Kurtosis-Based Feature Selection Method using Symmetric Uncertainty to Predict the Air Quality Index

Usharani Bhimavarapu, M. Sreedevi

Abstract

Feature selection is vital in data pre-processing in machine learning, and it is prominent in datasets with many features. Feature selection analyses the relevant, irrelevant, and redundant features in the dataset. Feature selection removes the irrelevant features, which improves both the accuracy and prediction performance. The significant advantages of reducing the number of features from the dataset are reducing the training time, reducing overfitting, decreasing the curse of dimensionality, and simplifying the prediction model. The filter feature selection techniques can handle the issues with the high number of features, and this paper uses the symmetric uncertainty coefficient to verify the relevance of the independent features. In this paper, a new feature selection method named as kurtosis-based feature selection has been proposed to select the relevant features which affect the air pollution. Kurtosis-based feature selection is compared with seven filter feature selection techniques on air pollution dataset and validated the performance of the proposed algorithm. It has been observed that the kurtosis-based feature selection extracts only PM2.5 as the key feature and has been compared to the accuracy of the five existing methods. The experimental results illustrate that the kurtosis-based feature selection algorithm reduces the original feature set up to 91.66%, but the existing filter feature selection techniques reduce the feature set to only 50%.

Keywords: Air Pollution, Air quality index, Correlation coefficient, Feature selection, Filter techniques, Symmetric uncertainty.

doi:10.56415/csjm.v30.19

1 Introduction

The features in the dataset may be repeated and noisy, and these repeated reduce the learning model's performance. This paper explains how to choose the relevant features to predict the air quality index and ignore the irrelevant features. The main advantages of feature selection (FS) are improving the prediction performance by removing the irrelevant and redundant features and reducing the computational cost [1].

There are two types of FS techniques: classifier independent (filter) and classifier dependent (wrapper, embedded) [2]. Filter FS techniques give the grade for each feature and select the top k features from the learning model. Some examples of the filter FS techniques are the symmetric uncertainty [3], Relief, Fisher, Mutual Information [4], recursive feature elimination, minimum redundancy maximum relevance, Distributed FS [5]. The classifier-dependent techniques are the time taking approaches as it needs a few learning algorithms to select the best features, which reduces the accuracy and the performance [6].

We have assessed the performance of the filter FS techniques using the machine learning techniques like linear regression, decision tree, random forest, XGBoost, lasso regression, and clustering techniques. Clustering techniques help to maintain the stability of a filter FS technique to perform similarity while selecting the subset of features with the same cardinality. In this paper, we selected the best features by measuring the stability of a method as its kurtosis of effectiveness across the features in the dataset, and we compare the stability and the performance of the seven filter FS methods. In this paper, we are using the target feature to remove the irrelevant features.

The summary of the proposed work is:

- We performed filter FS techniques using the symmetric uncertainty and found the minimal subset of features.
- We applied the threshold and constructed the correlation coefficient matrix to measure the dependency between the features.
- We have selected the best features by congregating the features into clusters.
- We reduced the features and hence, the overfitting that improves the prediction accuracy.

The structure of the rest of the paper is as follows. Section 2 reviews the antecedents of the FS techniques in various fields and discusses some existing FS techniques. Section 3 discusses the novel FS algorithm that helps find the minimal subset of features to predict the air quality index. Section 4 describes the experimental results, analyses the proposed algorithm's behavior, and presents the obtained results. Section 5 concludes the paper.

2 Related work

FSs affect the model construction, and the predefined criterion evaluates the optimization of the feature subset. FS approaches consist of selection, evaluation stopping criterion, and validation. The filter FS measures the relevance between the features, different metrics used for this are correlation-based FS algorithms [7], distance-based FS algorithms [8], statistics-based FS algorithm [9], information theory-based FS algorithms [10]. The FS algorithms are divided into linear [11] and non-linear FS algorithms [12]. Fran et al. [13] proposed the conditional mutual information FS technique that is weakly dependent. Bennasar et al. [14] proposed the joint mutual information, which extends the conditional mutual information, and used the maximum, minimum criteria. Zeng et al. [15] proposed interaction weight FS, which dynamically influences the mutual information of the features and the class labels. Hu et al. [16] proposed the dynamic relevance and joint mutual information to remove the redundant features. Kolli et al. [17] proposed a granular feature multi-variant clustering-based genetic algorithm for feature subset selection. This technique uses the granularity of neighborhood-based rough sets and the fitness values as the threshold to subset features. Sai Prasad et al. [18] proposed a novel left-to-right and right-to-left framework to reduce the features and generate a finite number of unique features.

