



An Analysis of Feature Selection Techniques in Machine Learning

Author : **Gargi Janak Bapodariya**

Program: MSC Data Science

College: Kandivali Education Society's BK Shroff College of Arts and MH Shroff College of Commerce
(Autonomous)

bapodariyagargi12@gmail.com

Abstract

Machine learning is widely used for solving problems in different areas by learning patterns from data. However, real-world datasets may contain a large number of features, including irrelevant, redundant, or less informative features. Using all available features can increase computational requirements, make models more complex, and sometimes negatively affect prediction performance. Feature selection is an important data preprocessing technique that focuses on selecting the most relevant features from a dataset while removing unnecessary features. This research paper focuses on understanding and analyzing different feature selection techniques used in machine learning. It discusses the major categories of feature selection methods, including filter methods, wrapper methods, and embedded methods. Techniques such as correlation-based selection, chi-square, mutual information, Recursive Feature Elimination (RFE), Lasso, and tree-based feature importance are considered. The study also examines the advantages, limitations, computational requirements, and influence of different feature selection approaches on machine learning models. The main purpose of this study is to understand how selecting relevant features can reduce data complexity, improve model interpretability, reduce computational requirements, and potentially improve machine learning performance. The study uses descriptive and comparative analysis based on existing research and experimental results obtained from selected datasets and machine learning models.

How to Cite this Article:

Bapodariya, G. J. (2026). An Analysis of Feature Selection Techniques in Machine Learning. International Journal of Creative and Open Research in Engineering and Management, 2(10), 1-9.
<https://doi.org/10.55041/ijcope.v2i9.248>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i9.248>

Keywords: Feature Selection, Machine Learning, Feature Engineering, Filter Methods, Wrapper Methods, Embedded Methods, Dimensionality Reduction, Classification.

Introduction

Machine learning is an important field of Data Science that allows computers to learn patterns from data and use those patterns to make predictions or decisions. A dataset used for machine learning generally contains multiple features or variables that describe the observations. As datasets become larger and more complex, the number of available features can also increase. Not every feature in a dataset is equally useful for predicting the target variable. Some features may contain useful information, while others may be irrelevant, redundant, noisy, or highly correlated with other features. Including unnecessary features can increase computational requirements and make the machine learning model more difficult to understand.

Feature selection is the process of selecting a subset of relevant features from the original dataset. Unlike feature extraction, feature selection keeps the original features and removes those that are considered less useful. Research commonly categorizes feature selection techniques into three major groups: filter methods, wrapper methods, and embedded methods. Filter methods generally evaluate features using statistical or information-based criteria before the machine learning model is trained. Wrapper methods evaluate different feature subsets according to the performance of a selected machine learning model. Embedded methods perform feature selection during the model



training process.

Feature selection can therefore be useful in machine learning because it can reduce the number of input variables, decrease computational requirements, improve interpretability, and in some cases improve predictive performance.

This research focuses on analyzing the major feature selection techniques and understanding their characteristics, advantages, limitations, and influence on machine learning classification performance.

Literature Review

Feature selection has been widely studied in machine learning and pattern recognition. Earlier research established the importance of removing irrelevant and redundant variables, particularly as datasets became increasingly high-dimensional. A widely used classification of feature selection techniques divides them into filter, wrapper, and embedded approaches. Filter methods evaluate features independently of a particular machine learning model. They generally use statistical measures or information-based criteria to rank features. Examples include correlation, chi-square and mutual information. These methods are usually computationally less expensive and can be applied before model training.

Wrapper methods use the performance of a machine learning model to evaluate feature subsets. Techniques such as Recursive Feature Elimination and sequential feature selection belong to this category. Wrapper methods can identify feature combinations that work well for a particular model, but they may require greater computational time because several feature subsets and model configurations need to be evaluated. Embedded methods perform feature selection as part of the model training process. Examples include Lasso-based feature selection and feature importance generated by tree-based models. These approaches can provide a balance between the computational simplicity of filter methods and the model-specific selection of wrapper methods.

Recent research continues to examine feature selection for high-dimensional and complex datasets. Studies also consider additional issues such as stability, computational cost, predictive performance, redundancy and the effect of feature selection on different machine learning algorithms. Feature selection is also different from dimensionality reduction techniques such as Principal Component Analysis. Feature selection retains selected original variables, whereas dimensionality reduction may transform the original variables into a new set of components.

Research Gap

Feature selection has been extensively studied; however, there is still a need to understand how different feature selection approaches behave when applied to different datasets and machine learning algorithms. Many studies focus on developing a particular feature selection algorithm or applying it to a specific domain. However, for Data Science students and practitioners, it is also important to understand the practical differences between commonly used techniques. Different feature selection methods may select different features from the same dataset. A method that performs well with one machine learning algorithm may not necessarily provide the same results with another algorithm. Computational requirements can also differ considerably between filter, wrapper and embedded methods. Therefore, this research focuses on providing a descriptive and comparative understanding of commonly used feature selection techniques. The study examines their working principles, advantages, limitations, computational requirements, number of selected features and influence on machine learning classification performance.

