Automatic Feature Discovery: A comparative study between filter and wrapper feature selection techniques

Andrei Mânăstireanu[1]

Supervisor(s): Asterios Katsifodimos[1]**, Andra Ionescu**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Andrei Mânăstireanu
Final project course: CSE3000 Research Project
Thesis committee: Asterios Katsifodimos, Andra Ionescu, Elvin Isufi

Abstract

The curse of dimensionality is a common challenge in machine learning, and feature selection techniques are commonly employed to address this issue by selecting a subset of relevant features. However, there is no consistently superior approach for choosing the most significant subset of features. We conducted a comprehensive analysis comparing filter and wrapper techniques to guide future work in selecting the most appropriate method based on specific circumstances. We quantified the performance of these techniques using a diverse collection of datasets. We utilised simple decision trees, linear machine learning algorithms, and support vector machines to assess the performance with varying percentages of features selected by the filter and wrapper techniques. The findings demonstrate that filter methods (Chi-Squared and ANOVA) perform better than wrapper methods (Forward Selection and Backward Elimination) regarding the classification accuracy, regression root mean squared error, and runtime.

1 Introduction

The emergence of the big data era, driven by the exponential growth of the Internet, has brought about significant advancements in the Machine Learning community. This community widely acknowledges that increasing the number of samples available for training can lead to improved model performance. However, adding more features to a dataset does not guarantee a substantial enhancement in the model performance. Additionally, real-world datasets frequently contain irrelevant or redundant features that hinder the efficiency of data analysis and machine learning tasks. The former features offer no valuable information for the specific problem, while the latter do not provide novel or additional insights.

The curse of dimensionality occurs when a model becomes excessively complex due to the high number of features [11]. This complexity often leads to overfitting of the training data, resulting in poor performance when applied to unseen data [11]. Moreover, the curse of dimensionality also has implications for ”memory storage requirements and computational costs for data analytics” [13].

This research's motivation is mitigating the previously-mentioned challenges associated with high-dimensional data by utilising dimensionality reduction techniques, with a specific focus on feature selection. The study excludes feature extraction [30] due to its added complexity. Feature selection involves selecting a subset of relevant features, leading to plain and more interpretable machine learning models [13]. Additionally, it improves the performance of data mining operations, enhances data organisation, and promotes efficient learning [13]. Objective measures enhanced by including a feature selection step are as follows: ”predictive accuracy, comprehensibility, learning efficiency, compact models, and effective data collection” [14]. Feature selection finds applications in various domains, including gene selection and text classification [8], remote sensing, intrusion detection, image retrieval [11], medical diagnosis, and prognosis [14].

The main focus of this research paper is to conduct an extensive comparative analysis between filter and wrapper feature selection methods. The goal is to answer the research question:

> *How do different feature selection techniques for categorical and numerical data influence the performance of simple decision trees, linear machine learning algorithms and support vector machines?*

This paper seeks to contribute to the advancement of knowledge in the field of feature selection and its implications for machine learning processes by providing a thorough examination of the differences and similarities between filter and wrapper feature selection methods. The findings of this study can be valuable in guiding the adoption of suitable feature selection techniques and optimizing the performance of commonly used machine learning algorithms.

The research paper is structured as follows. The subsequent section provides an overview of the related work conducted in feature selection. Section 3 delves into the methods, formulas, and requirements relevant to the study. Section 4 outlines the methodology employed, encompassing analysis of the datasets, hypotheses, and data preprocessing techniques. In Section 5, we present the primary contribution of this research: an empirical evaluation of the feature selection techniques. Section 6 provides a discussion of the results and addresses the limitations of the study. Conclusions and future work are discussed in Section 7, while Section 8 focuses on responsible research practices.

2 Related Work

We commonly classify feature selection techniques into three main categories: filter, wrapper, and embedded methods. Some studies have proposed a fourth category, hybrid models, combining multiple feature selection algorithms to leverage their complementary strengths [17]. Embedded and hybrid techniques are deliberately excluded from this study as they represent a middle ground between filter and wrapper methods, rendering their inclusion unnecessary for this investigation.

Filter feature selection techniques operate independently of a specific learning algorithm and instead focus on exploiting the inherent characteristics of the training data [13]. These methods are computationally efficient since they do not involve training a model, but they may produce less accurate and robust results [21].

Wrapper feature selection techniques, on the other hand, incorporate a learning algorithm as a black box to evaluate different feature subsets iteratively [13]. Wrapper methods tend to provide more accurate and robust results at the expense of increased computational complexity compared to filter methods [21] due to the usage of a learning model.

In a previous study [29], the impact of filter and wrapper feature selection techniques on logistic regression was explored, focusing on various metrics. The study employed feature selection methods on three different datasets and found that wrapper methods, specifically Sequential Forward Selection and Sequential Backward Elimination, outperformed the filter methods when applied to datasets with continuous features. In contrast, the present study expands upon these findings by investigating a broader range of machine learning models and considering not only continuous features but furthermore categorical and discrete ones.

Several existing studies [22], [3], [2], [15] have highlighted the absence of a universally optimal feature selection method. Therefore, the focus of this study is to delve into the intricacies of the datasets and contribute to the ongoing discourse on feature selection. The findings can serve as a valuable resource for future researchers, helping them make informed decisions in selecting the most appropriate feature selection method based on specific constraints and requirements.

