Chapter: 08

AN APPROACH TO USING FEATURE SELECTION FOR CLASSIFICATION AND REGRESSION DILEMMA THROUGH FILTER METHOD

Mr. Anuj Kumar*
Faculty, Glocal School of Science and Technology,
Glocal University, Saharanpur, U.P.
**Correspondence to: anuj.kumar@theglocaluniversity.in*

Mohd Hyder Gouri
Faculty, Glocal School of Science and Technology,
Glocal University, Saharanpur, U.P.

DOI: https://doi.org/10.52458/9789388996747.nsp2023.eb.ch-08

Ch.Id:-GU/NSP/EB/EFMLDSP/2023/Ch-08

ABSTRACT

Given the prevalence of extremely complex collections of data, feature selection – one of the essential challenges of machine learning – has received greater emphasis. It shows the vital aspects of a specific problem. Traditionally, it has been used to solve many different issues in fields such as banking (e.g., detecting fraud), surveillance systems, healthcare institutions (e.g., cancer detection), and biological data processing. It helps to get rid of all irrelevant variables and boosts classifier efficiency and precision. Initially, the model's construction should be made simpler by having fewer parameters, taking less time to train, improving applicability to prevent excessive filling, and staying away from the curse known as dimensionality. Filter methods have significance over feature selection since they can be used with any kind of machine-learning model and significantly shorten the time it takes for machine-learning algorithms to run. The

assessments are intended to review the operation of various filter methods, evaluate how effectively they perform in terms of run time and estimated efficiency, and offer recommendations for usage.

Keywords: *Feature selection, Filter method, Variance Threshold, SelectKBest, Information Gain, Select Percentile.*

8.1 INTRODUCTION

The procedure for identifying the most effective and relevant combination of attributes in a set of data to improve a machine learning model's efficacy and precision is known as feature selection. Because we now live in a digital age, there is an obvious rise in the amount of data generated by various applications both horizontally and vertically in terms of rows and columns. It puts a strain on analytics and makes machine learning algorithms that recognize patterns more difficult to use [1]. The growing amount of healthcare data gathering in recent years has allowed professionals to diagnose individuals with greater accuracy. Machine learning is emerging as a key tool in the healthcare sector to help with illness diagnostics. Whenever an endeavor is massive or difficult to program, like interpreting health records data, establishing epidemic forecasting, or even evaluating information about genes, machine learning is an effective tool for analysis [2]. Social networking sites, geographic-based services, multimedia platforms, along online sales are just a few of the most recent applications that have arisen as centers for information collection and sharing over the past ten years due to innovations of technological advances [3]. Merely noteworthy characteristics are chosen using feature selection, which does not create new features; rather, instead assesses features using evaluation functions. Real-world consequences follow the final feature subsets that are acquired through feature selection. The features that were eliminated included excessive & unnecessary aspects, whereas those that were previously identified as significant elements are also referred to as relevant features. Repetitive features include excessive and unneeded information, whereas irrelevant features are those that are unrelated to the operations at concern [4]. Based on an association among classification algorithms, there are three primary methods exist for selecting features: filter method, wrapper method, and embedded method. Filter method, To be able to select (filter) the variables that are input that will be the ones employed by the model, filter feature selection methods employ statistical approaches to assess each input variable's relevance to the target variable. These evaluations serve as the foundation for the selection process [5]. Wrapper method, when attempting to pick features that produce the most efficient model based on an outcome metric, wrapper feature selection methods generate several models using various subgroups of input features. These

approaches may become expensive to compute, but they don't care about the variable types. One effective wrapper feature selection technique is Recursive Feature Elimination (RFE) [6]. Embedded method, this method, combines the benefits of both filter and wrapper methods.

8.2 LITERATURE REVIEW

In 2019, B. Venkatesh et al. studied that data is produced continuously and increases quickly with the rise of the Internet of Things and applications that are web-based, increasing the likelihood of jumbled data and degrading computational efficiency. The adaptability of the feature selection approaches is compromised as the data set grows in size. They concluded that the implementation of feature selection by using the filter method is faster or more accurate as compared with the wrapper method [1].

