Feature Selection Methods to Detect Faking in Psychological Tests

Can Abdullah Camuz
canabdullah.camuz@studenti.unipd.it

Desaru Abhishek Varma
abhishekvarma.dasaraju@studenti.unipd.it

Luis Marcos López
luismarcos.lopezcasines@studenti.unipd.it

Abstract

"The identification of crucial features to spot faking in psychological tests is addressed in this study. The issue of feature stability in machine learning models is acknowledged and several model-independent feature selection techniques are examined for diverse psychological tests. These model-independent methods have been found to be more stable than the machine learning models, though they may not completely agree on the selected features. It is found that selecting 20% of the features using these techniques serves as a good balance between the number of items and the relative change in accuracy."

1. Introduction and motivation

Feature selection is the process of identifying and selecting a relevant subset of features to be used in model construction. It helps to simplify the model, increase its interpretability, decrease the computation, and improve the ability of generalizing unseen data.

Excessive use of features can lead to overfitting, where the model performs well on the training data but poorly on out of sample data. This can occur when the model learns irrelevant noise or random variations in the data, instead of focusing on the underlying relationships. By carefully selecting a subset of relevant features, we can prevent overfitting and improve the model's generalization ability.

Furthermore, by using a smaller set of features, the model becomes more interpretable [1], making it easier to understand how it makes predictions and to communicate the results to others. There are various methods for feature selection such as statistical tests, model-based techniques and optimization methods [2].

The purpose of this research is to evaluate various feature selection techniques for their ability to identify the key questions in psychological tests that can distinguish between genuine and faked symptoms. To accomplish this, we will use a diverse set of datasets containing answers to psychological test questions and measure the performance of both model-dependent and model-independent feature selection methods. The evaluation will focus on two aspects: the stability of the selected features and the achieved accuracy in recognizing faked responses.

2. Faking in psychological tests

Psychological tests are standardized methods for measuring various mental abilities, emotional states, and personality traits. They are used in a wide range of settings, such as in clinical and research settings, to help diagnose mental disorders, evaluate cognitive and emotional functioning, and to assess personality characteristics.

Faking, also known as "response bias" or "social desirability bias", refers to the act of intentionally responding to psychological test items in a way that presents oneself in a favorable light, rather than providing an honest or accurate response. Faking can take two forms: faking good and faking bad. Faking good refers to exaggerating one's strengths or abilities, while faking bad refers to exaggerating one's weaknesses or limitations.

The main challenge when using psychological tests to identify faking is that many of the model-dependent methods are known to identify different features every time they are applied to the same dataset, which makes it difficult to establish a stable and reliable process for identifying fakers. As a solution, researchers are now investigating model-independent feature selection methods, which may be more robust and stable than model-dependent methods. The goal is to find a small amount of features that can help quickly identify faking in psychological tests, regardless of the specific model used.

3. Datasets

The research we are conducting involves a total of 16 datasets, which are divided into two groups: faking good and faking bad. For each dataset, subjects were first asked to answer honestly to a set of questions and then asked to fake a corresponding pathology or goal during a second test.

The faking good datasets include a range of tests such as the dark triad test applied to a job interview scenario and a cause for the custody of children, caregivers' ability to mentalize with their children, Big Five personality traits, a full dark triad test, and a five dimensions of human personality test in the context of a job interview for a salesperson position, a humanitarian organization and obtaining a child custody in the context of a litigation.

Similarly, the faking bad datasets include tests such as a questionnaire on prospective and retrospective memory, Post Traumatic Stress Disorder (PCL and IES-R versions), victims of mobbing, Anxious-depressive syndrome, mental disorders (short and long versions), and Adjustment Disorder test. The basic information about the datasets is found in Table 1.

Given the large number of datasets, the results of only one or two tests will be presented where relevant, and the rest of the results can be found in the appendices.

4. Brief EDA

As a way to understand the data better, the mean value of answers of each question in a psychological test and its standard deviation were plotted. This was done for both the honest and dishonest answers separately. Fig. 1 shows the results for the anxious depressive test (from now on dataset 5) and the five dimensions of the human personality test in the context of a job interview for a humanitarian organization (from now on dataset 15). The former is an example of faking bad, while the latter corresponds to faking good.

When individuals are attempting to fake bad, they tend to exaggerate symptoms more, as evidenced by a larger gap in the mean values of their answers. For example, in dataset 15, which is an example of faking good, there is a smaller gap between the means of honest and dishonest answers compared to dataset 5. The Likert scale used in the psychological tests we work with is designed such that higher answers correspond to the presence of the symptom/competence in question. However, this is not the case for all datasets. In datasets 1 (shortDT) and 11 (DDDT), dishonest answers tend to be lower because the presence of the symptom or competence is associated with a lower answer.

For the machine learning models and model-independent selection methods that are described in the following sections, no distinction is made between faking good and faking bad datasets, since we consider the differences they may have are unimportant for our study.