3 Preliminaries

We need to focus on specific feature selection techniques and machine learning algorithms for evaluating the effectiveness of filter and wrapper methods. In this study, we have considered widely used feature selection techniques, namely Chi-Squared and ANOVA for filter techniques and (Sequential) Forward Selection and (Sequential) Backward Elimination for wrapper methods.

3.1 Chi-Squared (Chi2 / χ^2)

The Chi-Squared test is a statistical analysis tool utilised for assessing the correlation between an independent categorical variable and a dependent one [12]. The null hypothesis asserts that there is no significant relationship between the independent and dependent variable [12].

We can compute the Chi-Squared value using the formula:

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} \tag{1}$$

In equation 1, χ^2 is the Chi-Squared test, O_i represents the observed frequency in each category i, and E_i denotes the expected frequency in each category i [10].

Using equation 1, we can compute the p-value for the Chi-Squared distribution with degrees of freedom [26]. If the resulting p-value is lower than or equal to a predetermined threshold, it is appropriate to reject the null hypothesis. This rejection indicates a statistically significant relationship between the independent and dependent variables [12].

Lastly, as shown in equation 1, the Chi-Squared test is based on frequency counts. Intuitively, counts cannot be negative. Therefore, this method cannot handle negative values.

3.2 ANOVA

The analysis of variance (ANOVA) is a statistical method employed to systematically compare the means of two or more independent groups by assessing their variability within and between them [25]. It is a widely utilised technique in statistical analysis, typically requiring the independent variable to be categorical and the dependent variables to be continuous in nature [27]. The null hypothesis in ANOVA states that there is no significant difference in means among the groups being investigated [9]. The hypothesis test in ANOVA uses an F-test and assumes that the data is sampled from a population that follows a normal distribution [24].

We can calculate the F-statistic as follows:

$$F = \frac{MSB}{MSE} \tag{2}$$

In equation 2, F is the F-statistic of the ANOVA test, MSB denotes the mean of squares between groups, and MSE denotes the mean of squares within groups [25]. We can reject the null hypothesis if the resulting F-statistic is greater than the critical value obtained from an F-distribution with the degrees of freedom. This rejection indicates that there is a significant difference in means between the groups under study [25].

3.3 Forward Selection and Backward Elimination

In consideration of the exponential nature of the search space in wrapper feature selection, this study adopts a sequential strategy using Forward Selection and Backward Elimination variants [11]. These variants involve iteratively adding or removing one feature at a time until we reach a predetermined number of features or the performance improvement ceases [11].

Nonetheless, a notable limitation of these sequential methods is their inability to reassess the significance of previously selected features after adding new features or the potential relevance of removed features after eliminating others [29]. This trade-off is considered acceptable for this research, as we expect the sequential feature selection techniques to yield satisfactory results within a relatively short time frame compared to exhaustive alternatives.

This study utilises, as estimators, logistic regression and linear regression for Forward Selection and Backward Elimination to reduce the number of dependent variables. Logistic regression is employed for classification problems, while we use linear regression for regression problems. As a result of utilising a limited number of estimators, we can conduct the comparative study more effectively, allowing for a focused analysis of the feature selection techniques.

3.4 Machine Learning Models for Evaluating the Performance of Feature Selection Techniques

The machine learning models used in this study are simple decision trees, linear regression, logistic regression, and support vector machines. They were selected due to their widespread usage and extensive documentation, allowing focused investigation into the impact of different feature selection techniques for categorical and numerical data on their performance.

Simple Decision Trees

The simple decision tree algorithm employs an iterative procedure to partition a given dataset into distinct nodes [7]. Each node represents a subset of the original dataset, and specific criteria determine the membership of individual data points [7].

Different variations of simple decision trees, Gradient Boosting Machine (GBM), Random Forest (RF), and Extreme Gradient Boosting (XGB), are utilised because they are readily available in the Autogluon [6] package, which we use to evaluate their performance. Additionally, Autogluon provides a convenient mechanism for hyperparameter tuning, making the experimentation process more streamlined.

Linear Regression

Linear regression is a statistical analysis technique that explores and models linear associations between the predictor(s) and the response variable [16]. The predictors follow a normal distribution [31] and can encompass both numerical and categorical variables [23].

Additionally, Autogluon [6] also provides a built-in model for Linear Regression (LR) and offers convenient features for hyperparameter tuning.

Logistic Regression

Logistic regression is a statistical technique employed to model the relationship between one or more predictor variables and a response variable, typically categorical, that can take two or more possible values [4]. This technique estimates the probability of the response variable belonging to a particular category based on the values of the predictor variables [4]. The predictors encompass numerical and categorical variables [23], and they do not need to adhere to a normal distribution [31].

Furthermore, Autogluon [6] provides a built-in model for Logistic Regression (LR) and offers convenient features for hyperparameter tuning.

Support Vector Machine

Support vector machines (SVMs) are powerful machine learning algorithms that classify data points into different classes by identifying boundaries that separate the data points effectively [19]. The main objective of SVMs is to find the hyperplane which maximises the margin between the closest data points of different classes, effectively creating a decision boundary [19]. The data points resting adjacent to the decision boundary, known as support vectors, play a crucial role in defining the hyperplane [19].