3 Methodology

In this section, we proposed a novel filter FS algorithm, which combines the symmetric uncertainty and the kurtosis to select the features and obtain the low redundant features. Figure 1 shows the proposed technique diagram.

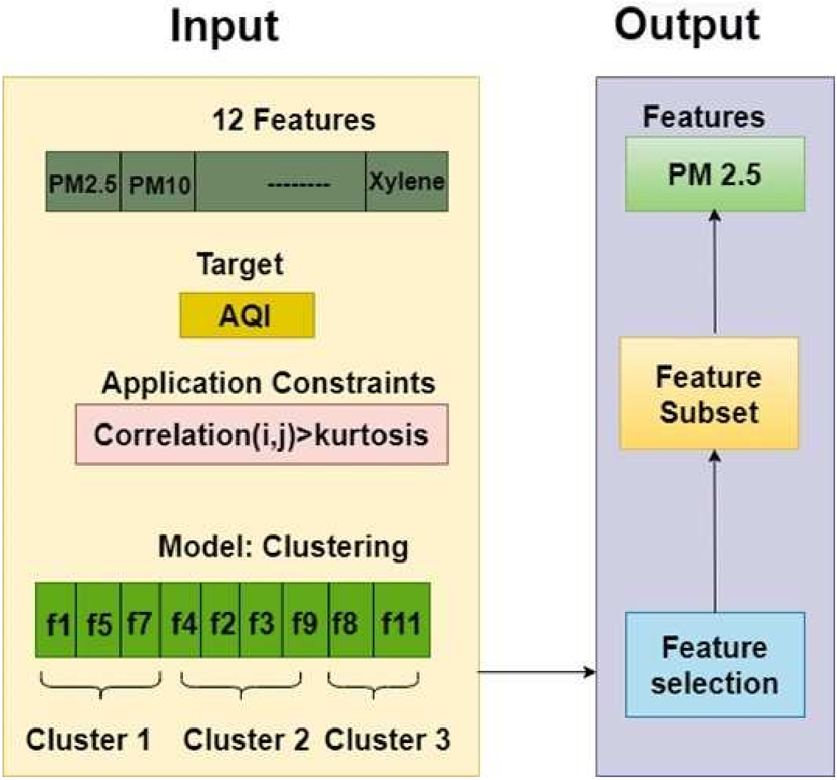


Figure 1. Proposed Block Diagram

In algorithm 1, first, we calculate the symmetric uncertainty values in step-1. In step-2, we calculated the kurtosis of all these symmetric uncertainty values and used this kurtosis value as the threshold. In step-3, we generated the correlation coefficient matrix of the original dataset, and in step-4, we constructed the binary matrix by applying the threshold in the correlation matrix. After generating the binary matrix, step-5 counted the values equal to 1 and stored these values in the $t + 1$ column in the binary matrix. Step-6 generates the clusters based on the binary $t + 1$ columns values, i.e., cluster all the similar count areas in the single cluster. In step-7, we remove the redundant features and maintain the relevant ones as the final subset.

Algorithm 1.