In 2020, Anna Karen Garate-Escamila et al. employed a classification-based model to detect heart disease by implementing feature selection and principal component analysis. They proposed a method to reduce dimensionality & feature selection strategy to identify heart disease attributes [2].

In 2018, Xuelian Deng et al. examined feature selection techniques for the classification of text. By implementing the wrapper approach, embedded approach, filter approach, and hybrid approach, they found out that the approach that works best is the filter approach, which has also been thoroughly studied in text classification [3].

In 2021, Hongfang Zhou et al. implemented a correlation coefficient to determine how various features relate to one another, adding bilateral data as well as the correlation coefficient. The significance of the repetitive item indicated by the interconnected data within the assessment criteria was determined by taking the mean of the relationship factor among two distinct attributes [4].

In 2019, Andrea Bommert et al. studied the functionality of several filter techniques, contrasting their results in terms of run time as well as estimated precision along with providing recommendations regarding deployments are the objectives of the assessments [5].

In 2019, Haoyue Liu et al. employed a comprehensive approach to feature selection for unbalanced data categorization, which is frequently found in datasets from the actual economy. They established an embedded feature selection method by using the weighted Gini index. After that, they compared the results with the parameter selection techniques of Chi2, F-statistic, and Gini index. F-statistic and Chi2 perform best

when just a small number of characteristics are chosen. The strategy they suggest has the greatest chance of producing the best possible outcomes when the number of selected attributes expands [6].

In 2019, Beatriz Remeseiroa et al. examined a variety of modern methods for picking features created to be used in healthcare applications, encompassing hot topics in research like genome microarray data interpretation, physiologic processing of signals, or medical image processing [7].

In 2020, Rung-Ching Chen et al. chose essential elements for machine learning-based categorization of information. To get the most effective proportion correctness, they also examine the dataset's outcome irrespective of a choice of critical features using the Random Forest techniques, and Recursive Feature Elimination [8].

In 2018, Rui Zhang et al. explained that feature-level fusion, which is additionally referred to by the low-level fusion approach, continues to be investigated as the standpoint from the fundamental notion interpretation, and apps utilized within investigations based on the feature selection techniques previously discussed [9].

In 2022, Kamal A. ElDahshan et al. addressed the difficulties with big data representation by combining both the feature selection approaches which are filtered and embedded, to preserve the integrity of information while simultaneously lowering big data density & classifier duration for training [10].

In 2019, Saúl Solorio-Fernández et al. examined the majority of unsupervised feature selection techniques such as embedded, filter, hybrids, or wrapper, necessitate identification of extreme parameters, such as the variety during features, the number of groupings, and additional factors that vary depending on the technique for choosing features employed through each approach, concerning their analysis of the literature. Yet, this expertise does not exist in routine, which means it is frequently hard to determine the ideal parameter values for every given dataset. As a result, choosing the optimal parameter values automatically is still an ongoing issue [11].

In 2019, Veronica Bolon-Canedo et al. gave the users fundamental ideas required to assemble a combination of feature selections besides evaluating recent advancements while offering commentary regarding forthcoming developments that remain being addressed [12].

In 2019, Utkarsh Mahadeo Khaire et al. compiled remedies to the numerous points of unstable, on the basis of the feature data approach, feature proximity, sampling

weighing, parameter to be used improvement, rigorous search method, cluster feature selection, or collective feature selection [13].

In 2019, Gang Koua et al. studied feature selection techniques for text categorization employing numerous parameters on limited datasets establishing choices. To examine the suggested methodology, a trial is planned that combines five 'multiple-criteria-decision-making' methods with ten feature selection methods, nine different binary classification appraisal measures, a total of seven multi-class categorization evaluation measures, and three classifiers using ten limited data sets [14].

