Filtering Knowledge: A Comparative Analysis of Information-Theoretical-Based Feature Selection Methods

Kiril Vasilev[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: Kiril Vasilev
Final project course: CSE3000 Research Project
Thesis committee: Asterios Katsifodimos, Andra Ionescu, Elvin Isufi

Abstract

The data used in machine learning algorithms strongly influences the algorithms' capabilities. Feature selection techniques can choose a set of columns that meet a certain learning goal. There is a wide variety of feature selection methods, however, the ones we cover in this comparative analysis are part of the information-theoretical-based family. We evaluate MIFS, MRMR, CIFE, and JMI using the machine learning algorithms Logistic Regression, XGBoost, and Support Vector Machines.

Multiple datasets with a variety of feature types are used during evaluation. We find that MIFS and MRMR are $2-4$ times faster than CIFE and JMI. MRMR and JMI choose columns that lead to significantly higher accuracy and lower root mean squared error earlier. The results we present here can help data scientists pick the right feature selection method depending on the datasets used.

1 Introduction

Machine learning algorithms depend highly on the data they are trained with. However, often only limited data is available and data scientists need to look into external sources that could potentially augment a given table, named further base table. A potential solution to this problem is the employment of dataset discovery systems that are capable of discovering semantically related tables [23]. Such an example is Auctus, which is a dataset search engine that enables users to find datasets that can be joined together with a base table [6].

Example. To present the challenges related to feature discovery, let us consider the well-known Titanic table and a set of tables on the topics of discrimination, crime data, and geothermal energy resources (shown in Figure 1). Although the topics look unrelated to the Titanic dataset, it may be possible to meaningfully join some of these datasets with the Titanic one to increase a learning model's performance.

We believe there is a meaningful connection between *DiscriminationCases* and *Titanic* tables based on a composite key made of *Age* and *Sex* columns. It is possible that individuals of a certain race or disability could have potentially been discriminated against during the evacuation of Titanic passengers. Moreover, we consider the relation between *Titanic* and *CrimeData* on the feature *Name* as semi-meaningful, due to the low probability of having passengers with crime records on Titanic. Finally, an example of an undesired join would be between *Titanic* and *GeothermalResources* based on columns *Parch* and *Class*, since there is no semantic connection between these features.

Challenges. All of this analysis, however, would require manually performing the joins and training models, which would take tremendous effort for data scientists. Moreover, not all features coming from other tables are useful, because they may bring unnecessary noise on which learning models overfit, or they might increase the runtime computation costs. To tackle this problem, feature selection algorithms can choose a subset of important features given a base table.

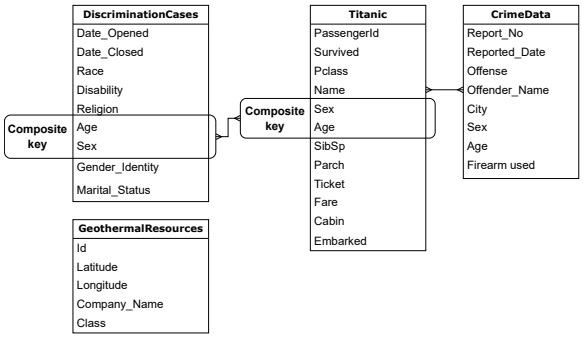

Figure 1: An example of an augmentation scenario for "Titanic" table

Nonetheless, choosing an appropriate filter selection method can be a challenge due to the abundance of approaches. We aim to compare widely used feature selection methods to help data scientists pick an appropriate one for their domain.

Research question. We focus on comparing the information-theoretical-based feature selection methods: Mutual Information Feature Selection (MIFS), Minimum Redundancy Maximum Relevance (MRMR), Conditional Infomax Feature Extraction (CIFE), and Joint Mutual Information (JMI), due to their applications in many domains as we show in Section 2. This paper answers the following research question:

How do the information-theoretical-based feature selection methods MIFS, MRMR, CIFE, and JMI compare in runtime and accuracy / RMSE for Decision trees, linear Machine Learning algorithms, and Support Vector Machines?

Although there are surveys that introduce feature selection methods like [4][20][28], no conclusion has been given on which of the methods is performing better than another and under which condition. The comparison there has mostly been made on the properties of the feature selection algorithms. Unfortunately, none of these surveys measures the runtime performance of the approaches or provides clear recommendations for use. Therefore, we plan to address this lack of knowledge in our research paper.

Contributions. In short, the main contributions can be summarised as follows:

- We display the importance of using complex entropy estimators on continuous features.
- We show that MIFS and MRMR have $2-4$ times lower runtime costs than CIFE and JMI.
- We extensively compare the methods to discover that the use of MRMR and JMI leads to models with higher effectiveness regardless of the type of features.
- We show that there are cases when relying on Information Gain feature selection is faster and more effective.

Outline. The paper builds upon the already existing literature discussed in Section 2. The notion of information theory and the four methods we compare are introduced in Section 3. Section 4 describes the steps we took to conduct this analysis. The experimental setup is presented in Section 5, which has led to our findings in Section 6. They are further discussed in Section 7. The limitations are mentioned in Section 8 and the main findings are concluded in Section 9.

2 Related work

Filter methods. Feature selection helps in improving a learning model's performance by "reducing irrelevant or redundant features" [15, p. 5]. The four methods evaluated in this paper are part of the filter methods family. Therefore, they focus only on the internal structure of the dataset and explore the relation between the features and a target column [7]. The methods have been previously outlined in surveys related to feature selection algorithms [4][20][28], however, there, they have only been introduced and briefly compared based on accuracy. Hence, we aim to introduce comparison not only on accuracy but also on root mean squared error.

Information theory methods. Each of the four approaches has found its applications in specific domains. For instance, MRMR is used in the field of biology [5][19], while CIFE was found applicable in image recognition [13][22]. In more recent papers MIFS is also part of a pipeline to diagnose Covid-19 based on X-ray images [26]. Nonetheless, JMI has been employed in urban arterial traffic detection in [21]. The justification there for using each method is not based on the properties of the datasets. We aim to investigate the relation between feature selection methods and types of features.

Comparative analyses have been created on information theory approaches. For instance, [12] evaluates JMI and MRMR in the domain of image classification. However, it does not do this extensively and does not include the methods MIFS and CIFE. JMI and MRMR are further compared in acoustic event detection tasks, where the former was found better performing [16]. Neither of the two analyses looks into the runtime of the methods. We believe this is worth investigating in our research for all four approaches. Moreover, not only on one learning algorithm but on multiple models.

Entropy estimators. Information theory methods depend on their entropy estimators. To address this, studies on estimating continuous entropy have been conducted in [11] and [30], where the learning error of different discretisation techniques is visualized. Apart from learning model capabilities, in our paper, we further compare the estimators based on runtime.

3 Preliminaries

In this section, we first introduce the notion of relevance and redundancy. Next, we show the main concepts of information theory and then we describe each of the four methods.

3.1 Relevance and Redundancy

The four information theory feature selection methods compared here aim to optimise the following criteria:

- **Relevance** measures the linear or non-linear relation between a random variable X_k (or a set of variables $\mathcal{S}$) and a class column Y. The higher the relevance, the more shared information there is between them [25]. Higher feature relevance is preferred by the four information theory feature selection approaches.
- **Redundancy** measures the amount of information a random variable X_k shares with the set of selected features $\mathcal{S}$. If a feature is too similar to already selected ones, then it does not bring sufficient new information and is considered redundant [22]. Therefore, the approaches aim to minimise this redundancy and bring features that share fewer characteristics with the ones in S.

3.2 Information theory concepts

Information-theoretical-based feature selection algorithms are built on top of a framework first introduced in Shannon's information theory paper [27]. To formally present the four methods, it is vital to briefly describe the main concepts of information theory on which these approaches rely.

Entropy is the first central concept and measures the uncertainty of a discrete random variable X with values $x_i \in X$ and probability density function $P(x_i)$ as shown [20]:

$$H(X) = -\sum_{x_i \in X} P(x_i) \cdot \log(P(x_i)) \tag{1}$$

Entropy can be thought of as a measurement of diversity in the values of a feature. The more diverse range of values there is, the higher the entropy of the feature would be [4]. Inversely, if a variable is highly biased towards certain values, the entropy would be low. To compute the entropy of a feature, one would need to estimate the density distributions $P(X_k)$ for the values of each feature X_k in a dataset. If the features are discrete, this is equivalent to computing their frequencies. If that is not the case, then one possible solution is the application of data discretisation techniques as a preprocessing step. For example, by using density estimation with Parzen windows, as suggested in earlier work in [18].