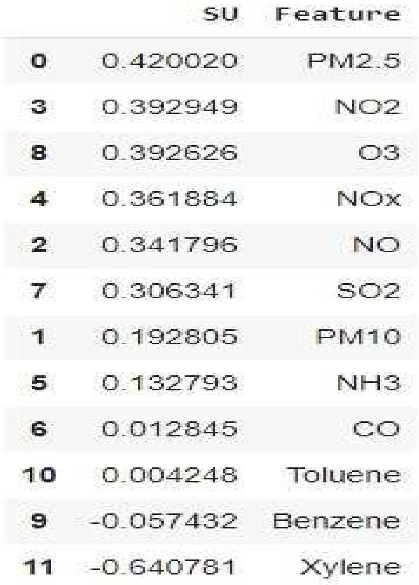
```
Input: Data set D, feature set F= $f_1, ...f_k$
Output: Selected feature set S

  1. Calculate the symmetric uncertainty of each feature and arrange all the features in descending order based on the symmetric uncertainty values.
  2. Choose the kurtosis value as the threshold.
  3. Find the correlation coefficient symmetric matrix for the original dataset.
  4. Apply the threshold value of the generated correlation coefficient matrix.
     (a) If the individual value of the correlation coefficient matrix is greater than the threshold, place the value equal to 1; otherwise, the value equal to 0.
     (b) Repeat the procedure for all the features in the correlation matrix.
  5. Calculate the total number of ones in each row.
  6. Combine all the features and form the clusters that have the same weights.
  7. Choose the highest symmetric uncertainty feature in each cluster.
```

4 Results

This paper uses the data collected from 270 monitoring stations from the Indian government website CPCB (Central Pollution Control Board); these stations automatically collect hourly air quality 24 hours per day. The data are open to the public. We collected significant air pollutants, i.e., PM2.5, PM10, CO, NO2, SO2, O3 data, from January 1, 2015, to September 1, 2019 [26]. We used Keras deep learning application programming interface with TensorFlow back end, and we implemented an improved algorithm using the IDE Anaconda.

We applied the proposed kurtosis-based FS(KBFS) algorithm to the collected dataset and first calculated the symmetric uncertainty on the collected air pollution dataset. We considered 12 features (PM 2.5, PM10, CO, NO, NO2, NOx, SO2, O3, NH3, Benzene, Toluene, Xylene) as the input and AQI as the target feature. Figure 2 shows the symmetric uncertainty values of the input features concerning the target feature.



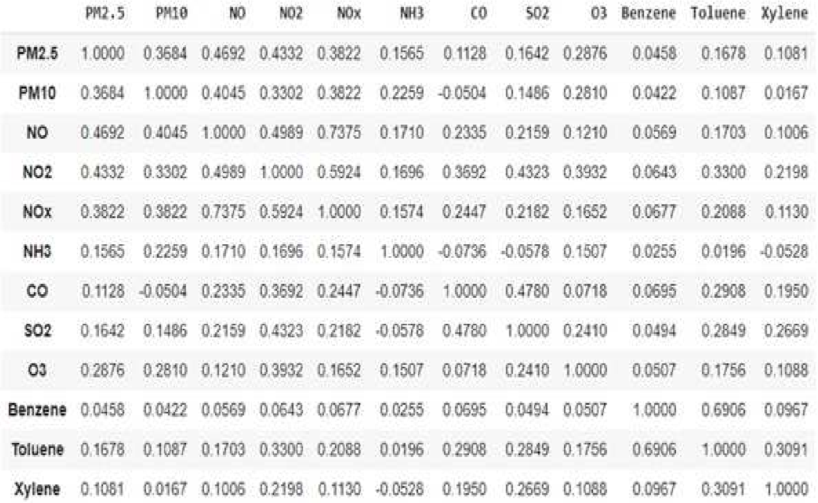
	SU	Feature
0	0.420020	PM2.5
3	0.392949	NO2
8	0.392626	O3
4	0.361884	NOx
2	0.341796	NO
7	0.306341	SO2
1	0.192805	PM10
5	0.132793	NH3
6	0.012845	CO
10	0.004248	Toluene
9	-0.057432	Benzene
11	-0.640781	Xylene

Figure 2. Sorted symmetric uncertainty

After calculating the symmetric uncertainty, we generated the correlation coefficient matrix on the original dataset. Figure 3 shows the correlation coefficient matrix for the air pollution dataset.