In 2021, Mohammad Pourhomayoun et al. employed a machine learning model by examining feature selection approaches to estimate the risk of death in COVID-19 patients to support healthcare industries. The most concerning signs as well as traits were identified throughout this investigation [15].

8.3 METHODOLOGY

Filtering techniques are usually utilized as an initial phase. The machine learning approaches provide minimal impact on the feature selection process. Rather, attributes have been selected depending on how well they correlate alongside the final parameter according to what comes out of several statistical analyses.

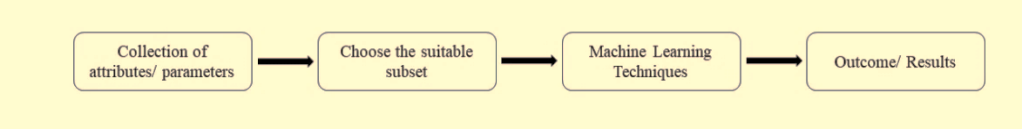

Figure-8.1: Filter Method Approach

Some of the main properties of the filter method are the following:

- Filter methods don't involve machine learning algorithms; instead, they depend on the features (or attributes) of the information being analyzed.
- Filter methods don't care about models.
- They are typically less expensive to compute.
- Generally speaking, filter methods perform less well in predictions compared to wrapper techniques.
- Filter methods perform effectively towards swiftly examining while eliminating things that aren't needed.

8.4 IMPORTANCE OF FEATURE SELECTION

a. Dimensionality Curse

Whenever there are too many attributes, and when we possess additional attributes compared to what is needed for developing the model, it means that we find ourselves unable to train the model effectively since such attributes would negatively impact the model or reduce its precision. For example, suppose we are developing a model to predict heart diseases and have features like cp, BMI, resting, old peak, fertilizers, rainfall, and pH value. At this point, we can observe that the dataset contained a large number of irrelevant variables, such as rainfall, pH value, and fertilizers, which prevents us from feeding it to our model for use in training.

There are two potential sources of the dimensionality curse.

- Extraneous details
- Superfluous features (i.e., height in meters and height in centimeters provided as separate features)

b. Solutions to Dimensionality Curse

- **Feature Selection:** Here, we determine our most valuable features, and, more accurately, we identify the attributes that will improve our precision and outcomes.
- **Feature Extraction:** The term "feature extraction" refers to techniques that combine and choose variables to create features, which reduces the quantity of information required to get evaluated yet retains the precision and comprehensiveness of the initial set of data.

8.5 FILTER METHOD

a. Data Set

In this study, we used two datasets (i.e., clinical records of heart failure and car price) that are accessible publicly on Kaggle. Additionally, both datasets fall under data with varying with number of instances and features. The dataset Clinical Records of Heart Failure has 209 instances and 13 features whereas the dataset car price has 205 instances and 13 features.

b. Variance Threshold

We can establish a minimal threshold for an acceptable deviation in each feature by using the variance threshold. Features that have a nothing (zero) deviation (identical values across all instances) from the mean are constant and can be eliminated.

Table-8.1 (a) & (b) demonstrates the elimination of those features which have low variance. As you can see none of the used features had a zero variance, so we are not going to remove any feature.

Table-8.1: Elimination of Features that have Low Variance

(a) Clinical Record for Heart Failure

	Feature	**Variance**
10	smoking	0.21
5	high_blood_pressure	0.23
9	Sex	0.23
3	Diabetes	0.24
1	anaemia	0.24
7	serum_creatinine	1.00
8	serum_sodium	18.56
0	Age	55.00
4	ejection_fraction	60.00
11	Time	281.00
2	creatinine_phosphokinase	7838.00
6	platelets	559000.00

(b) Car Price

	Feature	**Variance**
6	boreratio	0.073
7	stroke	0.094
2	carwidth	4.73
3	carheight	6.12
10	citympg	25.00
0	wheelbase	29.00
11	highwaympg	31.00

1	carlength	58.00
8	horsepower	236.00
5	enginesize	256.00
4	curbweight	2247.00
9	peakrpm	2450.00

8.6 SELECTKBEST METHOD – CLASSIFICATION DILEMMA

By using the k maximum scores, the SelectKBest technique chooses the features. You can use the approach with either classification or regression data by adjusting the 'score_func' option. When it comes time to prepare a huge dataset during training, choosing the optimal features is a crucial step in the process. It facilitates the elimination of less crucial data and shortens the training period.