Conditional entropy is the second main concept and denotes the entropy of a random variable X given knowledge of another variable Y, where $P(x|y)$ represents the conditional probability of a value $x \in X$ on a value $y \in Y$:

$$H(X|Y) = -\sum_{y_i \in Y} P(y_i) \sum_{x_j \in X} P(x_j|y_i) \cdot \log P(x_j|y_i) \tag{2}$$

The conditional entropy would always be less or equal to the initial entropy unless the features X and Y are independent [2]. This stems from $P(X|Y) = P(X)$ for independent random variables X and Y.

Information gain (IG) is the combination of Equation 1 and Equation 2 for discrete random variables X and Y and is denoted by $I(X;Y)$. The information gain is also known as **mutual information** and is a value quantifying the dependence between random variables X and Y [13]:

$$I(X;Y) = H(X) - H(X|Y) \tag{3}$$

Through algebraic manipulations Equation 3 extends to the following formula:

$$I(X;Y) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \cdot \log \frac{P(x_i, y_j)}{P(x_i) \cdot P(y_j)} \tag{4}$$

If X and Y are independent, the information gain would be 0, since then $P(x_i, y_j) = P(x_i) \cdot P(y_j)$ and this results in calculating $I(X;Y) = 0$. Hence, $H(x|y) = H(x)$ and $H(y|x) = H(y)$, and the two features have no overlap. In addition, information gain satisfies a symmetry property. Namely, $I(X;Y) = I(Y;X)$.

Conditional information gain is another concept that two of the information-theoretical-based methods rely on. It measures the information gain between discrete random variables X and Y given the knowledge of a random variable Z [20]:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{5}$$

3.3 Information-theoretical-based feature selection methods

The four approaches select columns based on scores calculated with a function J, where at each iteration the feature with the highest score gets selected. All four of them share the same scoring function $J(X_k)$ [20]:

$$\begin{aligned} J(X_k) = I(X_k;Y) + \alpha \sum_{X_j \in \mathcal{S}} I(X_j;X_k|Y) \\ - \beta \sum_{X_j \in \mathcal{S}} I(X_j;X_k), \end{aligned} \tag{6}$$

where $\mathcal{S}$ denotes the set of currently selected features, X_k is a non-selected feature we are calculating the score J for, Y is the target column, and α and β are hyperparameters that influence the strength of the two components. Furthermore, when no features have been selected, $|\mathcal{S}| = 0$. Hence, the first selected feature by the methods depends solely on the information gain between feature X_k and class variable Y.

Although the four feature selection algorithms share the same scoring function, the values of α and β can significantly influence their efficiency and effectiveness.

Mutual Information Feature Selection

Mutual Information Feature Selection (MIFS) considers only feature relevance to a class variable and shared information between selected and unselected features [2]. As such it aims to maximise feature information and minimise selected features redundancy. This relation is shown in the following equation for a candidate feature X_k:

$$J(X_k) = I(X_k;Y) - \beta \sum_{X_j \in \mathcal{S}} I(X_j;X_k), \tag{7}$$

If the feature X_k is relevant to the target column Y and there is little relation between it and the selected features in $\mathcal{S}$, then it has a higher score J. Values for β between 0.5 and 1 are preferred for many classification tasks [2].

Minimum Redundancy Maximum Relevance

Minimum Redundancy Maximum Relevance (MRMR)[25] is a special case of MIFS, where β is defined as $\frac{1}{|\mathcal{S}|}$. Therefore, with more features added to $\mathcal{S}$, the influence of the β coefficient decreases:

$$J(X_k) = I(X_k;Y) - \frac{1}{|\mathcal{S}|} \sum_{X_j \in \mathcal{S}} I(X_j;X_k) \tag{8}$$

Conditional Infomax Feature Extraction

Conditional Infomax Feature Extraction (CIFE) aims to not only maximise feature relevance and minimise feature redundancy, but this approach also looks towards the joint information that can be collected between selected and non-selected features [22]. Thus, it explores how multiple features improve the class variable's information gain. Equal weight is placed on the feature information gain, the joint information gain, and the feature redundancy between the candidate feature and already selected features. If the information gain between X_k and the already selected features in $\mathcal{S}$ is too high, then X_k has a lower score J:

$$\begin{aligned} J(X_k) = I(X_k;Y) + \sum_{X_j \in \mathcal{S}} I(X_j;X_k|Y) \\ - \sum_{X_j \in \mathcal{S}} I(X_j;X_k) \end{aligned} \tag{9}$$

Joint Mutual Information

Joint Mutual Information (JMI) approach suggests that the feature redundancy and joint information gain between features should be inversely scaled by the size of the set of already selected features $\mathcal{S}$ [4]. This is a generalisation of the approach introduced in [32], where the focus is on visualising high-dimensional data as 2-D projections. Therefore, both α and β from Equation 6 are defined as $\frac{1}{|\mathcal{S}|}$:

$$\begin{aligned} J(X_k) = I(X_k;Y) + \frac{1}{|\mathcal{S}|} \sum_{X_j \in \mathcal{S}} I(X_j;X_k|Y) \\ - \frac{1}{|\mathcal{S}|} \sum_{X_j \in \mathcal{S}} I(X_j;X_k) \end{aligned} \tag{10}$$

4 Methodology

The previous section introduced the theoretical concepts behind the four feature selection methods. The steps undertaken to complete this study shall hereby be described. The choices of datasets, machine learning algorithms for evaluation, and metrics shall help us answer the paper's research question.

4.1 Datasets

To empirically compare the feature selection algorithms multiple datasets were carefully chosen. An outline of them is provided in Table 1, where we show the name of the dataset, the number of instances, the number of continuous/discrete/categorical features, and the machine learning task the dataset was designed for. The distributions of the features of the datasets were also inspected, such that we have columns with unimodal and multi-modal distributions. We present these value distributions in Appendix A.

The datasets have been collected from publicly available sources. We have three datasets with non-continuous features as *Gisette*, *Internet advertisements*, and *Census Income* and two datasets with only continuous features such as *Steel plates faults* and *Breast cancer*. Finally, *Housing prices* and *Bike sharing* contain a mix of all three feature types.

4.2 Machine learning algorithms

Due to the variety of datasets chosen, we used both classification and regression machine learning algorithms to measure the performance of the feature selection methods. We relied on linear and non-linear models: Logistic Regression, XGBoost, and Support Vector Machines. We briefly describe the characteristics of each of the algorithms.

Table 1: Datasets used for comparative analysis and their characteristics

Dataset name	#Instances	#Features	#Continuous	#Discrete	#Categorical	Task type
Gisette	6000	5000	0	5000	0	Binary classification
Internet advertisements	3279	1558	1	2	1555	Binary classification
Housing prices	1460	80	33	3	44	Regression
Bike sharing	17379	16	7	8	1	Regression
Census Income	32560	14	0	6	8	Binary classification
Steel plates faults	1941	33	33	0	0	Binary classification
Breast cancer	569	31	31	0	0	Binary classification

Logistic Regression is the linear model we have chosen. It aims to find a function, which accurately approximates the probability of a label [24]. The model can find the relation between dependent and independent variables in a classification problem. An example of a logistic function is the sigmoid function $\sigma(h(x)) = \frac{1}{1+e^{-h(x)}}$, where $h(x)$ is the function we are trying to approximate, and x is an input feature vector.

XGBoost is the Decision tree implementation that we rely on. XGBoost is a system that provides an implementation of tree boosting models [8]. It is an ensemble learning algorithm similar to a random forest that combines multiple decision trees to provide a more accurate prediction. An advantage of XGBoost compared to other Gradient Boosting Tree (GBT) solutions is that it can build trees in parallel, which provides scalability to learning models [8].

Support Vector Machines aim to maximize the margin between classes [3]. If we consider a space with linearly separable classes, it means that we consider values on the one side of the line to belong to one class, while on the other side to another. If the hyperspace between classes is not linearly separable, a different kernel function has to be used. For example, the Radial Basis Function [3].

4.3 Metrics

We have evaluated the four feature selection methods based on efficiency and effectiveness. In the context of this paper, we refer to these terms as follows:

Efficiency shall measure the runtime of the algorithms. Entropy estimation can be a time-consuming process and as such it was important to compare the methods on time spent computing. Therefore, we measured the runtime of feature selection algorithms in seconds. We considered methods with lower runtime duration to score better in terms of efficiency.

Effectiveness shall measure the learning performance of machine learning algorithms trained on the resulting datasets after feature selection has been performed. We measured accuracy for classification tasks and estimated root mean squared error (RMSE) for regression tasks. Feature selection algorithms that create models with high accuracy and low RMSE were considered to have higher effectiveness in achieving their tasks. We now briefly introduce the two metrics.