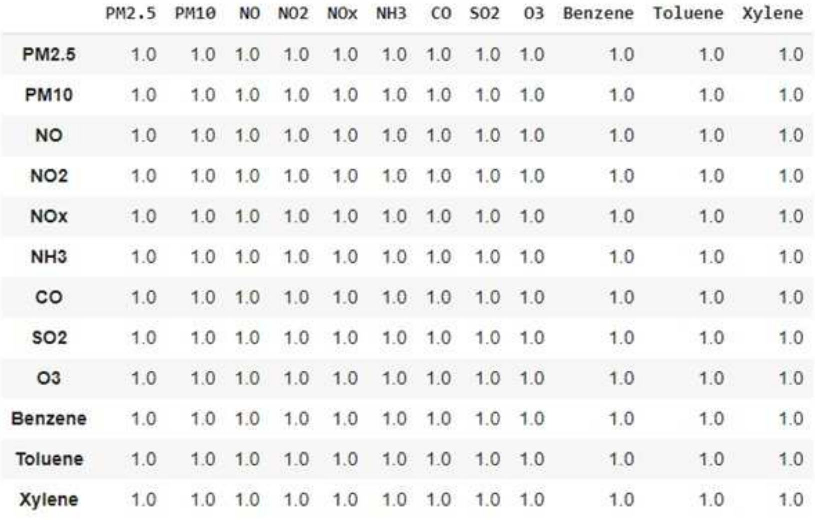
After generating the correlation coefficient matrix, calculate the kurtosis for the symmetric uncertainty values according to Figure 2. Set this value as the threshold value. Apply this threshold value to the correlation coefficient matrix. If the coefficient matrix value is greater than the threshold, then set value one; otherwise, set value zero. Figure 4 shows the generated binary matrix.

Count the number of ones in the binary matrix and record those



	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
PM2.5	1.0000	0.3684	0.4692	0.4332	0.3822	0.1565	0.1128	0.1642	0.2876	0.0458	0.1678	0.1081
PM10	0.3684	1.0000	0.4045	0.3302	0.3822	0.2259	-0.0504	0.1486	0.2810	0.0422	0.1087	0.0167
NO	0.4692	0.4045	1.0000	0.4989	0.7375	0.1710	0.2335	0.2159	0.1210	0.0569	0.1703	0.1006
NO2	0.4332	0.3302	0.4989	1.0000	0.5924	0.1696	0.3692	0.4323	0.3932	0.0643	0.3300	0.2198
NOx	0.3822	0.3822	0.7375	0.5924	1.0000	0.1574	0.2447	0.2182	0.1652	0.0677	0.2088	0.1130
NH3	0.1565	0.2259	0.1710	0.1696	0.1574	1.0000	-0.0736	-0.0578	0.1507	0.0255	0.0196	-0.0528
CO	0.1128	-0.0504	0.2335	0.3692	0.2447	-0.0736	1.0000	0.4780	0.0718	0.0695	0.2908	0.1950
SO2	0.1642	0.1486	0.2159	0.4323	0.2182	-0.0578	0.4780	1.0000	0.2410	0.0494	0.2849	0.2669
O3	0.2876	0.2810	0.1210	0.3932	0.1652	0.1507	0.0718	0.2410	1.0000	0.0507	0.1756	0.1088
Benzene	0.0458	0.0422	0.0569	0.0643	0.0677	0.0255	0.0695	0.0494	0.0507	1.0000	0.6906	0.0967
Toluene	0.1678	0.1087	0.1703	0.3300	0.2088	0.0196	0.2908	0.2849	0.1756	0.6906	1.0000	0.3091
Xylene	0.1081	0.0167	0.1006	0.2198	0.1130	-0.0528	0.1950	0.2669	0.1088	0.0967	0.3091	1.0000