Table-8.2 shows the selected features from all the features that reside in our dataset (Clinical record for heart failure) along with the KBest score.

Table-8.2: Selection of features using KBest Classifier from the existing features

	Feature	**KBest Classification Score**	**Selected Feature**
0	age	19.13	✔
1	anaemia	0.13	
2	creatinine_phosphokinase	0.10	
3	diabetes	0.02	
4	ejection_fraction	13.35	✔
5	high_blood_pressure	0.46	
6	platelets	0.09	
7	serum_creatinine	27.95	✔
8	serum_sodium	6.24	✔
9	sex	1.08	
10	smoking	0.35	✔
11	time	82.18	✔

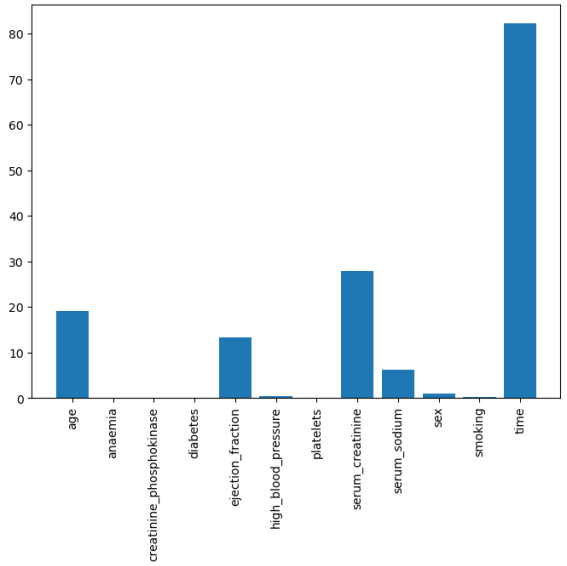

Figure-8.2: Feature KBest Classification Score from dataset Clinical record for heart failure

As shown in the above Figure-8.2, we got only six features out of twelve features which are age, ejection_fraction, serum_creatinine, serum_sodium, sex, and, time by implementing the SelectKBest method.

After employing the logistic regression technique as a foundation model, the comparison by using all of the features (getting a score of 0.76%) in the dataset, we achieved a higher score by utilizing just six of them (getting a score of 0.78%).

8.7 SELECTKBEST METHOD – REGRESSION DILEMMA

Table-8.3 shows the selected features from all the features that reside in our dataset (i.e., car price) along with the KBest score.

Table-8.3: Selection of features using KBest Regressier from the existing features

	Feature	KBest Regression Score	Selected Feature
0	wheelbase	61.27	
1	carlength	116.94	
2	carwidth	206.71	✔
3	carheight	0.51	

4	curbweight	335.17	✔
5	enginesize	468.58	✔
6	boreratio	59.07	
7	stroke	0.21	
8	horsepower	271.30	✔
9	peakrpm	1.67	
10	citympg	167.59	✔
11	highwaympg	180.56	✔

As shown in Figure-8.3, by implementing the SelectKBest method for regression dilemma, we retrieve only 6 features which are carwidth, carweight, enginesize, horsepower, citympg, and, highwaympg.

Despite utilizing only 6 features, we ultimately failed to accomplish an improved score. We found all features scored 0.79% and only selected features scored 0.76%.