- **Accuracy** measures the percentage of correctly classified target columns of unseen data. Although simplistic, it is commonly used as an evaluation metric in machine learning algorithms due to its ease of interpretability. Accuracy is bounded between 0% and 100% and generally higher accuracy is preferred.
- **RMSE** was used when provided with a regression task to estimate performance. RMSE can be defined using the following formula $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y_i})^2}$, where y_i denotes the actual value and $\hat{y_i}$ denotes the predicted value by the machine learning algorithm. Therefore, the lower the RMSE, the better the model performs. The range of values for RMSE is between 0 and ∞.

5 Experimental setup

In this section, we present the experimental setup used to answer the research question of this paper. The four methods have been compared against multiple datasets with several machine learning algorithms. Since all four of them have a common term in their scoring function - the information gain (IG), we have made use of it as a baseline feature selection algorithm in the comparison. IG can be seen as a special case of the general feature selection function introduced in section 3.3 with $\alpha = 0$ and $\beta = 0$.

5.1 Missing data and encoding data

Some of the datasets used contain missing values, which meant that they require data preprocessing. Simply dropping data entries with missing values is dangerous because it can introduce a class imbalance. Class imbalance occurs when the target column values are highly skewed towards only certain values (or a single class) [14]. This can lead to poor learning model performance because algorithms would focus on learning the majority class and disregard the minority one.

We relied on an automated machine learning (AutoML) framework called AutoGluon [10] to train and evaluate models. The framework provided several benefits such as automatic hyperparameter tuning, imputation of missing values, and encoding of non-numeric features with only a few lines of code. Therefore, using AutoGluon eased the evaluation process and allowed us to focus on comparing the feature selection methods directly. For Support Vector Machines, all features were scaled between -1 and 1 and missing entries were replaced with 0.

5.2 Entropy estimation

To reiterate, the four information-theoretical-based feature selection methods depend highly on the entropy estimators used. Two discretisation approaches were evaluated. The first is a frequency-based approach that measures the number of occurrences of feature values, referred to as a simple approach. The second one is a k-nearest neighbour (KNN) entropy estimator, which is suitable for continuous features in datasets [17], referred to as a complex entropy estimator.

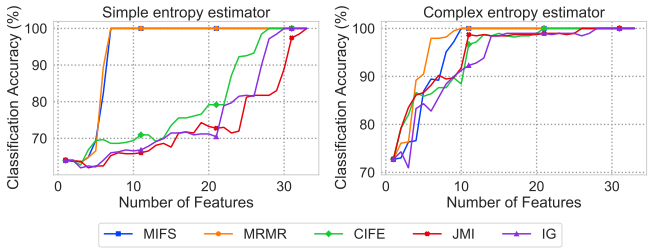

Figure 2: Accuracy comparison of entropy estimators on Steel plate faults dataset with Logistic Regression

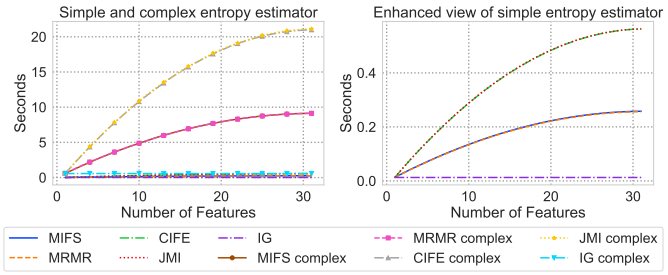

Figure 3: Runtime comparison of entropy estimators on Breast cancer dataset

5.3 Implementations and configuration

In Section 3.3 we showed the main similarity between the four feature selection methods. Since they all have a similar structure, they share the same implementation, too.

A limitation of accurate continuous entropy estimation is its computation costs. To tackle this problem, a smart approach (implemented in [20]) was used to memorise intermediate relevance and redundancy summations and as such reduce the necessity to recalculate values several times. For each of the candidate features X_k, we store their relevance and redundancy sums and update them until these features themselves are selected. This way, it is not needed to recalculate the entire relations, but only the mutual information and conditional mutual information between candidate features and the newly selected columns, which are added to $\mathcal{S}$.

To provide consistent results, the experiments have been computed on the same machine: AMD EPYC 7413 2.65GHz 24-Core Processor and all datasets fit in memory. During the evaluation, the datasets were initially split into an 80% training set and a 20% test set. The feature selection algorithms have been run on the training set and were later evaluated on the test set using AutoGluon. The machine learning models' hyperparameter tuning was performed on the entire dataset with AutoGluon. Moreover, the random state has been set to a specific seed (42) before each feature selection over the datasets. This was necessary due to the generation of random values during the entropy estimation process.

6 Results

Our main findings have been grouped based on the type of dataset used, which allowed us to control for the potential different behaviour of feature selection methods on the type of features. This enabled us to see whether there is a particular trend in the specific feature types used: continuous, discrete, and categorical. Since we are not evaluating machine learning algorithms against each other the focus on the results is more towards the efficiency and effectiveness of the feature selection methods themselves.

Firstly, we make a short comparison between the use of a simple and a complex entropy estimator. We observed that if we are dealing with continuous features the information-theoretical-based methods are dependent on the entropy estimator and the complex one performs 30% better in terms of effectiveness but has $50 - 100$ times higher runtime costs. Secondly, we present a brief comparison of the runtime costs (efficiency) of the approaches. Namely, we have discovered that MIFS and MRMR are $2 - 4$ times faster than CIFE and JMI because the former do not need to estimate conditional information gain. Next, we show the effectiveness of the approaches against the algorithms and datasets introduced in Section 4. A commonly occurring pattern is that IG, MRMR, and JMI are performing significantly better in terms of accuracy and RMSE regardless of the type of dataset and the machine learning task used. Finally, we compare different values for hyperparameter β in MIFS and show that IG can also be a viable feature selection method.

6.1 Entropy estimation

Entropy estimation plays an important role in the effectiveness of the four feature selection methods.

Let us consider the Logistic Regression model on the *Steel plate faults* dataset in Figure 2. Using the complex entropy estimator allows the methods to select features with high effectiveness quicker. For instance, instead of selecting 30 features with CIFE or JMI, or IG to reach 100% accuracy, we can select just $10 - 15$ columns with their complex entropy implementation. Furthermore, with simple entropy estimation, the methods CIFE, JMI, and IG underperform with nearly 30% at 20 selected features (as seen in the left plot of Figure 2), due to the estimator's limitations in approximating conditional information gain. Therefore, complex entropy estimation should be used for datasets with continuous features.

6.2 Runtime costs

To measure the runtime costs, the four methods were tasked with selecting columns one by one until all columns have been selected. We have run this experiment on all datasets from Table 1 except for *Gisette* and *Internet advertisements* due to the computation costs. Moreover, we have also added IG to our analysis. The results found are shown in Table 2. The complex entropy estimator is $50 - 150$ times slower than the simple one. This runtime difference becomes apparent very early in the feature selection process as shown in the left plot in Figure 3. There, we visualise how long it takes to select a certain number of features. Choosing the first feature is 40 times more expensive when the complex entropy estimator is used due to the complexity of calculating KNN.

In the right plot of Figure 3, we show a comparison between the feature selection algorithms' runtimes using a simple entropy estimator. MIFS and MRMR take roughly the same time. This also applies to the pair CIFE and JMI. To reiterate Section 3.3, what the first two methods have in com-

Table 2: Comparison of runtime costs (in seconds) for the four feature selection algorithms and information gain

	Simple entropy estimator					Complex entropy estimator				
Dataset name	**IG**	**MIFS**	**MRMR**	**CIFE**	**JMI**	**IG**	**MIFS**	**MRMR**	**CIFE**	**JMI**
Housing prices	**0.13**	2.41	2.40	8.04	8.10	**4.66**	179.41	183.25	368.85	369.09
Bike sharing	**0.30**	1.80	1.83	5.29	5.31	**18.08**	163.92	163.98	302.66	321.80
Census Income	**0.29**	1.69	1.71	4.28	4.25	**25.28**	232.02	231.81	639.27	638.78
Steel plates faults	**0.08**	0.71	0.71	1.70	1.70	**2.40**	40.72	40.94	100.29	100.40
Breast Cancer	**0.02**	0.26	0.26	0.56	0.56	**0.57**	9.13	9.13	20.99	21.12

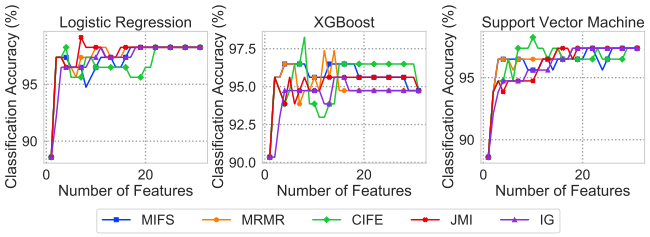

Figure 4: Accuracy of methods on Breast cancer dataset

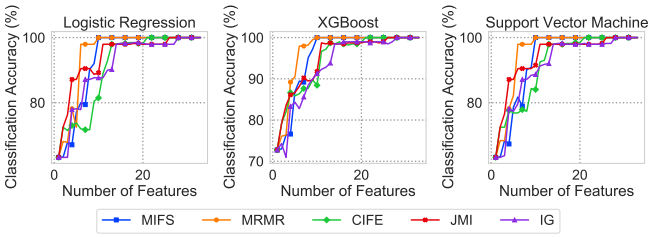

Figure 5: Accuracy of methods on Steel plates faults dataset

mon is that $\beta = 0$ in the general scoring function J. Therefore, they have fewer entropy estimations to perform and are $2-4$ times faster than CIFE and JMI. However, MIFS and MRMR are $6-40$ times slower than IG. This makes IG the fastest feature selection approach of all five.