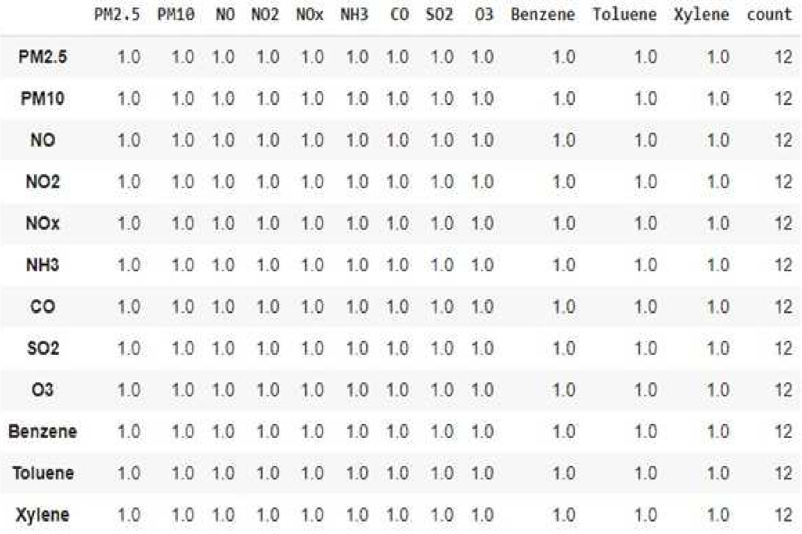
Figure 3. Correlation matrix



	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
PM2.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
PM10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
NO	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
NO2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
NOx	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
NH3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
CO	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SO2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
O3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Benzene	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Toluene	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Xylene	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Figure 4. Binary matrix

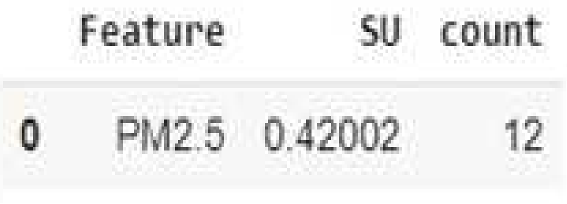
counts in the t+1 column in the binary matrix. Figure 5 shows the count for the binary matrix for each input feature.



	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	count
PM2.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
PM10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
NO	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
NO2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
NOx	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
NH3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
CO	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
SO2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
O3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
Benzene	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
Toluene	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12
Xylene	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	12

Figure 5. Count in the binary matrix using the kurtosis as the threshold

Generate the clusters that have similar counts and select the relevant feature from each cluster. We sorted all the features in the cluster as per its symmetric uncertainty, and then we selected the best symmetric uncertainty as to the relevant feature. Figure 6 shows the final subset of features.



	Feature	SU	count
0	PM2.5	0.42002	12

Figure 6. Subset of features

Table 1 tabulates different statistical tests used to find the threshold we applied to the correlation matrix. These experimental results showed that the final subset of features is only one for kurtosis, so it is the best statistical test measure to set as the threshold.

Table 1. Different statistical tests used for threshold in symmetric uncertainty

Statistical Techniques	**Total Features**	**Selected Features**
Mean	12	8
Median	12	6
Mode	12	6
Variance	12	6
Standard deviation	12	6
kurtosis	12	1

Table 2 tabulates the comparison of the different filter FS techniques. From the results, we observed that selected features using the symmetric uncertainty are smaller than the remaining filter FS techniques.

Table 2. Comparison of different filter FS techniques

FS Techniques	**Total Features**	**Selected Features**
Anova	12	6
Correlation Feature selector	12	5
mRmR	12	6
Fisher	12	5
Mutual Information	12	2
Information Gain	12	4
Gain ratio	12	3
RelieF	12	3
Variance	12	5
Symmetric Uncertainty	12	1

We reduced the number of features using the filter FS methods, keeping the relevant features that help predict the accurate air quality index. In comparison, the proposed technique selects the most relevant features, and the remaining FS techniques keep an average of 42% of the original features, but the kurtosis-based FS technique reduces the features up to 83.33%, whereas the remaining filter feature selection reduction rate ranges from 40% to 50%. Figure 7 shows the comparison of reduced rates of features using various FS techniques.