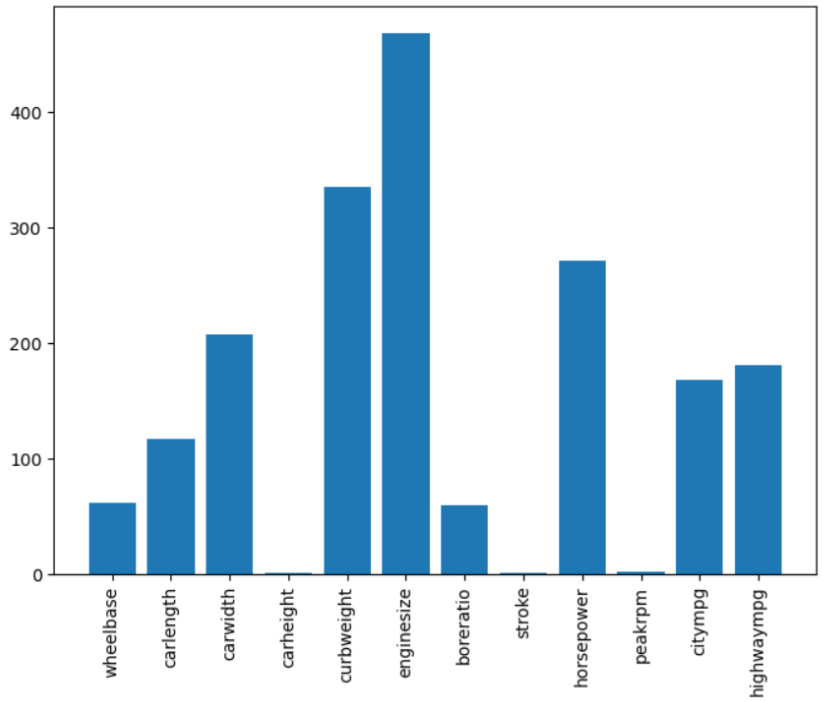

Figure-8.3: Feature KBest Regression Score from dataset car price

8.8 INFORMATION GAIN METHOD – CLASSIFICATION DILEMMA

The decrease in complexity resulting from a dataset's modification is computed as information gain. By assessing every parameter's, information gain concerning the target variable, it can be further applied to feature selection.

Table-8.4 Selection of features Information Gain Classifier from the existing features

	Feature	Information Gain Classifier score	Selected Feature
0	age	0.09	✔
1	anaemia	0.0	
2	creatinine_phosphokinase	0.001	
3	diabetes	0.015	
4	ejection_fraction	0.071	✔
5	high_blood_pressure	0.009	✔
6	platelets	0.018	
7	serum_creatinine	0.101	✔
8	serum_sodium	0.062	✔
9	sex	1.076	
10	smoking	0.0	
11	time	0.236	✔

As shown in Figure-8.4, by implementing the Information Gain method for the classifier dilemma, we retrieve only 6 features that are age, ejection_fraction, high_blood_pressure, serum_creatinine, serum_sodium, and time. We found all features scored 0.75% and only selected features scored 0.78% which is far better as compared to all features scored.

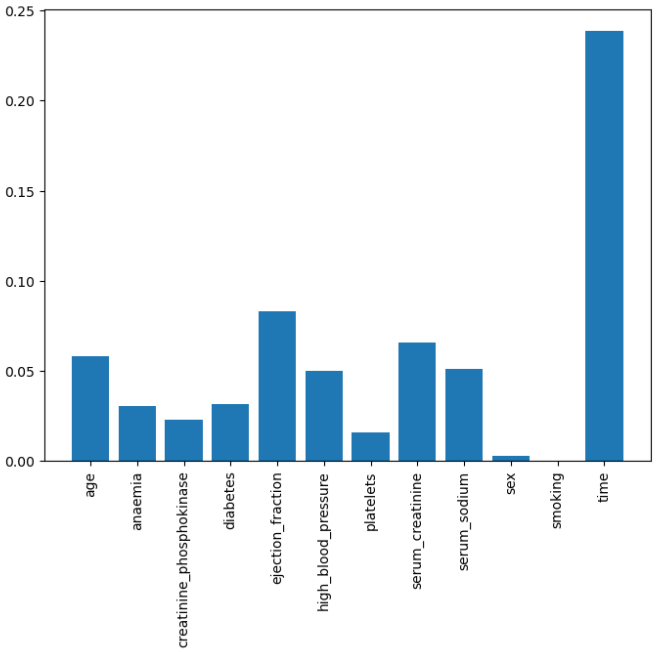

Figure-8.4: Feature Information Gain Score from the dataset Clinical record for heart failure