Unfortunately, Autogluon [6] does not provide a built-in model for SVM. Therefore, we utilise the Support Vector Classification and Support Vector Regression variants from the sklearn package [18]. These implementations are widely used and offer reliable performance. Furthermore, the sklearn package also includes a model for conducting grid search, allowing for efficient hyperparameter tuning for SVMs.

4 Methodology

It is crucial to gather datasets that are both large in size and diverse in nature to ensure a thorough comparative analysis between filter and wrapper methods. However, these datasets often require preprocessing techniques to address their unprocessed state and ensure optimal performance. Therefore, this section focuses on describing the unique characteristics of each dataset, including assumptions about the expected performance of each method when applied to these datasets, as well as the necessary transformations needed for their implementation.

Within the experimental framework, we consider several independent variables. These variables encompass distinct actions performed on the datasets, including:

1. Performing no alteration to the datasets.
2. Applying preprocessing techniques on the datasets.
3. Removing features that do not align with the expected type for each feature selection technique.
4. Partitioning the datasets into categorical, discrete, and continuous subsets of features.

Conversely, the dependent variable under examination is as follows:

- The percentage of selected features indicating the proportion of the ones chosen by the feature selection techniques.

4.1 Datasets

We collected the datasets from reputable repositories such as the UCI Machine Learning Repository [5], Kaggle[1], and OpenML [28]. Factors considered for dataset selection included instance count, feature count, and feature type (categorical or numerical). The selection process also accounts for the machine learning tasks applicable to each dataset.

A fundamental aspect of the nature of the datasets is that categorical data can be classified as ordinal (e.g., Likert scale) or nominal (e.g., gender), while numerical data can be discrete (e.g., integers) or continuous (e.g., floating-point numbers) [1]. These distinctions are important as they may impact the performance of different feature selection techniques.

For in-depth details regarding these datasets, including their characteristics and attributes, we encourage interested readers to visit Appendix A and the dedicated GitHub repository[2] created for this research project.

4.2 Hypotheses

We anticipate the Chi-Squared test to exhibit superior performance when applied to categorical or discrete data rather than numerical data.

On the other hand, we expect the ANOVA test to yield more reliable outcomes when employed on continuous data, as opposed to discrete or categorical data.

Regarding the wrapper feature selection techniques, both Forward Selection and Backward Elimination are versatile and accommodating for numerical and categorical data. However, we expect numerical data to yield more favourable results, given that it often provides richer information and a higher potential for meaningful feature selection.

In general, we expect wrapper methods to outperform filter methods. The reason for this is that wrapper methods can incorporate more computationally complex operations to assess the importance of each feature.

[1] https://www.kaggle.com/

[2] https://github.com/delftdata/bsc_research_project_q4_2023/blob/filter_vs_wrapper_methods/datasets/datasets_summary.md

4.3 Data preprocessing

In this study, different methods require varying degrees of data preprocessing. The following table presents a summary of the preprocessing techniques applicable to the datasets based on the chosen feature selection technique:

Table 1: Preprocessing applied to feature selection techniques

Technique	Misc*	Min-Max	Binning	Normalization
Chi-Squared	X	X	X	
ANOVA	X			X
Forward Selection	X			X
Backward Elimination	X			X

For more detailed information about the preprocessing techniques utilised in this study, please refer to the Appendix B.

5 Evaluation

This section details the employed metrics, the experimental goals, the implemented experimental setup, and the obtained results. It emphasizes the main contribution to the ongoing discourse surrounding feature selection.

5.1 Metrics

Various metrics provide an objective assessment of the performance exhibited by each of the feature selection techniques under scrutiny. The selection of these metrics is primarily driven by their widespread usage and acceptance within the field, ensuring consistency and reliability across evaluations. For classification problems, we deem the metric of accuracy, while for regression tasks, we choose the root mean squared error (RMSE) as the indicative criterion.

The experimentation involves varying the percentage of selected features from 0 to 100 to facilitate a meaningful comparison among the methods. This range was divided into intervals of 10, ensuring a step-wise progression that allows for a systematic evaluation of the techniques.

Additionally, we take into account the runtime of each feature selection technique. It is well-known that wrapper methods typically require more computing power than filter methods. Therefore, we consider carefully the runtime aspect to evaluate the trade-off between computational costs and performance improvement.

5.2 Experiment goals

The primary objective of these experiments is to validate the hypotheses presented in Section 4.2. By substantiating these hypotheses, we address the research question, *How do different feature selection techniques for categorical and numerical data impact the performance of simple decision trees, linear machine learning algorithms, and support vector machines?*. The outcomes of the experiments provide a clear insight into the influence of filter and wrapper methods on the performance of the machine learning models mentioned in Section 3.4, enabling an understanding of their respective efficacy and suitability for handling categorical and numerical data.

Furthermore, we designed these hypotheses mainly to provide guidance and direction rather than being seen as rigid goals. Thus, disproving certain suppositions can still yield valuable insights that contribute to addressing the research question at hand.

5.3 Experimental setup

We dedicate this section to introducing the experimental setup. The software that we implemented for this research project, along with instructions on how to use it, can be easily accessed within the dedicated folder named ”filter_vs_wrapper_methods” in the publicly available GitHub repository[3]. We created this repository to validate the hypotheses of this study and provide transparency and convenient access to the essential tools and accompanying guidelines.

