	1	1	0	0	0
Amaravati	2	1	0	1	2	0	2	0	0	1	1	1
Amritsar	2	2	0	0	0	1	2	1	1	1	2	1
Bengaluru	2	3	1	1	1	1	2	1	1	1	1	0
Bhopal	2	1	0	1	1	1	2	1	2	1	1	1
Brajrajnagar	3	2	1	2	1	1	2	1	1	1	1	1
Chandigarh	2	1	0	0	0	0	2	0	0	1	2	1
Chennai	3	1	1	1	1	1	2	1	2	1	1	0
Coimbatore	1	1	1	0	0	1	2	1	1	1	1	0
Delhi	1	1	1	1	1	1	2	1	0	2	1	1
Ernakulam	0	0	0	1	0	0	2	0	0	1	0	1
Gurugram	3	0	0	1	0	0	2	1	1	1	1	1
Guwahati	1	2	1	1	1	3	3	2	1	1	1	1
Hyderabad	3	1	0	0	0	0	2	1	0	1	0	1
Jaipur	3	3	1	0	0	1	2	1	1	1	1	0
Jorapokhar	2	2	1	2	1	1	2	1	1	1	1	1
Kochi	2	2	0	0	1	1	2	1	1	1	2	1
Kolkata	1	1	1	1	1	1	2	2	0	0	1	1
Lucknow	3	1	2	3	1	1	2	1	1	1	1	1
Mumbai	2	1	0	1	1	2	2	0	0	1	0	0
Patna	3	0	0	1	0	1	1	0	0	1	0	1
Shillong	0	0	2	1	0	2	1	1	1	0	0	0
Talcher	2	3	1	1	1	1	2	1	1	2	1	0
Thiruvananthapuram	1	1	2	1	2	1	2	1	1	1	1	1
Visakhapatnam	2	1	1	0	0	0	0	1	0	1	0	0

Fig.8. Dataframe of total votes for every pollutant, for every city(26 cities in total)

The above dataframe will serve us in deletion of columns and to assign weightage to the retained pollutants. This ultimately will help to construct the pie-charts of every city. We then delete the redundant columns according to threshold value(as discussed in the methodology).

```
In Ahmedabad we delete : ['NH3']
In Aizawl we delete : ['NOx', 'Benzene', 'Toluene', 'Xylene']
In Amaravati we delete : ['NO', 'NH3', 'SO2', 'O3']
In Amritsar we delete : ['NO', 'NO2', 'NOx']
In Bengaluru we delete : ['NO', 'NO2', 'NOx', 'NH3', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Bhopal we delete : ['NO']
In Brajrajnagar we delete : ['NO', 'NOx', 'NH3', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Chandigarh we delete : ['NO', 'NO2', 'NOx', 'NH3', 'SO2', 'O3']
In Chennai we delete : ['PM10', 'NO', 'NO2', 'NOx', 'NH3', 'SO2', 'Benzene', 'Toluene', 'Xylene']
In Coimbatore we delete : ['NO2', 'NOx', 'Xylene']
In Delhi we delete : ['O3']
In Ernakulam we delete : ['PM2.5', 'PM10', 'NO', 'NOx', 'NH3', 'SO2', 'O3', 'Toluene']
In Gurugram we delete : ['PM10', 'NO', 'NO2', 'NOx', 'NH3', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Guwahati we delete : ['PM2.5', 'NO', 'NO2', 'NOx', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Hyderabad we delete : ['PM10', 'NO', 'NO2', 'NOx', 'NH3', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Jaipur we delete : ['NO', 'NO2', 'NOx', 'NH3', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Jorapokhar we delete : []
In Kochi we delete : ['NO', 'NO2']
In Kolkata we delete : ['O3', 'Benzene']
In Lucknow we delete : ['PM10', 'NOx', 'NH3', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Mumbai we delete : ['NO', 'SO2', 'O3', 'Toluene', 'Xylene']
In Patna we delete : ['PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene']
In Shillong we delete : ['PM2.5', 'PM10', 'NOx', 'Benzene', 'Toluene', 'Xylene']
In Talcher we delete : ['NO', 'NO2', 'NOx', 'NH3', 'SO2', 'O3', 'Toluene', 'Xylene']
In Thiruvananthapuram we delete : []
In Visakhapatnam we delete : ['NO2', 'NOx', 'NH3', 'CO', 'O3', 'Toluene', 'Xylene']
```

Fig.9. Deletion of columns

Then we perform RF regression for every city and record the R2 score, along with a plot to visualize the prediction and a pie chat to show the extent of contribution of every retained pollutant in the AQI prediction. This is done for all 26 cities. All the individual prediction accuracies are then used to construct a bar graph(accuracy against individual cities).Below is the prediction outcome for Mumbai.

```
Initiating prediction for city : Kolkata
R2 Score of Kolkata prediction : 0.9945093069807532
```