8.9 INFORMATION GAIN METHOD – REGRESSION DILEMMA

Table-8.5: Selection of features Information Gain Regressier from the existing features

	Feature	Information Gain Regression Score	Selected Feature
0	wheelbase	0.52	
1	carlength	0.62	✔
2	carwidth	0.55	
3	carheight	0.23	
4	curbweight	0.92	✔
5	enginesize	0.76	✔
6	boreratio	0.44	
7	stroke	0.27	
8	horsepower	0.80	✔
9	peakrpm	0.19	

10	citympg	0.85	✔
11	highwaympg	0.84	✔

Figure-8.5 illustrates that after the implementation of the Information Gain method for the regression dilemma, we found only 6 features which are carlength, carweight, enginesize, horsepower, citympg, and, highwaympg. We found whole features scored 0.79% and only selected features scored 0.78%. We ultimately failed to accomplish an improved score. It might be conceivable that a few of the features are oblique and don't reveal anything concerning the desired variable.

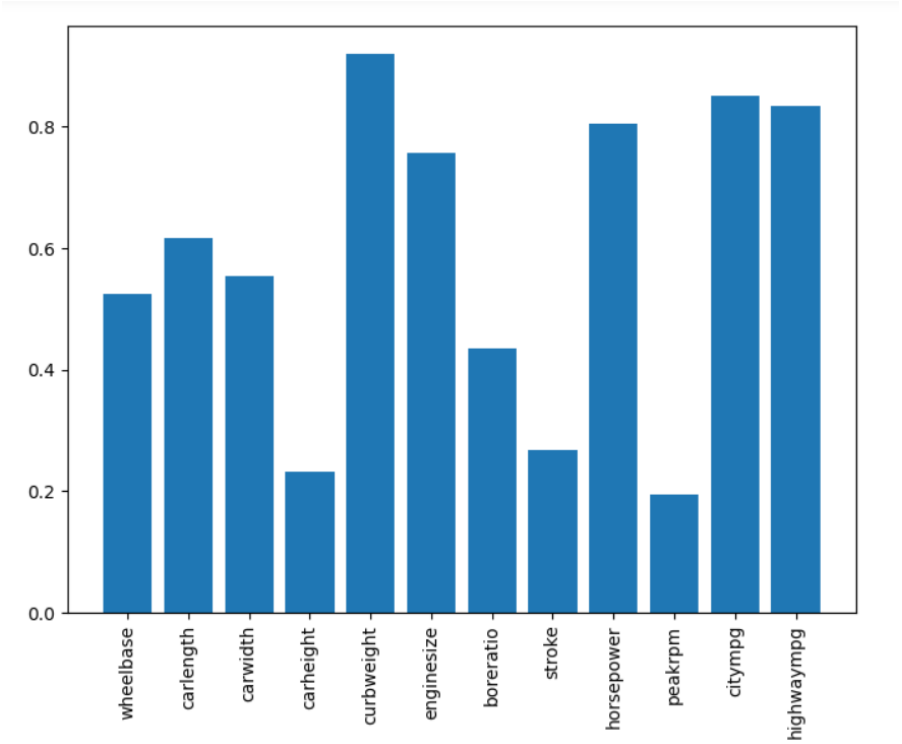

Figure-8.5: Feature Information Gain Regression Score from dataset car price

8.10 SELECTPERCENTILE METHOD

In this version using the K-Best feature selection method, the most outstanding variable's % of the highest-scoring features is chosen. Thus, if the specific variable in our example is 50%, we want to choose attributes that fall in the top 50 percentile according to their evaluations. The one difference between the procedure with SelectKBest has to do with we supply a percentile in the SelectPercentile object that is

generated. After the implementation of SelectPercentile method on the dataset Clinical record for heart failure, we found only 6 selected features which are age, time, ejection_fraction, sex, serum_creatinine, and serum_sodium.