6.3 Effectiveness

After showing the importance of deciding on an entropy estimator and considering the efficiency of the feature selection methods, we now focus on their effectiveness in terms of accuracy and RMSE. We analyse how the four methods perform against each other and against their baseline IG method on multiple datasets with varying feature types.

Datasets with continuous features

Our first measurement of effectiveness is on the datasets *Breast cancer* and *Steel plates faults*.

The results of the *Breast cancer* dataset in Figure 4 identify how the five approaches select features. All five would pick the same first feature since they only estimate the information gain between the target column and all candidates. Afterwards, the approaches would split into three groups in selecting the second feature - those which estimate conditional information gain (CIFE and JMI), those that do not, but estimate the redundancy between target and candidate features (MIFS and MRMR), and information gain (IG). The three machine learning algorithms show a noisy performance based on the subset of selected features. For the Logistic Regression model, MRMR and JMI seem to have $1-2\%$ higher accuracy than the rest, while for XGBoost, MIFS and CIFE have

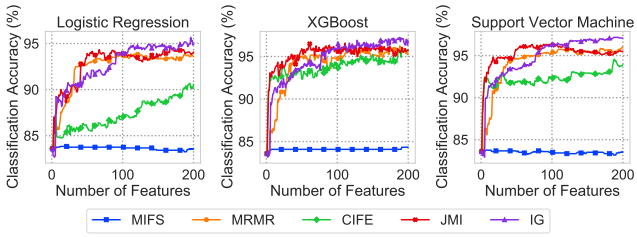

Figure 6: Accuracy of methods on Gisette dataset

$1-2\%$ better effectiveness. In terms of Support Vector Machines, MRMR, JMI, and IG have 1% superior accuracy.

The second dataset with continuous features, *Steel plates faults*, exhibits a more apparent trend (Figure 5) compared to *Breast cancer* table. Namely, for Logistic Regression and Support Vector Machines, MRMR and JMI achieve twice better accuracy at 5 selected features already. However, when $10-15$ features are selected, MIFS and CIFE overtake JMI in effectiveness with 2%. Solely relying on IG is disadvantageous in this dataset as it needs to select $5-10$ more features than MRMR or JMI to reach an accuracy close to 100%.

Summary for datasets with continuous features: MRMR and JMI select the more effective features twice earlier than MIFS, CIFE, and IG for the three learning algorithms.

Datasets with discrete and categorical features

We now take a look at datasets with discrete and categorical features. Due to the computation costs of the complex entropy estimator, we used the simple one on the *Gisette* and *Internet advertisement* datasets. We believe that this does not negatively influence the performance of the feature selection algorithms as we are dealing with discrete features and the simple entropy estimator based on frequencies should suffice.

The results in Figure 6 are clear - MRMR, JMI, and IG achieve nearly 95% accuracy with a small subset of features, while MIFS does not exceed 83% accuracy. Furthermore, CIFE only reaches 92% for Logistic regression and Support Vector Machines at 200 features.

For all three models, MIFS does not select features that increase the accuracy of the algorithms. The reason is that the method does not estimate the conditional mutual information between already selected features and the candidate column. On the contrary, CIFE estimates this and outperforms MIFS. However, neither of the two approaches changes the hyperparameters α and β during the feature selection process, which leads to their underperformance. The baseline method, IG, ignores the relations between features and assumes they are independent. Ultimately, it outperforms MRMR and JMI on all three models at $100-150$ selected features.

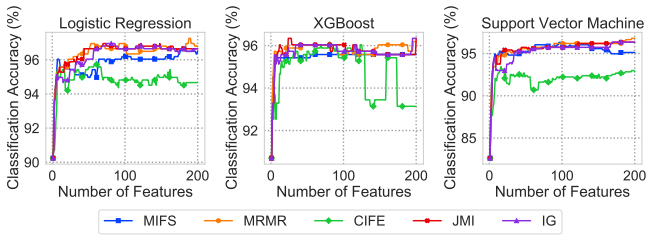

Figure 7: Accuracy of methods on Internet advertisements dataset

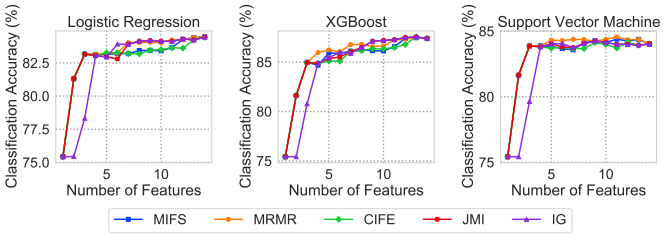

Figure 8: Accuracy of methods on Census Income dataset

The evaluation on *Internet Advertisements* dataset in Figure 7 confirms the tendency found for *Gisette* - MIFS and CIFE initially select more redundant features for the three learning models, as opposed to MRMR and JMI, and perform $1-2\%$ worse than them for Logistic Regression. Moreover, IG has a 0.5% worse accuracy than MRMR and JMI.

For the *Census Income* dataset, visualised in Figure 8, the four main approaches select the three most useful features together that bring the effectiveness of the machine learning model close to 83%. Beyond this point, the performance improvement is marginal and only within the $2-3\%$ range. Hence, the comparison between the approaches cannot be objectively established due to potential noise in the learning models. It is seen, however, that IG performs poorly during the selection of the first few features with only 80% accuracy.

Summary for datasets with discrete/categorical features: MRMR, JMI, and IG are more effective feature selection methods for all three machine learning models.

Mixed datasets

We now compare the methods using tables that contain features of all three types - continuous, discrete, and categorical.

The findings for *Housing prices* are displayed in Figure 9. There, MRMR and JMI get up to 50% lower RMSE than MIFS and CIFE with approximately 20 selected features. An explanation for that could be that they find a better balance between relevance and redundancy as they rely on dynamic values for α and β for scoring function J. These findings are confirmed for all three learning models. However, it is shown that IG has a nearly 50% lower RMSE than MRMR and JMI at 10 columns, potentially due to the independence of features. Hence, IG has a lower RMSE than the four methods.

The *Bike sharing* dataset (Figure 10), albeit having a small number of features, builds on top of the previous observations that IG, MRMR, and JMI perform better as the plot for Logistic Regressions indicates. It is interesting to observe that XGBoost achieves an optimal RMSE of 3.5 with only 2 selected features. This occurs because the five feature selection methods would pick the same 2 features and the model generalises well on them. Thus, no further difference is observed.

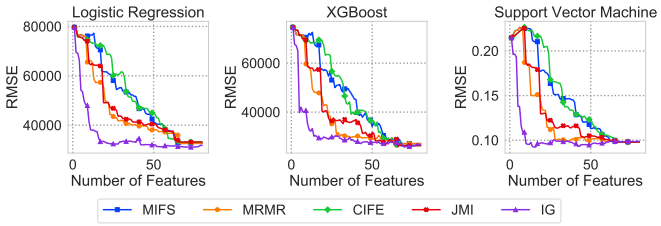

Figure 9: RMSE of methods on Housing prices dataset

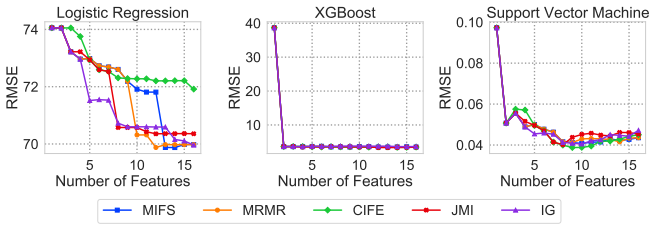

Figure 10: RMSE of methods on Bike sharing dataset

Summary for datasets with mixed feature types: IG outperforms all four methods with $50-100\%$. Nonetheless, MRMR and JMI are better in effectiveness than MIFS and CIFE for datasets with mixed feature types.