Areas to focus:

This research focuses on 5 main areas that are studied in this paper:

1. **Feature Relevance:** Understanding how relevant features are identified.
2. **Feature Redundancy:** Understanding how unnecessary or repeated information can be reduced.
3. **Method Comparison:** Comparing filter, wrapper and embedded techniques.



4. **Model Performance:** Examining the effect of feature selection on machine learning performance.
5. **Computational Efficiency:** Understanding the relationship between feature selection and computational requirements.

Research Objective

The main purpose of this research is to analyze different feature selection techniques used in machine learning and understand their effect on classification performance and computational requirements.

The study aims to achieve the following objectives:

1. To understand the concept and importance of feature selection in machine learning.
2. To identify and study different categories of feature selection techniques.
3. To analyze filter, wrapper and embedded feature selection methods.
4. To compare commonly used feature selection techniques based on selected features and model performance.
5. To examine the effect of feature selection on machine learning classification models.
6. To understand how feature selection can reduce computational complexity.
7. To identify the advantages and limitations of different feature selection approaches.
8. To contribute to existing research by providing a practical comparison of commonly used feature selection techniques.

Research Methodology

This study follows a descriptive and comparative research methodology to analyze feature selection techniques used in machine learning. The research is divided into two major parts. The first part is based on secondary data collected from research papers, journals, academic articles and reliable online sources. This information is used to understand the concepts, categories, working principles, advantages and limitations of feature selection methods. The second part can involve an experimental comparison using publicly available machine learning datasets. A suitable classification dataset is selected and basic preprocessing is performed. Different feature selection techniques are then applied to the same dataset. The selected features are used to train machine learning classification models. The results are compared with the results obtained using the complete feature set.

The comparison can consider:

- Number of original features
- Number of selected features
- Accuracy
- Precision
- Recall
- F1-score
- Training time
- Model complexity

The general research workflow is:

Dataset → Data Preprocessing → Feature Selection → Machine Learning Model → Model Evaluation → Comparison → Analysis



Types/Methods of Feature Selection

The following major categories of feature selection methods are considered in this research.

1. Filter Methods

Filter methods select features using statistical or information-based criteria without depending strongly on a particular machine learning algorithm.

Common examples include:

- Correlation-based selection
- Chi-square test
- Mutual information
- Variance-based selection

Filter methods are generally simple and computationally efficient. They can be useful when datasets contain a large number of features.

2. Wrapper Methods

Wrapper methods use a machine learning algorithm to evaluate different subsets of features. The model is trained using different combinations of features and the subset that produces suitable performance is selected.

Examples include:

- Recursive Feature Elimination
- Sequential Forward Selection
- Sequential Backward Selection

Wrapper methods can provide model-specific feature subsets, but they may require more computational resources because multiple models need to be trained and evaluated.

3. Embedded Methods

Embedded methods perform feature selection during the model training process. Examples include:

- Lasso Regression
- Decision Tree Feature Importance
- Random Forest Feature Importance
- Regularization-based methods

These methods integrate feature selection with model training and can provide a useful balance between filter and wrapper approaches.

4. Correlation-Based Feature Selection

Correlation measures the relationship between variables. If two features contain highly similar information, one of them may be considered redundant. Correlation analysis can therefore be used to identify highly correlated features and reduce unnecessary variables. However, correlation mainly captures certain forms of relationships and should not be treated as the only criterion for selecting features.



5. Chi-Square Feature Selection

The chi-square test can be used to examine the relationship between categorical features and a categorical target variable. Features showing stronger association with the target can receive higher importance during selection.

6. Mutual Information

Mutual information measures the amount of information that one variable provides about another variable. It can identify relationships that may not be captured through simple correlation. Therefore, it can be useful for selecting informative features.

7. Recursive Feature Elimination

Recursive Feature Elimination, commonly known as RFE, repeatedly trains a model and removes less important features. The process continues until the required number of features remains. RFE is useful when the researcher wants to select a specific number of features, but it can require more computational time than simple filter methods.

8. Lasso-Based Feature Selection

Lasso applies regularization that can shrink some model coefficients toward zero. Features with coefficients reduced to zero can consequently be excluded from the model. This provides an embedded approach to feature selection.

9. Tree-Based Feature Selection

Decision Trees and ensemble tree models can provide measures of feature importance. Features that contribute more strongly to model decisions can receive higher importance values. Random Forest is one example where feature importance can be used for feature selection.

Problems

Several problems can occur when working with feature selection techniques.

1. Irrelevant Features

A dataset may contain features that do not provide useful information for the target variable. Keeping such variables may increase model complexity.

2. Redundant Features

Some features may contain similar or overlapping information. Including many redundant features may not provide additional useful information.

3. High Dimensionality

Datasets containing a large number of features can require more computational resources and may make machine learning models more difficult to train.

4. Overfitting

Selecting features using the same data repeatedly without appropriate validation can result in overfitting. A feature subset may perform well on one dataset split but fail to generalize to new data.



5. Computational Cost

Wrapper methods can become computationally expensive because many different feature subsets may need to be evaluated.