8.11 CONCLUSION AND FUTURE SCOPE

Those interested in conducting filtering might use this work to be a source of information. This might help users in selecting the right filters per their particular application circumstances, and existing computing resources, among other factors. The filter model performs feature selection in an initial processing phase. We haven't put forth an effort to enhance efficiency. We choose an area with characteristics that optimize the simulation performance by applying the filter model. Finding a feature subset with minimal feature recurrence as well as significant relevance to the classification is the aim of this study. It is possible that some of the features fail to inform us enough about the target factor and thus remain uninformative. In the future, we will conduct another study on feature selection by comparing the filter method with the different kinds of feature selection methods which are wrapper method, embedded method, and hybrid method.

REFERENCES

1. *B. Venkatesh, J. Anuradha (2019), A Review of Feature Selection and Its Methods, Bulgarian Academy of Sciences, Cybernetics and Information Technologies, Volume 19.*
2. *Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres (2020), Classification models for heart disease prediction using feature selection and PCA, Informatics in Medicine Unlocked, ScienceDirect,*
3. *Elsevier Xuelian Deng, Yuqing Li, Jian Weng, Jilian Zhang (2018), Feature selection for text classification: A review, Multimed Tools Appl, Springer.*
4. *Hongfang Zhou, Xiqian Wang, Rourou Zhu (2021), Feature selection based on mutual information with correlation coefficient, Applied Intelligence, Springer*
5. *Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, Michel Lang, (2019), Benchmark for filter methods for feature selection in high-dimensional classification data, Computational Statistics and Data Analysis, ScienceDirect, Elsevier.*
6. *Haoyue Liu, Student Member, MengChu Zhou, Qing Gary Liu (2019), An Embedded Feature Selection Method for Imbalanced Data Classification, IEEE/CAA Journal of Automatica Sinica.*

7. *Beatriz Remeseiroa, Veronica Bolon-Canedob (2109), A review of feature selection methods in medical applications, Computers in Biology and Medicine, ScienceDirect, Elsevier.*

8. *Rung-Ching Chen, Christine Dewi, Su-Wen Huang, Rezzy Eko Caraka (2020), Selecting critical features for data classification based on machine learning methods, Journal of Big Data, Springer Open.*

9. *Rui Zhang, Feiping Nie, Xuelong Li, Xian Wei (2018), Feature Selection with Multiview Data: A Survey, Information Fusion, ScienceDirect, Elsevier.*

10. *Kamal A. ElDahshan, Abdullah A. AlHabshy and Luay Thamer Mohammed (2022), Filter and Embedded Feature Selection Methods to Meet Big Data Visualization Challenges, Computers, Materials & Continua.*

11. *Saul Solorio-Fernandez, J. Ariel Carrasco-Ochoa, Jose Fco. Martínez-Trinidad (2019), A review of unsupervised feature selection methods, Artificial Intelligence Review, Springer.*

12. *Veronica Bolon-Canedo, Amparo Alonso-Betanzos (2019), Ensembles for feature selection: A review and future trends, Information Fusion, ScienceDirect, Elsevier.*

13. *Utkarsh Mahadeo Khaire, R. Dhanalakshmi (2019), Stability of feature selection algorithm: A review, Journal of King Saud University –Computer and Information Sciences, ScienceDirect.*

14. *Gang Koua, Pei Yanga, Yi Pengb, Feng Xiaoa, Yang Chena, Fawaz E. Alsaadic (2019), Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods, Applied Soft Computing Journal.*

15. *Mohammad Pourhomayoun, Mahdi Shakibi (2021), Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making, Smart Health, ScienceDirect, Elsevier.*