6.4 Tuning hyperparameter of MIFS method

Finally, we look at the effectiveness of MIFS under different values for the hyperparameter β (default being 0.5). Having $\beta = 0$ is equivalent to relying only on IG for feature selection.

Gisette (Figure 11), *Internet advertisements*, *Housing prices*, and *Bike sharing* datasets illustrate that the lower β is, the higher the model accuracy would be.

In contrast, *Steel plates faults* (Figure 12) and *Breast cancer* datasets show that having higher values for β leads to higher learning model accuracy.

Census Income dataset provides inconclusive results on the values of β. The graphs that illustrate the findings for the hyperparemeter β for all datasets can be found in Appendix B.

7 Discussion

The evaluation of the four information-theoretical-based feature selection methods in terms of efficiency and effectiveness has yielded several significant findings.

Efficiency of feature selection methods

We have empirically shown that MIFS and MRMR are $2-4$ times faster than CIFE and JMI for both the simple and complex entropy estimators. We attribute this to the fact that the former two do not deal with estimating conditional information gain and thus save computation time. What is more, IG can be $6-40$ times faster than MIFS and MRMR.

Effectiveness of feature selection methods

Regarding effectiveness, MRMR and JMI had better performance than the other two methods for all three types of datasets used - continuous, discrete/categorical, and mixed. We believe that the shortcomings of MIFS and CIFE come from the lack of flexibility in the methods as both have predefined values for α and β. This does not give space to consider the properties of the features in the dataset as more features

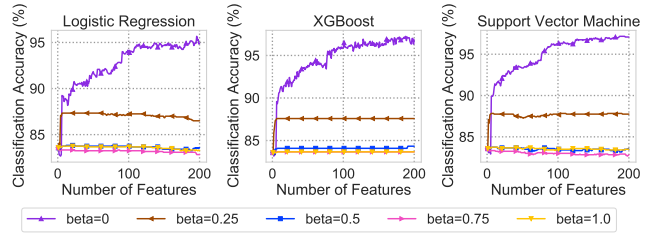

Figure 11: Accuracy of methods on Gisette dataset

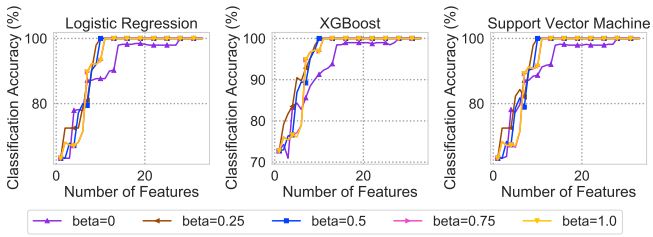

Figure 12: Accuracy of methods on Steel faults dataset

are added to the set of features $\mathcal{S}$. In contrast, MRMR and JMI dynamically update these hyperparameters. In particular, JMI starts with $\alpha = 1$ and $\beta = 1$ and decreases these values at each iteration. At the beginning of the feature selection process, they would be more strict in picking new features however as more features are added, the strengths of redundancy and relevance components would shift inversely. For example, with more features added, we penalise the candidates less for their redundancy and value their relevance more.

Thus, there is a trade-off to be made. For instance, MIFS takes less time than JMI to select columns but has worse model effectiveness. Provided that MRMR and JMI perform equally well in effectiveness and that MRMR has lower runtime, we would prefer to use MRMR for feature selection.

IG showed better accuracy and RMSE than any of the four approaches on some datasets. A potential explanation for that can be that the features in these datasets are independent of one another. Therefore, the estimations of joint information gain and conditional information gain do not bring sufficient insight into the feature selection. Nonetheless, the four methods compared in this paper have their advantages against their baseline counterpart as shown in certain other datasets.

Tuning hyperparameter of MIFS method

A linear trend can be observed regarding the values of β in MIFS. In some tables, with higher β, we receive higher effectiveness, but in some, the opposite is observed. In the original paper for MIFS, it is mentioned that generally values of 0.5 and higher lead to greater accuracy on classification tasks [2]. However, we empirically show that this is not always the case. In *Gisette* and *Internet advertisement* datasets, the lower β is, the higher the accuracy of learning models is.

8 Limitations

This comparative analysis was performed only on a limited set of tables. Therefore, our findings may not be generalisable for all datasets. However, given that we vary the types of features in them, we believe that we have mitigated this drawback. The wide range of metrics used allowed us to get a more broad view of the performance of the four approaches.

Another pressing limitation is that we have only used three machine learning algorithms during the evaluation. Furthermore, we have only applied hyperparameter tuning for the models on the entire dataset, rather than on each dataset composed of a subset of selected features. To tackle this limitation, we used both linear and non-linear models to introduce diversity in the comparison of the methods.

9 Conclusion & Future Work

The data used in machine learning algorithms strongly influences their performance. Using the entire dataset can increase the computation costs or bring noise to a model, which negatively impacts its learning capabilities. Feature selection algorithms can target this problem by selecting a subset of features that are capable of successfully achieving a given learning task. The focus of this research paper was to investigate what influences the performance of four information-theoretical-based feature selection methods MIFS, MRMR, CIFE, and JMI under the machine learning algorithms Logistic regression, XGBoost, and Support Vector Machines.

Entropy estimation. The study discovered that the four methods depend strongly on the entropy estimators they utilise. If the entropy estimator cannot successfully model the data, the estimations of information gain and conditional information gain become inaccurate. This results in method underperformance as high as 30% in accuracy. Therefore, to prevent this, continuous entropy estimators should be used when having continuous features.

Efficiency and effectiveness. The analysis has shown a huge imbalance in the efficiency of the approaches. CIFE and JMI are $2-4$ times slower than MIFS and MRMR in feature selection. However, the analysis of effectiveness has yielded different results - MRMR and JMI perform better than MIFS and CIFE on several datasets with distinct data characteristics. The computational limitations of JMI leads us to recommend MRMR as a feature selection method due to the displayed high efficiency and effectiveness. Nonetheless, for some datasets, IG leads to effective learning models. IG was also shown to be the fastest of all feature selection methods.

Future work. This paper briefly touched upon IG feature selection and further investigation into it can be beneficial. The performed study was only an empirical analysis of the four methods. Therefore, we have not looked from a theoretical perspective that systematically proves our findings. A future study on this can be performed to verify the conclusions we made in this paper. A limitation of the study was that only a few datasets and learning algorithms were used. Hence, a more elaborate comparison with additional tables and models can provide more in-depth results.

Acknowledgements

I am very grateful for the support I have received throughout the development of this research paper. I want to thank my supervisor Andra Ionescu and my responsible professor Asterios Katsifodimos for introducing me to the topic of data augmentation, for the inspiration they gave me through this course, and for the detailed feedback they provided me on my progress. It was a pleasure for me to work with you.

Responsible Research

The development of our comparative analysis has closely followed the principles introduced in the Dutch code of conduct for research integrity [29]. Namely: honesty, scrupulousness, transparency, independence, and responsibility.

Honesty. The datasets, algorithm implementations, and experiment results are publicly available in a repository on GitHub[1] in the folder *information-theoretical-based*. We adhere to the FAIR principles [31] and thus our research has the open Apache License 2.0 so that others can reuse our findings and modify our code with ease. A Docker file has been set up there, as well as, a requirements file, which ensures that the experiments are run in a standard environment with predefined versions of the Python packages. Therefore, all the findings here can be verified.

Scrupulousness and transparency. The datasets used in this research were collected from publicly available sources and were used as they were provided. Thus, no entries or features have been trimmed out. The choice of datasets was based on the types of features and not the actual topic or theme of the tables. The feature selection methods researched in this paper do not take into account the names of the features, their domain, or the origin of the data but purely rely on the distribution of values themselves. Given that the information-theoretical-based feature selection methods do not have the task to classify data or estimate values, we believe that their analysis does not introduce biases in the world.

Independence. The design methodology of the paper was chosen to be scientific and reproducible. During the comparison, we did not put any of the feature selection methods at a disadvantage. To mitigate the issues related to randomness, we reset the randomization seed of the feature selection algorithms every time a new run is performed. Therefore, the order of running the feature selection methods does not influence the result. Furthermore, all logs, results files, and graphs are available in the repository and no findings have been purposefully left out.

Responsibility. According to the tripartite model (users, engineers, politicians) [9], engineers are responsible for the creation of products and can be held accountable for breaking moral principles in that process. However, they are not blameworthy for issues that may arise from the improper use of their creations. Therefore, data scientists that decide to rely on the recommendations that we provide in this paper should carefully assess the potential ethical issues that can originate from their application in their respective domains.