6. Different Methods Select Different Features

There is no guarantee that different feature selection techniques will select the same features. Their results depend on the statistical criteria, search strategy and machine learning model being used.

7. Dataset Dependency

A feature selection technique that performs well on one dataset may not necessarily provide the same results on another dataset.

Solutions

The following approaches can help address the problems associated with feature selection.

1. Compare Multiple Techniques

Instead of depending on a single technique, researchers can compare filter, wrapper and embedded methods.

2. Use Appropriate Evaluation Metrics

Accuracy alone may not provide sufficient information. Precision, recall, F1-score and other suitable metrics can also be considered depending on the classification problem.

3. Use Training and Testing Data Properly

Feature selection should be performed carefully to avoid information leakage from the test data. The selection process should be based on the training data when evaluating generalization performance.

4. Remove Redundant Features

Correlation analysis and other redundancy measures can be used to identify features providing similar information.

5. Consider Computational Requirements

The selection method should be chosen according to the size and complexity of the dataset. Simple filter methods can be useful when computational resources are limited.

6. Compare Number of Features

The researcher should record the number of features before and after selection. This helps demonstrate whether the technique actually reduced the feature space.

7. Evaluate Model Performance

The selected feature set should be tested using suitable machine learning algorithms to determine whether removing features changes model performance.



Discussion

Feature selection is an important part of the machine learning workflow because the quality and relevance of input features can influence the learning process. The analysis of existing literature shows that feature selection methods can generally be grouped into filter, wrapper and embedded approaches. Filter methods are relatively simple and computationally efficient because they evaluate features using statistical or information-based criteria. They can therefore be useful as an initial feature selection step, particularly for datasets containing many variables. Wrapper methods evaluate feature subsets using the performance of a machine learning model. This can provide a feature subset that is more closely related to a particular model, but the process can require more computational resources.

Embedded methods perform selection during model training. They provide an alternative that combines feature selection with the learning process. An important point is that feature selection should not be evaluated only by whether it increases accuracy. A useful feature selection approach may also reduce the number of variables, reduce computational requirements, improve interpretability, or provide a simpler model. For example, if a dataset contains 30 features and a feature selection technique reduces them to 12 features while maintaining similar classification performance, the reduction itself can be considered useful because the model has fewer inputs to process.

The final results may differ depending on the dataset, feature characteristics and machine learning algorithm. Therefore, feature selection should be considered an experimental process rather than assuming that one technique is always suitable for every dataset.

Conclusion

In conclusion, feature selection is an important technique in machine learning for identifying relevant features and reducing unnecessary or redundant variables. As datasets continue to increase in size and complexity, selecting useful features can help make machine learning models more efficient and understandable. This research analyzed the major categories of feature selection techniques, including filter, wrapper and embedded methods. Techniques such as correlation, chi-square, mutual information, Recursive Feature Elimination, Lasso and tree-based feature importance were discussed. The analysis shows that different feature selection techniques have different strengths and limitations. Filter methods are generally simpler and computationally efficient, wrapper methods can provide model-specific feature subsets but may require more computation, and embedded methods perform feature selection during model training. Feature selection should therefore be selected according to the characteristics of the dataset, machine learning problem, computational resources and research objective. Comparing multiple techniques can provide a better understanding of which features are important and how feature reduction affects machine learning performance. Overall, feature selection can support the development of simpler, more efficient and interpretable machine learning models when it is applied carefully and evaluated using appropriate methodology.

References

- Chandrashekar, G., and F. Sahin. "A Survey on Feature Selection Methods." *Computers & Electrical Engineering*, vol. 40, no. 1, 2014, pp. 16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>. Accessed 28 Sept. 2026.
- Moslemi, Amir. "A Tutorial-Based Survey on Feature Selection: Recent Advancements on Feature Selection." *Engineering Applications of Artificial Intelligence*, vol. 126, 2023, article 107136. <https://doi.org/10.1016/j.engappai.2023.107136>. Accessed 28 Sept. 2026.
- Cai, Jie, et al. "Feature Selection in Machine Learning: A New Perspective." *Neurocomputing*, vol. 300, 2018, pp. 70–79. <https://doi.org/10.1016/j.neucom.2017.11.077>. Accessed 28 Sept. 2026.



- Bommert, Andreas, et al. "Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data." *Computational Statistics & Data Analysis*, vol. 143, 2020, article 106839. <https://doi.org/10.1016/j.csda.2019.106839>. Accessed 28 Sept. 2026.
- Villa-Blanco, Carlos, Concha Bielza, and Pedro Larrañaga. "Feature Subset Selection for Data and Feature Streams: A Review." *Artificial Intelligence Review*, vol. 56, 2023, pp. 1011–1062. <https://doi.org/10.1007/s10462-023-10546-9>. Accessed 28 Sept. 2026.
- Li, Jundong, et al. "Feature Selection: A Data Perspective." *ACM Computing Surveys*, vol. 50, no. 6, 2017, article 94. <https://doi.org/10.1145/3136625>. Accessed 28 Sept. 2026.
- Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, 1996, pp. 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>. Accessed 28 Sept. 2026.
- Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, 2005, pp. 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>. Accessed 28 Sept. 2026.