Experiment 1: Limitations of Feature Selection Techniques

The primary aim of this experiment is to document and examine the inherent constraints and limitations associated with the Chi-Squared, ANOVA, Forward Selection, and Backward Elimination feature selection techniques.

The experimental condition involves utilising the data in its raw state. This simple yet significant independent variable offers a compelling approach to gaining deep insights into the ability of the methods to handle missing values if such handling is feasible. Furthermore, it provides a valuable opportunity to evaluate the compatibility of the techniques with a wide range of feature types.

Experiment 2: Comparison between Filter and Wrapper Methods using Preprocessing

We design experiment 2 to be more practical and conclusive, as it ensures that all methods run successfully under experimental conditions.

The experimental conditions for this experiment involve the preprocessing step described in Appendix B. This crucial step guarantees the successful execution of each feature selection method. Furthermore, experiment 1 provides valuable insights that inform the necessary preprocessing actions required to ensure the proper functioning of the techniques.

In the context of this experiment, we can consider several experimental conditions, namely the inclusion of mean, median, or mode imputation strategy and the application of normalization. We expect that varying these experimental conditions will significantly impact the subset of features selected by each method. However, even observing no impact can be an important finding, as it would demonstrate the versatility of the techniques.

Imputation strategies depend on the type of features. We can employ for continuous data all imputations such as mean, median, mode (or most frequent), and constant imputation. Only the last two strategies are meaningful for categorical data, as mean and median calculations do not apply to this data type.

[3] https://github.com/delftdata/bsc_research_project_q4_2023

We expect the omission of normalization to have a particularly adverse effect on ANOVA for classification and regression tasks. Similarly, the wrapper methods may be adversely affected by the lack of normalization when applied to regression tasks, as these methods rely on linear regression as the underlying estimator.

Experiment 3: Comparison between Filter and Wrapper Methods using Limited Preprocessing

Experiment 3 represents a trade-off between experiment 1 and experiment 2, aiming to strike a balance between guaranteeing the successful application of feature selection methods and preserving the datasets in their unaltered, raw form. Subsets of features are carefully selected to align with the expected data types for each feature selection method. Furthermore, to handle missing values, we eliminate the dataset rows containing at least one missing value.

However, a drawback of this trade-off is that meaningful comparisons between the feature selection methods are hard to accomplish, as they operate on different subsets of the original datasets within this experimental setup. The only viable comparison lies in observing the effect of applying or not applying feature selection rather than directly comparing the performance of different methods.

Experiment 4: Comparison between Filter and Wrapper Methods using Preprocessing and Data Type Partitioning

The experimental setup of experiment 4 involves preprocessing techniques and three distinct partitions: categorical, continuous, and discrete. The first partition encompasses both ordinal and nominal features, as they are treated similarly by the feature selection techniques. The second partition comprises features encoded as float64[4], representing continuous numerical data. The last one consists of features encoded as int64[5], representing discrete numerical values.

With its emphasis on categorizing and evaluating feature selection methods based on the type of features, we anticipate experiment 4 to yield highly informative results. This approach eliminates the confounding effects of mixing categorical and numerical features, allowing for a focused evaluation of which feature selection methods are most effective for handling either categorical or numerical features.

5.4 Results

In this subsection, we present a comprehensive analysis of the results from the previously-mentioned experiments. The focus is on highlighting the most important findings.

During the preprocessing stage, unless specified otherwise, we apply default techniques based on the nature of the data. We use mean imputation to handle missing values for continuous data, while for categorical and discrete data, we apply mode imputation. Additionally, we perform normalization for ANOVA and wrapper methods.

For more details about the results, please visit the GitHub repository[6].

Experiment 1: Limitations of Feature Selection Techniques

In experiment 1, the results primarily focused on identifying the limitations of each method. These limitations serve as a foundational understanding upon which we build the subsequent experiments. The following are the most important findings:

- None of the methods are capable of handling missing values.
- All methods are unable to process strings.

The findings from experiment 1 played a crucial role in shaping the preprocessing techniques employed in experiment 2 and experiment 4. We specifically designed these techniques to address the limitations identified in experiment 1.

Experiment 2: Comparison between Filter and Wrapper Methods using Preprocessing

The main finding of this experiment is that overall filter methods outperform wrapper methods regarding the performance of the selected features on the considered machine learning models. However, we sometimes need to factor in the runtime of the techniques to come to this conclusion.

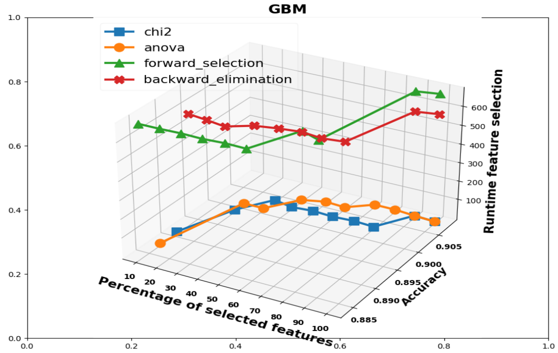

Figure 1: Accuracy of Gradient Boosting Machine for the bank marketing dataset

Bank Marketing

This dataset consists of numerous categorical and discrete features. Surprisingly, the ANOVA test stands out by efficiently selecting a subset of features that yields mostly the best classification accuracy across all considered machine learning models.