Summary. This comparative analysis clearly explains the steps that were undertaken to produce these results and argues about the limitations discovered in the entire process. Furthermore, recommendations for future work are provided. The source code has been carefully documented to allow researchers to obtain and validate our findings. We believe that this was a very crucial step in our research, given the existing reproducibility crisis in science [1].

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

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 05 2016.

[2] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 07 1994.

[3] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, page 144–152, 1992.

[4] Gavin Brown, Adam Craig Pocock, Mingjie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 01 2012.

[5] Yudong Cai, Tao Huang, Lele Hu, Xiaohe Shi, Lu Xie, and Yixue Li. Prediction of lysine ubiquitination with mrmr feature selection and analysis. *Amino Acids*, 42:1387–1395, 01 2011.

[6] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. Auctus. *Proceedings of the VLDB Endowment*, 14:2791–2794, 07 2021.

[7] Chengliang Chai, Jiayi Wang, Yuyu Luo, Zeping Niu, and Guoliang Li. Data management for machine learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35:1–1, 2022.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016.

[9] Van De and Lamber M. M. Royakkers. *Ethics, Technology, and Engineering: an Introduction*. Wiley-Blackwell, 2011.

[10] 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:2003.06505 [cs, stat]*, 03 2020.

[11] Jian Fang, Li-Na Sui, and Hong-Yi Jian. Comparative analysis of continuous entropy estimation with different unsupervised discretization methods. In *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pages 367–370. Atlantis Press, 2013/03.

[12] Yuanyuan Fu, Zhong-Ping Jiang, Wenjiang Huang, and Jihua Wang. A comparative analysis of mutual information based feature selection for hyperspectral image classification. *Proceedings of the 2014 IEEE China Summit International Conference on Signal and Information Processing (ChinaSIP)*, 07 2014.

[13] Baofeng Guo and M.S. Nixon. Gait feature subset selection by mutual information. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 39:36–46, 01 2009.

[14] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. *2008 Fourth International Conference on Natural Computation*, 4, 2008.

[15] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 01 2021.

[16] Eva Kiktova-Vozarikova, Jozef Juhar, and Anton Cizmar. Feature selection for acoustic events detection. *Multimedia Tools and Applications*, 74:4213–4233, 06 2013.

[17] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69, 06 2004.

[18] Nojun Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1667–1671, 12 2002.

[19] Bi-Qing Li, Le-Le Hu, Lei Chen, Kai-Yan Feng, Yu-Dong Cai, and Kuo-Chen Chou. Prediction of protein domain with mrmr feature selection and analysis. *Public Library of Science ONE*, 7:e39308, 06 2012.

[20] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection. *ACM Computing Surveys*, 50:1–45, 01 2018.

[21] Shen Li, Guofa Li, Yang Cheng, and Bin Ran. Urban arterial traffic status detection using cellular data without cellphone gps information. *Transportation Research Part C: Emerging Technologies*, 114:446–462, 2020.

[22] Dahua Lin and Xiaoou Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. *European Conference on Computer Vision*, 9:68–82, 01 2006.

[23] Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. Responsible data integration: Next-generation challenges. In *Proceedings of the 2022 International Conference on Management of Data*, page 2458–2464, 2022.

[24] Hyeoun-Ae Park. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43:154, 2013.

[25] Hanchuan Peng, 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*, 27:1226–1238, 08 2005.

[26] Kashif Shaheed, Piotr Szczuko, Qaisar Abbas, Ayyaz Hussain, and Mubarak Albathan. Computer-aided diagnosis of covid-19 from chest x-ray images using hybrid-features and random forest classifier. *Healthcare*, 11(6), 2023.

[27] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656, 10 1948.

[28] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24:175–186, 03 2013.

[29] KNAW NWO VH VSNU, NFU. Netherlands code of conduct for research integrity. 09 2018.

[30] Janett Walters-Williams and Yan Li. Estimation of mutual information: A survey. In *Rough Sets and Knowledge Technology*, pages 389–396, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[31] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 03 2016.

[32] Howard Yang and John Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *Proceedings of Advances in Neural Information Processing Systems*, volume 12, 1999.

A Distributions of datasets

We present the distributions of the features of all datasets from Table 1 except *Gisette* and *Internet advertisements*. We have excluded the distributions of these two datasets in this paper due to the high number of features present in them. However, their column distributions are publicly available on the GitHub repository of the project for interested parties.

For each of the datasets, a Kernel Density Estimation of their distribution has been performed for each feature, which was not deemed categorical. Therefore, one can observe the modality and variance of the values in every column.

The feature distributions of each table can be found as follows: *Housing prices* (Figure 13), *Bike sharing* (Figure 14), *Steel plates faults* (Figure 15), *Census Income* (Figure 16), and *Breast cancer* (Figure 17).