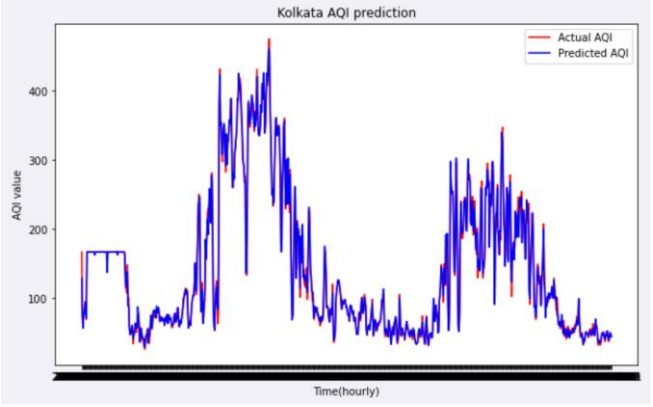

Fig.9 AQI prediction of Kolkata

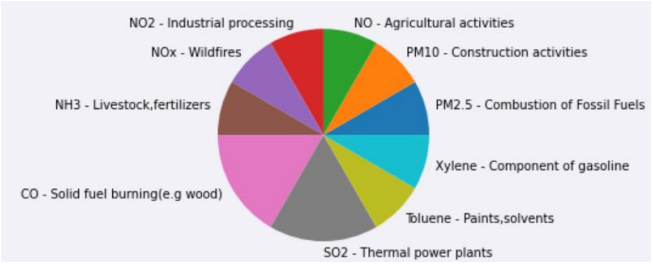

Fig.10 Pie chart of the contribution of every significant pollutant of Kolkata in its AQI prediction.

VIII Conclusion

By performing FV on each city dataset, we successfully predict the AQI values of every city. We have pinpointed the causes of each pollutant, indirectly providing solutions to mitigate the emission of the concerned pollutants for every city. This can help improve urban air quality if the results of this exercise are translated into policy by the concerned

authorities to help India to achieve net zero carbon emissions by 2070.

The project's approach to identify and relatively measure the individual contribution of different pollutants for 26 Indian cities can be applied on other cities and additional pollutants can be taken into consideration. This template can serve well for future attempts to mitigate urban air pollution via policy initiatives or legislation by urban local bodies.

References

[1] **Towards an Unsupervised Feature Selection Method for Effective Dynamic Features**

Naif Almusallam;Zahir Tari;Jeffrey Chan;Adil Fahad;Abdulatif Alabdulatif;Mohammed Al-Naeem

[2] **A Novel Framework of Two Successive Feature Selection Levels Using Weight-Based Procedure for Voice-Loss Detection in Parkinson's Disease**

Amira S. Ashour; Majid Kamal A. Nour; Kemal Polat; Yanhui Guo; Wafaa Alsaggaf; Amira El-Attar

[3] **Novel and efficient randomized algorithms for feature selection**

Zigeng Wang; Xia Xiao; Sanguthevar Rajasekaran Big Data Mining and Analytics, Year: 2020

[4] **Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection**

Anwar Ul Haq; Defu Zhang; He Peng; Sami Ur Rahman |

[5] **Bearing Fault Feature Selection Method Based on Weighted Multidimensional Feature Fusion**

Yazhou Li; Wei Dai; Weifang Zhang

[6] **Feature Redundancy Based on Interaction Information for Multi-Label Feature Selection**

Wanfu Gao; Juncheng Hu; Yonghao Li; Ping Zhang I

[7] **Prediction Analysis using Random Forest Algorithms to Forecast the Air Pollution Level in a Particular Location**

Puli Dilliswar Reddy; L. Rama Parvathy

[8] **Predictive Maintenance of Relative Humidity Using Random Forest Method**

Aji Teguh Prihatno; Himawan Nurcahyanto; Yeong Min Jang

[9] **Research and Application of Intelligent Community Based on Naive Bayes**

Lisha Yao; Kang Su

[10] **The Abstract of Thesis Classifier by Using Naïve Bayes Method**

Hairani Hairani;Anthony Anggrawan;Ahmad Islahul Wathan;Kumiadin Abd Latif;Khairan Marzuki;Muhammad Zulfikri

[11] **Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers Under Imbalanced Data Sets**

Christos K. Aridas; Stamatis Karlos; Vasileios G. Kanas; Nikos Fazakis; Sotiris B. Kotsiantis

[12] **Dynamic Feature Selection for Clustering High Dimensional Data Streams**

Conor Fahy ; Shengxiang Yang

[13] **Bayesian Penalized Method for Streaming Feature Selection**

Xiao-Ting Wang;Xin-Ze Luan

[14] **Data-Driven Air Quality Characterization for Urban Environments: A Case Study**

Yuchao Zhou;Suparna De;Gideon Ewa;Charith Perera;Klaus Moessner

[15] **Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations**

Ping-Wei Soh;Jia-Wei Chang;Jen-Wei Huang

[16] **Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities**

Saba Ameer;Munam Ali Shah;Abid Khan;Houbing Song;Carsten Maple;Saif Ul Islam;Muhammad Nabeel Asghar

[17] **Urban air quality based on Bayesian network**

Ruijun Yang;Feng Yan;Nan Zhao