The difference caused by faking good and bad was ignored since it did not affect those models and techniques.

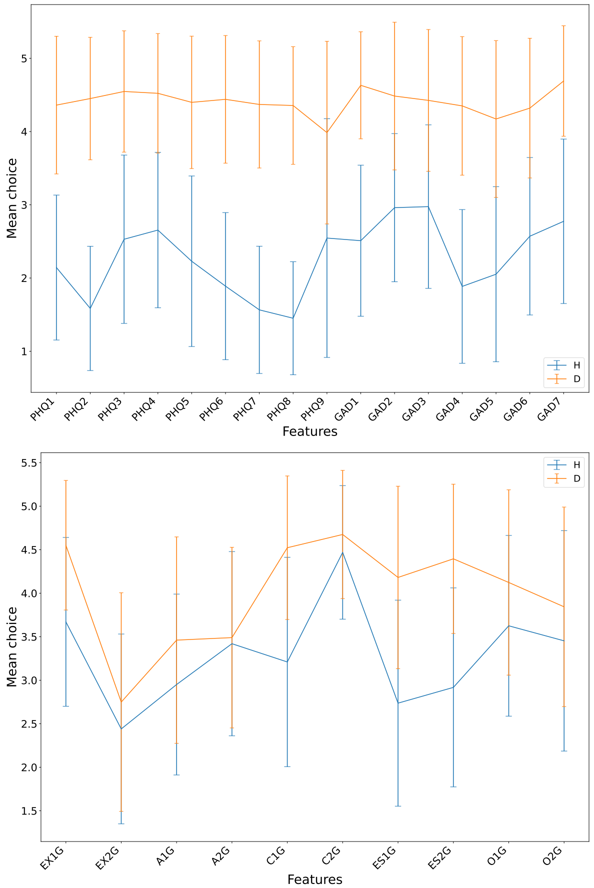


Figure 1: Mean value of answers for each question and its standard deviation in the anxious depressive test (top) and the 5 dimensions of human personality test in the context of a job interview for a humanitarian organization (bottom). Honest answers in blue, dishonest in orange.

5. Models, selectors and stability measures

5.1. Machine Learning Models

This paper focuses on five machine learning models. On one hand, we have tree based models, such as trees themselves, random forest and XGBoost.

A **tree** is a simple yet powerful machine learning algorithm used for both classification and regression tasks. It can be used to visualize and model complex decision processes in an easy-to-understand way. A tree is composed of a set of nodes and edges, where each node represents a test on an attribute, each branch represents the outcome of a test, and each leaf node represents a class label.

A **random forest** is an ensemble method that consists of multiple decision trees. It works by creating a large number of decision trees, and then averaging the predictions of all the trees to arrive at the final prediction. This averaging helps to reduce overfitting and improve the overall perfor-

Dataset Name	No of examples	No of items	Likert scale	Faking category
SD3	JI - 482 CC-864	27	1-5	Good
PMRQ	1404	16	1-5	Bad
PTSD	402	20	0-4	Bad
NAQ-R	712	22	1-5	Bad
PHQ9_GAD7	1118	16 (9+7)	1-5	Bad
PID5	824	220	0-3	Bad
s_PID5	1038	25	0-3	Bad
PRFQ	526	18	1-5	Good
IESR	358	22	0-4	Bad
Revised_NEO_PI	77687	120	1-5	Good
DDDT	984	12	1-5	Good
IADQ	450	9	1-5	Bad
BF	CC-486 JIS-442 JIHO-460	10	1-5	Good

Table 1: Information about the datasets.

mance of the model.

XGBoost (eXtreme Gradient Boosting) is a gradient boosting algorithm that is used for supervised learning tasks. It is designed to be efficient and scalable and is particularly good at handling large datasets and high-dimensional feature spaces. It's efficient because it uses gradient descent algorithm to minimize the loss and it reduces the complexity of the model by using regularization.

As for models that rely on coefficients, we will work with logistic regression and support vector machines. **Support vector machines** are supervised learning algorithm that can be used for classification and regression tasks. The algorithm finds the best boundary (or "hyperplane") that separates the different classes in the data. The data points that are closest to the boundary are called support vectors.

Logistic regression is a statistical method that is used for predicting binary outcomes based on one or more features. The model creates a mathematical equation that predicts the probability of a certain outcome based on the input variables. The model uses this probability to make a prediction about the outcome using a threshold (usually 0.5).

These models can determine the importance that each feature has for them. But they are, of course, biased (depends on the model).

5.2. Feature selection methods

As an alternative to choosing features with machine learning models, there are model independent feature selection methods. The ones we will work with are chi squared (χ^2), mutual information, permutation importance, ANOVA and principal component analysis (PCA) [3, 4, 5].