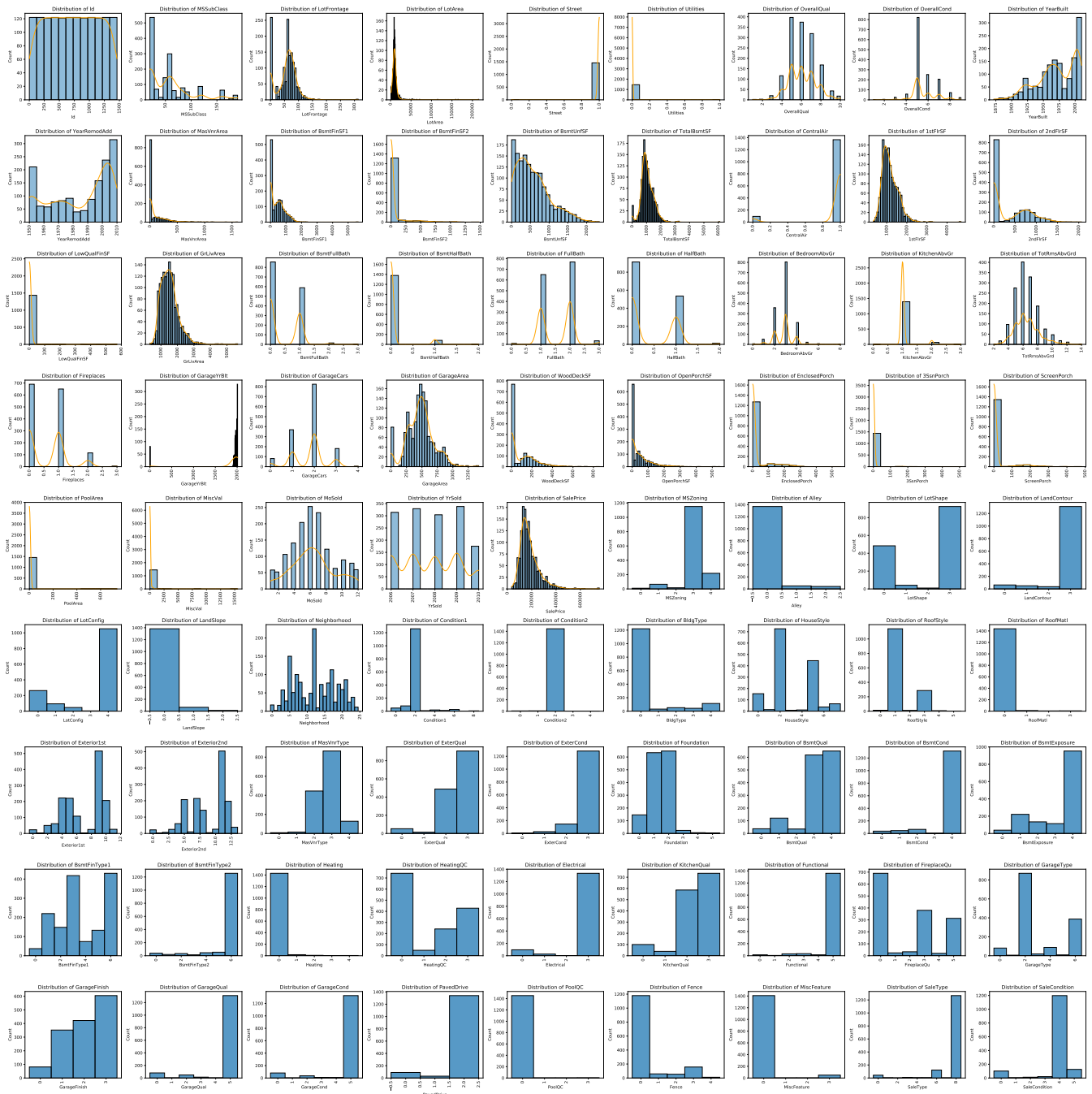

Figure 13: Feature distribution of Housing prices dataset

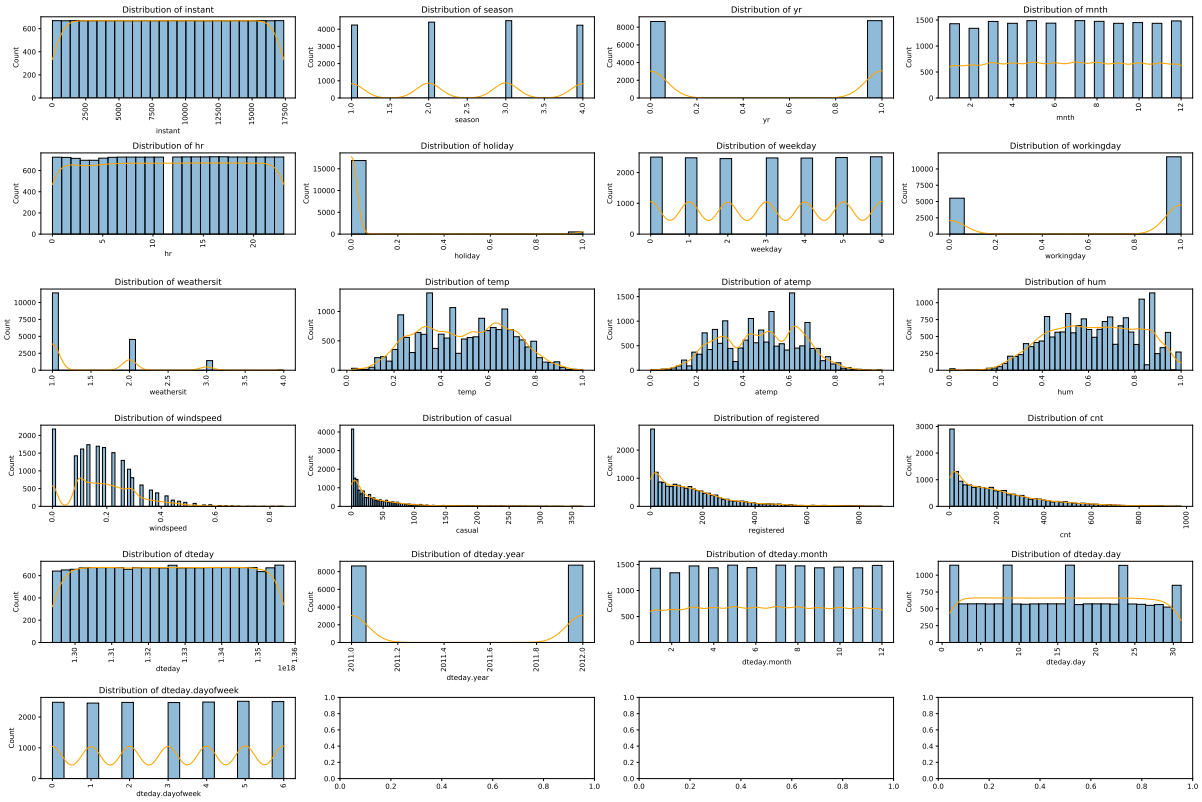

Figure 14: Feature distribution of Bike sharing dataset

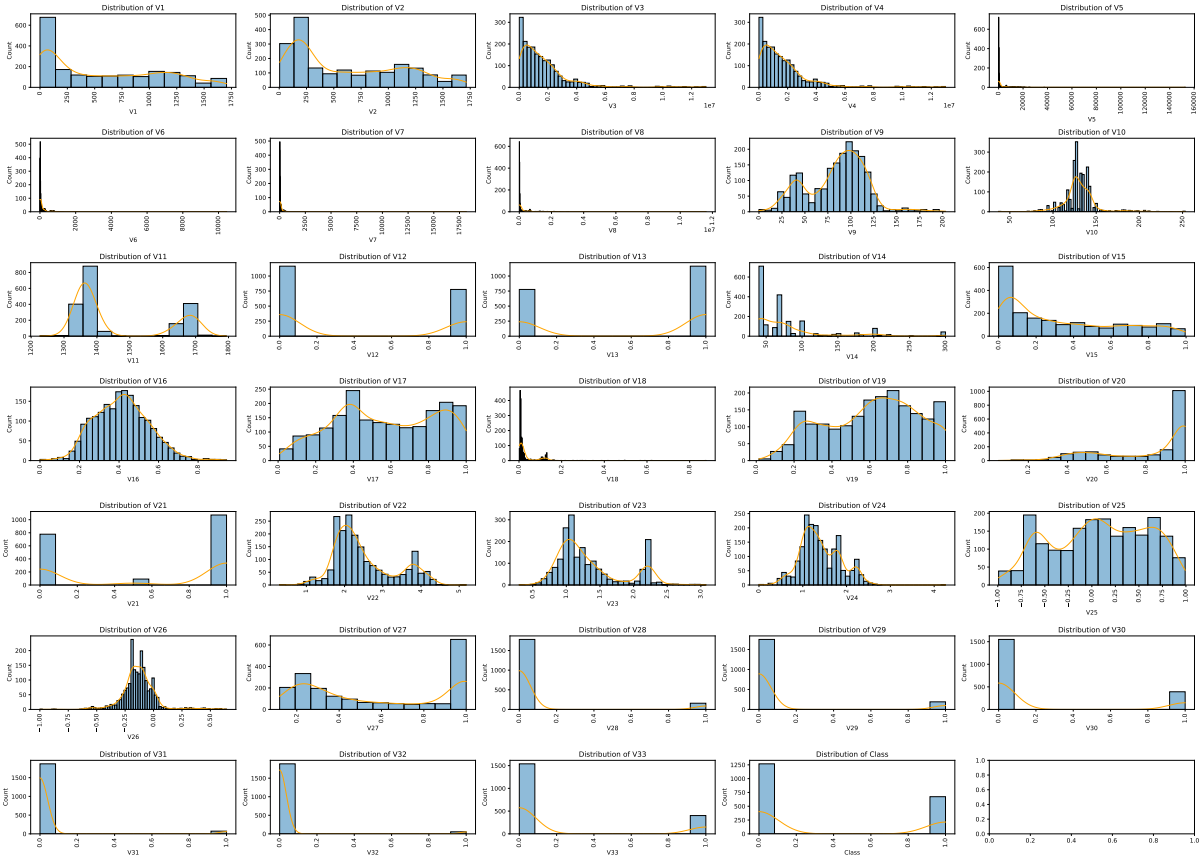

Figure 15: Feature distribution of Steel plates faults dataset

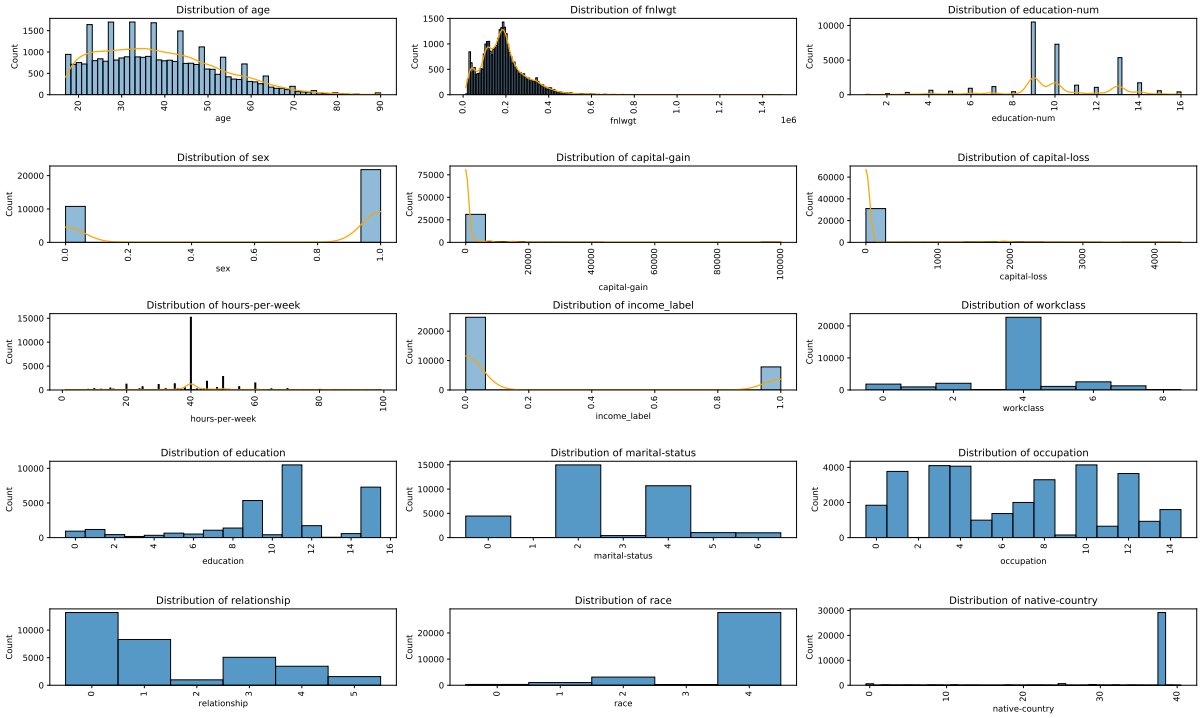

Figure 16: Feature distribution of Census Income dataset

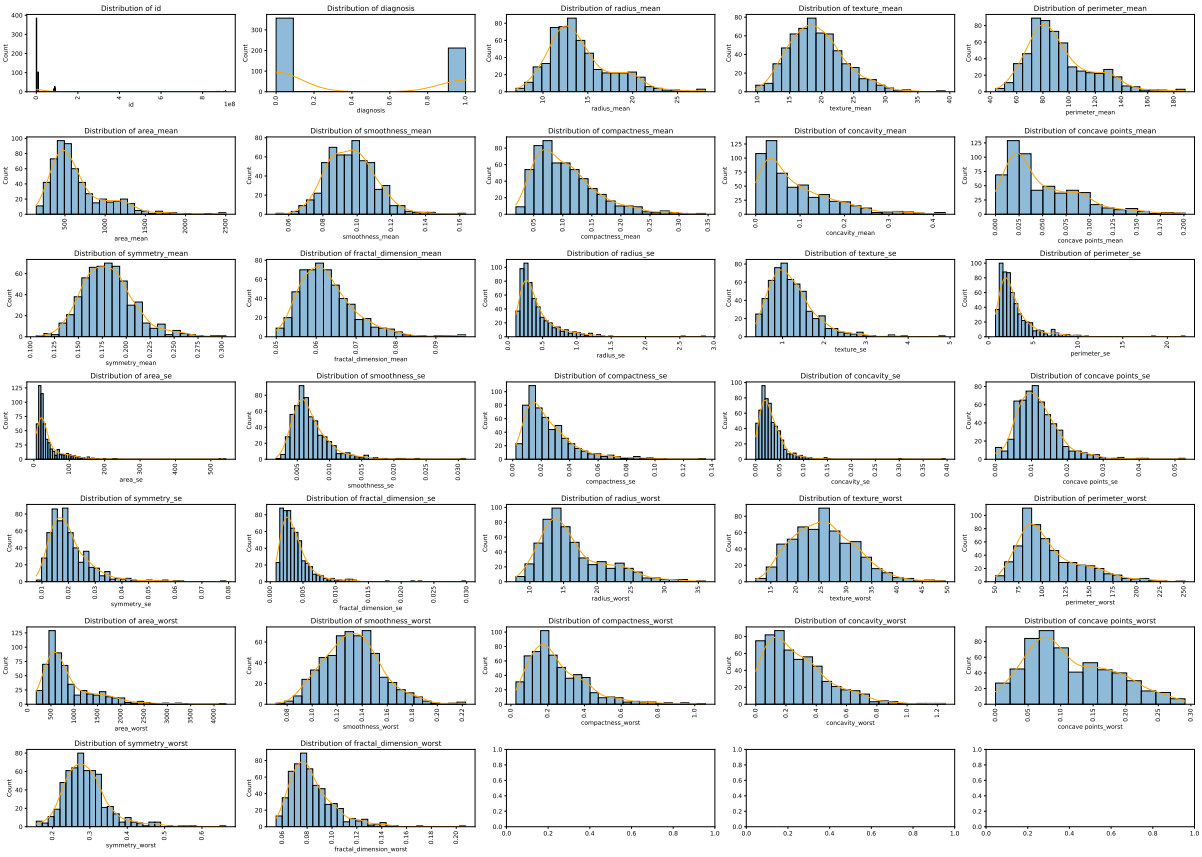

Figure 17: Feature distribution of Breast cancer dataset

B Hyperparameter tunings of MIFS

In this section, we provide the graphs that present the changes in effectiveness based on different values for β in MIFS method.

The performance for each dataset can be found as follows: *Gisette* (Figure 18), *Internet advertisements* (Figure 19), *Bike sharing* (Figure 