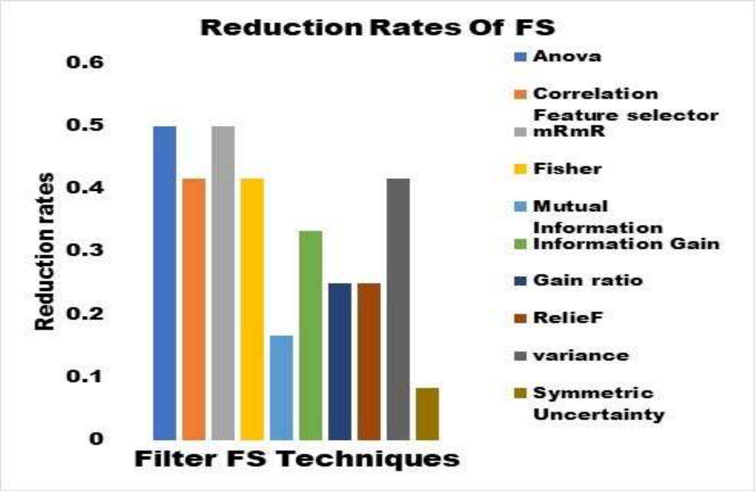


Figure 7. Comparison of reduction rates of features using various FS techniques

We analyzed the air quality dataset using machine learning techniques. We also analyzed the complete air pollution dataset, and the features were selected using the kurtosis-based FS by evaluating the performance metrics like correlation coefficient, root-mean-square error, and accuracy. We implemented the proposed kurtosis-based FS algorithm on different machine learning algorithms and tabulated it in Table 3.

We discussed the performance comparison of the machine learning classifiers after performing the FS. Table 4 compares the different FS techniques using the different classifiers. This paper uses the classifiers Random Forest(RF), Linear Regression(LR), and Principal Component

Table 3. Comparison of different machine learning algorithms with proposed technique

Classifiers	Without FS				With FS			
	r	R2	RMSE	ACC	r	R2	RMSE	ACC
Decision Tree	0.847	0.717	37.07	76.95	0.894	0.826	29.53	91.76
Linear Regression	0.827	0.736	35.45	74.53	0.874	0.825	31.42	89.34
Lasso regression	0.814	0.717	37.83	78.58	0.872	0.838	29.32	89.53
Random Forest	0.864	0.767	36.34	78.56	0.914	0.848	21.57	93.78
XGBoost	0.826	0.762	33.56	72.45	0.893	0.852	27.45	89.74
Support Vector machine	0.823	0.736	33.53	72.54	0.864	0.857	25.53	88.86

Analysis(PCA).

From the results, we observed that processing time is less for the proposed technique. We observed that accuracy significantly improved by applying the proposed technique and accomplished better performances for the symmetric uncertainty for the random forest classifier, and the processing time is 09ms.

We assessed the correlation between the major air pollutants PM2.5, CO, NO2, O3, SO2, and the air quality index and finally explored their relationship. Figure 8 shows the correlation between the major air pollutants and the air quality index.

This research finds the minimal subset of features that helps predict the accurate air quality index. From the observations, we found that PM2.5 is a relevant feature to predict the air quality index.

5 Conclusion

The objective of the FS is to select a minimal subset of relevant features. We proposed a kurtosis-based FS algorithm to reduce the dimensionality of the air pollution data by selecting the best features, which enhances the prediction performance. This paper discusses different filter FS techniques and various statistical tests to finalize the threshold to find the best fit subset of features. The authors performed the comparison for different filter FS techniques, the results showed that the proposed kurtosis-based FS algorithm improves the prediction performance.