These findings contradict the hypothesis that the Chi-Squared test would perform better due to the proportion of categorical features being higher than that of the numerical ones. The significant factor behind this observation is likely the statistical relevance of differences between means.

Interestingly, despite their computational requirements, Backward Elimination and Forward Selection underperformed compared to filter methods.

[4] https://numpy.org/doc/stable/user/basics.types.html

[5] https://doc.embedded-wizard.de/int-type

[6] https://github.com/delftdata/bsc_research_project_q4_2023

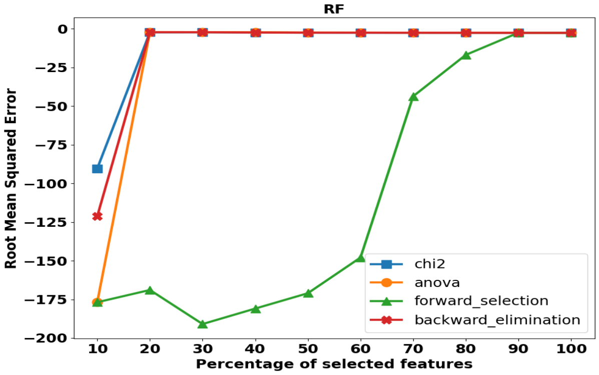

Figure 2: Root Mean Squared Error of Random Forest for the bike sharing dataset

Bike Sharing

This dataset consistently demonstrates similar results across all models, with Chi-Squared, ANOVA, and Backward Elimination yielding the same root mean squared error. On the other hand, Forward Selection consistently performs the worst in terms of root mean squared error.

The dataset primarily consists of discrete features, followed by continuous features, and a single categorical one. Filter techniques surpassed wrapper methods in terms of performance and execution time.

We can attribute the inferior performance of Forward Selection to robust correlations among certain features within the *Bike Sharing dataset*. Since Forward Selection selects one feature at a time, starting from an empty set of features, it does not account for these correlations, leading to suboptimal results.

Breast Cancer

In this dataset, the features selected by ANOVA and Forward Selection consistently demonstrate superior classification accuracy compared to the ones chosen by Chi-Squared and Backward Elimination across all considered machine learning models.

This dataset predominantly contains continuous features, with only one discrete and irrelevant one (the unique sample id). The relatively poorer performance of Chi-Squared is due to the conversion of floating-point data to discrete or categorical values, which may not fully preserve the underlying information, leading to a loss of accuracy. On the other hand, Backward Elimination cannot account for strongly correlated features after discarding them, resulting in inferior performance.

Census Income

In the case of the *Census Income dataset*, where categorical and discrete features are prevalent, Chi-Squared emerges as the most effective feature selection method, yielding subsets of features that result in the highest classification accuracy. ANOVA and Forward Selection also perform well, closely following Chi-Squared.

The poor performance of Backward Elimination is due to the elimination of features one by one based on their relevance. This approach may overlook the combined feature relevance, where features that may not appear individually relevant become significant when jointly considered.

Housing Prices

For this dataset, no particular feature selection method consistently outperforms the others. However, when considering the runtime a crucial factor, ANOVA appears to be a favourable choice due to its lower computational cost.

The dataset primarily comprises categorical features, discrete features, and a small number of continuous ones. In this context, the counterintuitive result of Chi-Squared yielding the worst performance is due to the importance of discrete values that closely resemble numerical data rather than categorical data.

Nasa Numeric

For the *Nasa Numeric dataset*, which primarily consists of categorical features with a small number of continuous ones, considering the runtime as a significant factor leads to the conclusion that we should prefer filter methods over wrapper methods.

Additionally, Chi-Squared lags in some cases, contradicting the hypothesis that categorical data should be more suitable for this feature selection technique. One possible explanation for this could be the small sample size of the dataset, which limits the ability to detect statistical significance accurately.

Steel Plates Faults

In the case of the *Steel Plates Faults dataset*, Backward Elimination consistently achieves the highest classification accuracy, closely followed by ANOVA. Furthermore, Chi-Squared outperforms Forward Selection.

This dataset consists of a majority of discrete features and a significant number of continuous ones. Contrary to expectations, Chi-Squared does not yield the best results, meaning that the floating-point features are more weighty, even though they are the minority or the discrete ones closely resemble numerical data.

While Backward Elimination achieves the best performance in terms of classification accuracy, it comes at a higher computational cost. Therefore, we can consider ANOVA more appropriate due to its low computational requirements and close performance to Backward Elimination.

Lastly, the poor performance of Forward Selection is on account of selecting features based on their relevance without considering their combined correlation with the target variable.

Variations of Experimental Conditions

In addition to the default mean imputation and normalization, we consider two more situations in the experiment: median imputation with normalization and mean imputation without normalization.

For the former, Chi-Squared performs better when selecting 10% of the features. However, apart from this difference, the results are similar to using mean imputation and normalization. We do not consider this setup in the subsequent experiments due to the similarity in performance and to maintain consistency.