20), *Housing prices* (Figure 21), *Steel plates faults* (Figure 22), *Breast cancer* (Figure 23), and *Census Income* (Figure 24).

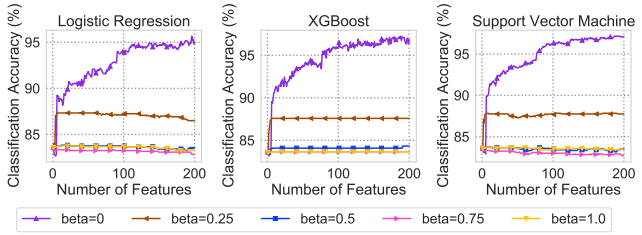

Figure 18: Accuracy of methods on Gisette dataset

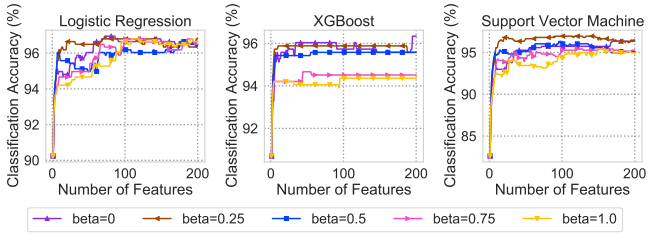

Figure 19: Accuracy of methods on Internet advertisements dataset

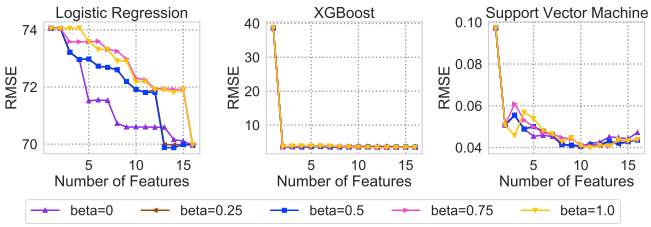

Figure 20: RMSE of methods on Bike sharing dataset

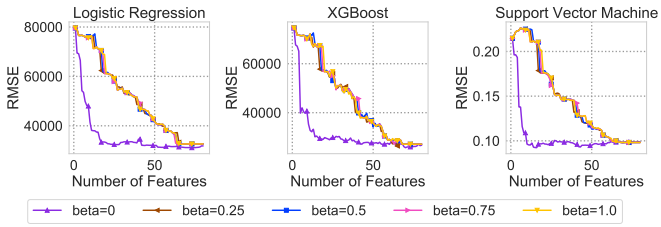

Figure 21: RMSE of methods on Housing prices dataset

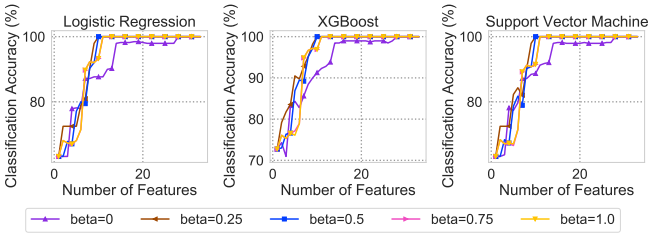

Figure 22: Accuracy of methods on Steel faults dataset

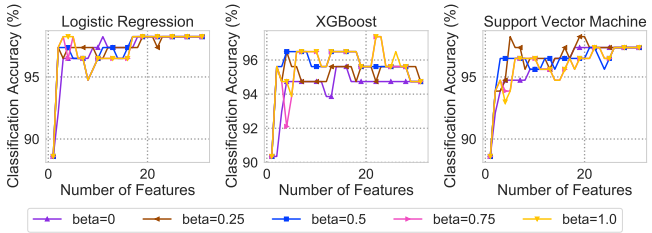

Figure 23: Accuracy of methods on Breast cancer dataset

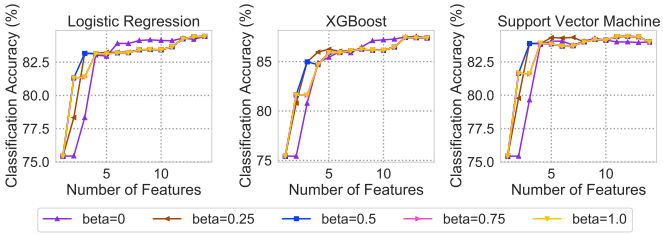

Figure 24: Accuracy of methods on Census Income dataset