Table 4. Comparison of different FS techniques using different classifiers

Classifier	Measure	r	R2	RSME	ACC	PT(ms)
RF	Chi-square [19]	0.825	0.701	36.54	85.78	13
	Relief [20]	0.802	0.703	37.73	84.62	14
	MultiSurf [21]	0.826	0.708	36.86	85.34	13
	Ensemble FS [22]	0.904	0.748	31.57	83.78	11
	Anova [23]	0.809	0.717	38.45	84.03	14
	M-Cluster FS [24]	0.895	0.826	22.45	91.34	10
	SU-MLP [25]	0.897	0.829	22.13	92.31	10
	Proposed	0.914	0.848	21.57	93.78	09
LR	Chi-square [19]	0.818	0.716	39.69	85.27	13
	Relief [20]	0.817	0.701	41.96	84.93	13
	MultiSurf [21]	0.814	0.701	40.51	85.34	12
	Ensemble FS [22]	0.910	0.741	36.54	83.53	11
	Anova [23]	0.813	0.704	39.34	85.45	14
	M-Cluster FS [24]	0.862	0.817	32.68	87.68	13
	SU-MLP [25]	0.869	0.821	31.79	88.95	13
	Proposed	0.874	0.825	31.42	89.34	12
PCA	Chi-square [19]	0.834	0.705	25.52	86.19	12
	Relief [20]	0.827	0.719	26.34	85.82	12
	MultiSurf [21]	0.829	0.715	25.64	86.45	13
	Ensemble FS [22]	0.918	0.749	22.49	86.57	12
	Anova [23]	0.827	0.721	27.54	85.73	12
	M-Cluster FS [24]	0.884	0.815	21.53	90.27	12
	SU-MLP [25]	0.889	0.821	20.84	90.95	12
	Proposed	0.894	0.826	20.53	91.76	11

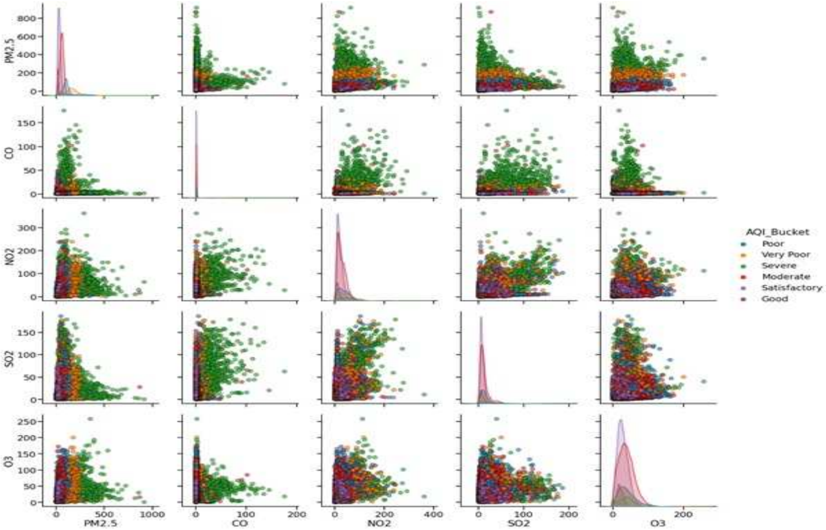


Figure 8. Correlation between the major air pollutants and the air quality index

References

[1] X. Wang, B. Guo, Y. Shen, C. Zhou, and X. Duan, "Input feature selection method based on feature set equivalence and mutual information gain maximization," *IEEE Access*, vol. 7, no. 1, pp. 151525–151538, 2019. DOI: 10.1109/ACCESS.2019.2948095.

[2] F. Macedo, M. R. Oliveira, A. Pacheco, and R. Valadas, "Theoretical foundations of forward feature selection methods based on mutual information," *Neurocomputing*, vol. 325, pp. 67–89, 2019.

[3] S. P. Potharaju and M. Sreedevi, "A novel cluster of quarter feature selection based on symmetrical uncertainty," *Gazi University Journal of Science.*, vol. 31, no. 2, pp. 456–470, 2018.

[4] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowledge-Based Systems*, vol. 140, pp. 103–119, 2018.

[5] S. P. Potharaju and M. Sreedevi, "Distributed feature selection (DFS) strategy for microarray gene expression data to improve

the classification performance," *Clinical Epidemiology and Global Health*, vol. 7, no. 2, pp. 171–176, 2019.

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[7] J. Xie, M. Wang, and Q. Hu, "The differentially expressed gene selection algorithms for unbalanced gene datasets by maximize the area under ROC," *Journal of Shaanxi Normal University (Natural Science Edition)*, vol. 1, no. 1, pp. 01–11, 2017.