In the case of mean imputation without normalization, Forward Selection performs better, while ANOVA does worse. This finding is counterintuitive to the hypothesis that Forward Selection should perform worse without normalization. Despite this unexpected result, the impact of not employing normalization is not considered significant enough for further exploration. The reason for this decision is the complexity introduced by the absence of normalization and the fact that Backward Elimination is not affected, even though it uses identical estimators as Forward Selection. Additionally, ANOVA, with and without normalization, still outperforms Forward Selection.

Experiment 3: Comparison between Filter and Wrapper Methods using Limited Preprocessing

The aim of experiment 3 is to showcase the potential performance improvement achieved through feature selection when the dataset structure aligns with the expected feature type for each method. To achieve this, columns with misalignments are removed from the dataset. Additionally, we remove rows containing missing values to minimize dataset alterations.

We should interpret with caution the apparent superiority of wrapper methods over filter methods in this experiment. The performance improvement of wrapper methods is owing to their ability to retain more features rather than indicating their overall superiority in feature selection. As mentioned earlier, experiment 3 does not objectively compare different feature selection methods. Instead, it compares the impact of applying feature selection methods versus not applying them. The results demonstrate that without performing significant alterations that may introduce confounding factors, using feature selection techniques can also lead to smaller datasets that retain features with at least the same information as the original datasets.

Experiment 4: Comparison between Filter and Wrapper Methods using Preprocessing and Data Type Partitioning

The main finding of this experiment indicates that, on the whole, filter methods exhibit better performance in feature selection compared to wrapper methods. However, it is crucial to consider the runtime of the techniques in specific cases to reach this conclusion.

Categorical partition

When considering datasets consisting solely of categorical data, both Chi-Squared and ANOVA consistently outperform wrapper methods in terms of accuracy and root mean squared error.

Moreover, ANOVA generally outperforms Chi-Squared in the categorical partition of the datasets suggesting that the difference in means within and between groups may be more appropriate in capturing the relationships between categorical features and the target variable, compared to the frequency-based approach of Chi-Squared.

Continuous partition

When dealing with datasets consisting of continuous data, the available options for feature selection are ANOVA, Forward Selection, and Backward Elimination. This partition successfully validates all of the initial hypotheses of this study.

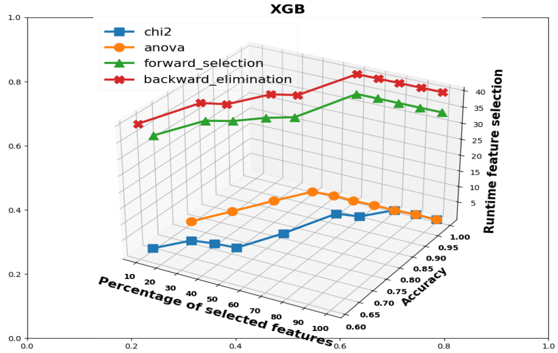

Figure 3: Accuracy of Extreme Gradient Boosting for the steel plates faults discrete subset

Given the potential limitations of wrapper methods in terms of their computational costs, ANOVA appears to be a preferred choice for feature selection in this context.

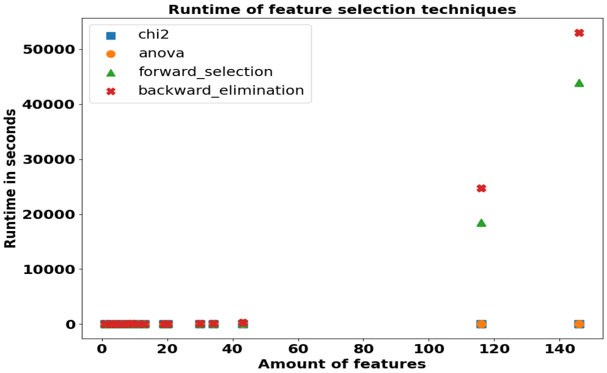

Figure 4: Runtime of feature selection techniques with respect to the number of features of each dataset

Discrete partition

The analysis of discrete subsets of datasets indicates a preference for filter techniques over wrapper techniques in most cases. However, there are certain edge cases where wrapper methods may outperform filter methods. These scenarios typically arise when the subset of discrete data is inadequate to accurately predict the outcome, resulting in low accuracy or high negative values of root mean squared error. Furthermore, despite the potential for performance improvement in some cases, the computational costs associated with wrapper techniques do not justify their adoption.

The suitability of Chi-Squared for feature selection increases as the degree to which the discrete data resembles categorical data rises with respect to the analysis of the *bike sharing dataset*, where the discrete features closely resemble categorical data (e.g., weekday). In this dataset, Chi-Squared outperforms ANOVA, indicating its effectiveness in selecting relevant features. On the other hand, for the *steel plates* *faults dataset*, the discrete features exhibits more of a numerical nature rather than resembling categorical data. As a result, ANOVA outperforms Chi-Squared in this case.

We cannot objectively quantify the degree to which discrete data resembles categorical data, and therefore both Chi-Squared and ANOVA should be executed to determine which method is more suitable for a given dataset containing discrete features.

6 Discussion and Limitations

Previous research studies focusing on filter and wrapper methods for feature selection identified wrapper methods as more suitable for feature selection tasks. However, the main critique we can attribute to these works is their lack of reproducibility.