[8] Y. Sun and J. Li, "Iterative RELIEF for feature weighting," in *Proceedings of the 23rd international conference on Machine learning*, vol. 1, no. 1, pp. 913–920, 2006.

[9] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Applied Soft Computing*, vol. 13, no. 1, pp. 211–221, 2013.

[10] S. Sharmin, M. Shoyaib, A. A. Ali, M. A. H. Khan, and O. Chae, "Simultaneous feature selection and discretization based on mutual information," *Pattern Recognition*, vol. 19, no. 29, pp. 162–174, 2019.

[11] J. Che, Y. Yang, L. Li, X. Bai, S. Zhang, and C. Deng, "Maximum relevance minimum common redundancy feature selection for nonlinear data," *Information Sciences*, vol. 409, pp.68–86, 2017.

[12] P. Zhang, W. Gao, and G. Liu, "Feature selection considering weighted relevancy," *Applied Intelligence*, vol. 48, no. 12, pp. 4615–4625, 2018.

[13] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine learning research*, vol. 5, no. 9, pp. 1–11, 2014.

[14] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[15] Z. Zeng, H. Zhang, R. Zhang, and C. Yin, "A novel feature selection method considering feature interaction," *Pattern Recognition*, vol. 48, no. 8, pp. 2656–2666, 2015.

[16] Hu, Liang and Gao, Wanfu and Zhao, Kuo and Zhang, Ping and Wang, Feng. “Feature selection considering two types of feature relevancy and feature interdependency,” *Expert Systems with Applications*, vol. 93, pp. 423–434, 2018.

[17] Kolli, Srinivas, M.Sreedevi, “A Novel Granularity Optimal Feature Selection based on Multi-Variant Clustering for High Dimensional Data,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol.12, no. 3, pp. 5051–5062, 2021.

[18] Potharaju, Sai Prasad and Sreedevi, M, “A novel LtR and RtL framework for subset feature selection (reduction) for improving the classification accuracy,” *Advanced Computing and Intelligent Engineering*, pp. 215–224, 2019.

[19] Jin, Xin and Xu, Anbang and Bie, Rongfang and Guo, Ping, “Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles,” *International workshop on data mining for biomedical applications*, pp. 106–115, 2016.

[20] Mesin, Luca and Orione, Fiammetta and Taormina, Riccardo and Pasero, Eros, M, “A feature selection method for air quality forecasting,” *International Conference on Artificial Neural Networks*, pp. 489–494, 2010.

[21] Urbanowicz, Ryan J and Meeker, Melissa and La Cava, William and Olson, Randal S and Moore, Jason H, “Relief-based feature selection: Introduction and review,” *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.

[22] Chen, Chih-Wen and Tsai, Yi-Hong and Chang, Fang-Rong and Lin, Wei-Chao, “Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results,” *Expert Systems*, vol.37, no. 15, pp. 1–15, 2020.

[23] Ding, Hui and Feng, Peng-Mian and Chen, Wei and Lin, Hao, “Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis,” *Molecular BioSystems*, vol.10, no. 8, pp. 2229–2235, 2014.

[24] Potharaju, Sai Prasad and Sreedevi, M, “A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for

Increasing Classification Accuracy of Medical Datasets," *Journal of Engineering Science & Technology Review*, vol.10, no. 6, pp. 1-8, 2019.

[25] Potharaju, Sai Prasad and Sreedevi, M and Amiripalli, Shanmuk Srinivas, M, "An ensemble feature selection framework of sonar targets using symmetrical uncertainty and multi-layer perceptron (su-mlp)," *Cognitive Informatics and Soft Computing*, pp. 247–256, 2019.

[26] Ministry of Environment Forest and Climate CHange, Government of India, "Central Pollution Control Board," *https://cpcb.nic.in/*, accessed December 7, 2020.

Usharani Bhimavarapu, M. Sreedevi

Received January 21, 2022
Accepted June 23, 2022

Usharani Bhimavarapu
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India.
E–mail: `ushareddy@kluniversity.in`

M. Sreedevi
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India.
E–mail: `msreedevi27@kluniversity.in`