That is why the present research project takes a different approach by emphasizing reproducibility and ensuring facile validation of the results in future experiments. The experimental findings provide evidence that filter methods outperform wrapper methods. Although the performance improvement of wrapper methods is present in some cases, it does not justify the extensive computational power required.

Despite the efforts made to ensure the reliability and reproducibility of the results, it is crucial to acknowledge the potential limitation of dataset selection bias in the study. While we use the number of samples and features as the primary selection criteria, it is possible that the chosen datasets unintentionally favoured filter methods over wrapper methods.

Another limitation of this study is the deliberate selection of logistic regression for classification tasks and linear regression for regression tasks as the underlying estimators for Forward Selection and Backward Elimination. The purpose behind this choice is to limit the scope of exploration and facilitate a more targeted analysis, given the limited timeframe of the project. Nonetheless, it remains plausible that alternative underlying estimators could offer superior performance enhancements, thereby showcasing the superiority of wrapper methods.

Lastly, delving into such an unfamiliar topic was an exceptional learning experience, requiring substantial effort to comprehend the concepts involved. We devoted significant efforts to implementing algorithms that align with the specifications of each feature selection technique, aiming to provide transparent and reliable results. The study's primary goal was to contribute to understanding which feature selection method is preferable under certain circumstances, with the hope of expanding knowledge in this domain.

7 Conclusions and Future Work

We meticulously addressed the research question regarding how different feature selection techniques for categorical and numerical data influence the performance of simple decision trees, linear machine learning algorithms and support vector machines by conducting a comparative study between filter and wrapper feature selection methods.

The study's central finding reveals that the filter methods, Chi-Squared and ANOVA, outperform the wrapper methods, (Sequential) Forward Selection and (Sequential) Backward Elimination, regarding classification accuracy, regression root mean squared error and runtime. When wrapper techniques show better accuracy or root mean squared error performance, the improvement is insubstantial to justify the increased computational requirements. Thus, the results contradict the study's chief hypothesis, which stated that wrapper methods would yield superior performance.

The analysis of dataset structure in terms of categorical and numerical features shows that Chi-Squared is particularly suitable for categorical data, aligning with one of the study's hypotheses. Additionally, ANOVA performs even better than Chi-Squared for categorical data, which we did not initially anticipate. For numerical data, the study differentiated between continuous and discrete cases. As expected, ANOVA performs better for continuous data. In the case of discrete data, the results vary depending on the degree to which we can interpret the discrete values as categorical. If the discrete data closely resembles categorical data, Chi-Squared is preferable. Otherwise, ANOVA proves to be more effective. Nevertheless, we cannot quantify the precise degree of resemblance without executing both Chi-Squared and ANOVA feature selection techniques on the specific dataset.

Future research could emphasize expanding the collection and analysis of datasets to account for the study's limitations. This analysis would help mitigate any unmeant bias towards filter methods that may have been present in the current study. Furthermore, investigating alternative underlying estimators for wrapper methods, specifically Forward Selection and Backward Elimination, could provide valuable insights into potential performance enhancements and improvements of the trade-off between computational costs and benefits.

8 Responsible Research

The principal aim of this research project is to align with the principles set forth by FAIR data [20] by ensuring the open availability of all datasets employed in the course of conducting experiments. These datasets, along with the accompanying code utilised to assess the effectiveness of feature selection techniques, can be accessed through the designated GitHub repository[7].

This study primarily focuses on utilising data for performing operations and deriving conclusions based on objective metrics, such as classification accuracy and root mean squared error. While favourable outcomes are preferred, the experimental design was not biased towards obtaining positive results. For instance, we specifically devised the first experiment to identify limitations in employing the Chi-Squared, ANOVA, Forward Selection, and Backward Elimination feature selection methods.

The integrity and reliability of the utilised sources preclude any possibility of data fabrication or falsification. Even if such malpractices were employed, they would not impact the results of the experiments, as the aforementioned objective metrics remain unaffected.

Additionally, we execute data trimming to eliminate irrelevant or non-influential data rather than filtering out undesirable outcomes.

[7] https://github.com/delftdata/bsc_research_project_q4_2023

One who would like to validate the results of this research project can reproduce them to a significant extent. We provide detailed instructions in Section 5.3. Moreover, interested readers can easily replicate the experiments by cloning the GitHub repository and executing the main.py or plot.py functions within the filter_vs_wrapper_methods/src folder.

Finally, it is imperative to acknowledge that the chief motivation behind this research project is solely to contribute to the ongoing discourse in the feature selection field. There is no intention to derive personal profit from the findings. The study outcomes are considered public property, intended to be openly utilised or subject to scrutiny in the event of any potential misconduct.

9 Acknowledgements

I want to express my sincere appreciation to all those who have supported and guided me throughout this research. Their valuable support and oversight have been instrumental in the study's successful completion.

First and foremost, I want to express my gratitude to my supervisor, Andra Ionescu. Her expertise, guidance, and prompt responses were invaluable. I am grateful for how our weekly meetings were conducted, as they provided me with valuable insights and direction. Additionally, her enlightening feedback regarding my progress and paper significantly enhanced the quality and depth of this research.

I would also like to thank my responsible professor, Asterios Katsifodimos, for his expertise and pragmatic approach in providing feedback regarding my progress and paper. His constructive criticism and suggestions have greatly enhanced the quality and clarity of this research.

Lastly, I want to express my gratitude to my fiancée, Maria-Elena. Her love and encouragement throughout this research project have been a constant source of strength. Her support has helped me stay motivated, making this achievement possible.

References

[1] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[2] Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143:106839, 2020.

[3] S DeepaLakshmi and T Velmurugan. A comprehensive survey on filter approach to feature selection methods for high dimensional data.

[4] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[5] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[6] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[7] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *icml*, volume 99, pages 124–133, 1999.

[8] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[9] Michael H Herzog, Gregory Francis, Aaron Clarke, Michael H Herzog, Gregory Francis, and Aaron Clarke. Anova. *Understanding Statistics and Experimental Design: How to Not Lie with Statistics*, pages 67–82, 2019.

[10] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. *Restorative dentistry & endodontics*, 42(2):152–155, 2017.

[11] Vipin Kumar and Sonajharia Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[12] Juan Laborda and Seyong Ryoo. Feature selection in a credit scoring model. *Mathematics*, 9(7), 2021.

[13] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[14] Huan Liu. *Feature Selection*, pages 402–406. Springer US, Boston, MA, 2010.

[15] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[16] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[17] Shahla Nemati, Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee, and Mehdi Hosseinzadeh Aghdam. A novel aco–ga hybrid algorithm for feature selection in protein function prediction. *Expert systems with applications*, 36(10):12086–12094, 2009.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] Derek A Pisner and David M Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020.

[20] B RDA FAIR Data Maturity Model Working Group et al. Fair data maturity model: specification and guidelines. *Res. Data Alliance*, 10, 2020.

[21] Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, and María Tombilla-Sanromán. Filter methods for feature selection–a comparative study. *Lecture notes in computer science*, 4881:178–187, 2007.

[22] Erik Schaffernicht, Robert Kaltenhaeuser, Saurabh Shekhar Verma, and Horst-Michael Gross. On estimating mutual information for feature selection. In *Artificial Neural Networks–ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings, Part I 20*, pages 362–367. Springer, 2010.

[23] Astrid Schneider, Gerhard Hommel, and Maria Blettner. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 107(44):776, 2010.

[24] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

[25] Lisa Sullivan. Hypothesis testing - analysis of variance (anova).

[26] Yale University. Two-way tables and the chi-square test, 1997-1998.

[27] Los Angeles University of California. What statistical analysis should i use? statistical analyses using spss, 2021.

[28] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[29] Yap Bee Wah, Nurain Ibrahim, Hamzah Abdul Hamid, Shuzlina Abdul-Rahman, and Simon Fong. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, 26(1), 2018.

[30] Xuechuan Wang and Kuldip K Paliwal. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition*, 36(10):2429–2439, 2003.

[31] Andrew Worster, Jerome Fan, and Afisi Ismaila. Understanding linear and logistic regression analyses. *Canadian Journal of Emergency Medicine*, 9(2):111–113, 2007.

A Datasets

Table 2: Dataset information as recorded by experiment 4

Dataset name	#Instances	#Continuous	#Discrete	#Categorical
Arrhythmia	452	116	146	0
Bank Marketing	45211	0	7	9
Bike Sharing	17379	4	11	1
Breast cancer	569	30	1	0
Census Income	48842	0	6	8
Character Font Images	745000	1	408	0
Housing prices	1460	3	34	43
Internet Advertisements	3279	0	1554	4
Nasa Numeric	93	3	0	19
Steel plates faults	1941	13	20	0

B Preprocessing Techniques

Table 3: Preprocessing applied to feature selection techniques

Technique	Misc*	Min-Max	Binning	Normalization
Chi-Squared	X	X	X	
ANOVA	X			X
Forward Selection	X			X
Backward Elimination	X			X

In table 3, *Misc** encompasses multiple preprocessing steps that we can apply to all feature selection techniques:

- Handling missing values: We can use imputation strategies (e.g., mean, median, constant, or mode imputation) to fill in the missing values. Alternatively, we can remove rows containing missing values. However, we should exercise caution when removing rows, as it may lead to insufficient data for effective feature selection if the remaining rows are fewer than the columns, limiting the exploration of all possible feature values.
- Handling data encoded as strings: If the dataset contains string-encoded data, a simple approach is to apply a bijective mapping, where each unique string value is assigned a unique natural number. For instance, the feature ”marital status” with values ”single,” ”married,” and ”divorced” can be mapped to 0, 1, and 2, respectively.
- Removing constant features: We can safely remove features having the same value for all rows, as they provide no discriminatory power.

These are the general aspects of preprocessing applicable to all methods. We now consider the specific preprocessing requirements for each technique.

Chi-Squared

We can apply a min-max scaling transformation to address negative values by mapping the values within a specified range, the default being (0, 1). For handling floating-point numbers, we can group continuous data into bins of arbitrary size (e.g., using binning techniques).

ANOVA

ANOVA assumes that the data follows a normal distribution, and therefore, a normalization step is necessary to ensure reliable results.

Forward Selection and Backward Elimination

Linear regression is the underlying estimator for Forward Selection and Backward Elimination techniques in regression tasks. As linear regression assumes that the data follows a normal distribution, it is essential to perform a normalization step to ensure reliable and accurate results.