The **chi-squared test** is a statistical test commonly used to select features in a dataset. It measures the association between a categorical variable and a target variable. The test compares the observed frequency of a category in a variable to the expected frequency of that category, and calculates a chi-squared statistic to indicate the strength of association. Features with a high chi-squared statistic are more likely to be relevant for the target variable.

Mutual Information is a non-parametric feature selection method that measures the dependence between two random variables. It is used to assess how much information the presence or absence of a feature provides about the outcome variable. Features with a high mutual information score are considered to be more important, as they contain a lot of information about the outcome variable. Mutual information is defined as the Kullback-Leibler divergence between the joint probability distribution and the product of the marginal probability distributions.

Permutation importance is a method to evaluate the importance of individual features in a model. It is a post-hoc technique that can be used with any model to assess feature importance (in our case KNN). It works by randomly permuting the values of a feature, and then measuring the effect of this random permutation on the model's performance. The idea behind this method is that if a feature is important, then randomly shuffling the values of that feature should have a large effect on the model's performance, as the model will lose access to that feature's information. By computing the change in the model's performance caused by permuting each feature, we can rank the features by their importance.

ANOVA (Analysis of Variance) is a statistical method that is commonly used for feature selection to test for the difference in means of multiple groups or levels of a categorical independent variable in relation to a continuous dependent variable. However, if the outcome variable is categorical binary output, one way to use ANOVA is to recode the binary categorical outcome variable into two separate continuous variables, one representing one class of the binary outcome and the other representing the other class. Then, ANOVA can be applied to test for differences in means of the predictor variables between the two classes.

Principal Component Analysis (PCA) is a technique for dimensionality reduction, that can be used as a preprocessing step before applying feature selection techniques. PCA is a linear transformation method that uses orthogonal transformation to convert a set of correlated variables into a set of linearly uncorrelated variables called Principal Components (PCs). The goal of PCA is to find a new set of features that can explain most of the variance in the original dataset. For that, one can choose the number of PCs based on the explained variance ratio, where a threshold is selected and the components that are above this threshold are kept. In a next step, by checking the correlation between components and features of the dataset, the most correlated features are selected to achieve the dimensionality reduction.

Here it's important to highlight that the principal components can be extracted from the entire dataset (honest and dishonest) or only for the honest answers. When PCA is performed with the whole dataset, it can be used to identify patterns of variation in the data that can be used to identify individuals who are likely to have faked their responses. On the contrary, the approach of performing PCA with only the honest answers is useful when the goal is to understand the underlying mechanisms of honest responding. In this work we compare both approaches.

5.3. Stability measures

When selecting features for a machine learning model, it is important to evaluate the stability of the feature selection process. One way to do this is to use accuracy and Jaccard similarity as stability measures.

Accuracy is a common metric used to evaluate the performance of a classification model. It is the proportion of correct predictions made by the model, compared to the total number of predictions. In other words, it is the number of correct predictions divided by the total number of predictions.

Jaccard similarity, also known as Jaccard index, is a measure of the similarity between two sets. It is defined as the size of the intersection of two sets divided by the size of their union. In the context of feature selection, Jaccard similarity can be used to measure the similarity between the selected features in different runs or folds of a cross-validation process.

By using accuracy and Jaccard similarity, we can have a better understanding of how stable the feature selection process is. High accuracy scores indicate that the model is making accurate predictions, while high Jaccard similarity scores indicate that the selected features are consistent across different feature selection methods.

6. Results

6.1. The Problem of Model Dependent Methods

To evaluate the performance of the models, the data was split into a training set and a test set, with the ratio being 80% for training and 20% for testing. Five machine learning models were employed to classify the data into two categories: honest and dishonest. For each model, the feature importance was calculated.

Tree-based methods determine feature importance by measuring the reduction in impurity (such as Gini impurity or information gain) in the target variable that is achieved by each feature at each split in the tree. On the other hand, coefficient-based methods determine the importance of each feature based on the value of the coefficient assigned to the predictor.

As seen in Figure 2, the normalized importance for each feature as calculated by the various machine learning models is depicted. The feature importance values obtained from tree-based models are similar among them, and the same holds true for coefficient-based models. However, there is a lack of consistency between the feature importance values yielded by the different model types, and even within the same model type.

Furthermore, this trend is consistent across all other datasets, regardless of whether they belong to the faking bad or faking good category. This highlights the fact that the selection of the most important features for predicting faking in a test will directly depend on the model used to train the data. It was also observed that the tree-based models were better for reducing the number of features since they selected fewer features as important. Yet coefficient-based models distribute the weight among the features. Hence, the main conclusion was that machine learning models were not enough to perform an efficient feature selection, and that was the reason for the necessity of applying model-independent methods.

6.2. Selection of 20% of Features

The most important 20% of features were selected with both the statistic based feature selection methods and the machine learning models. Fig. 3 shows the normalized importance that machine learning models gave to the features picked by χ^2. Again, there was no agreement between different machine learning models. And this holds for the features selected by other model-independent features selected

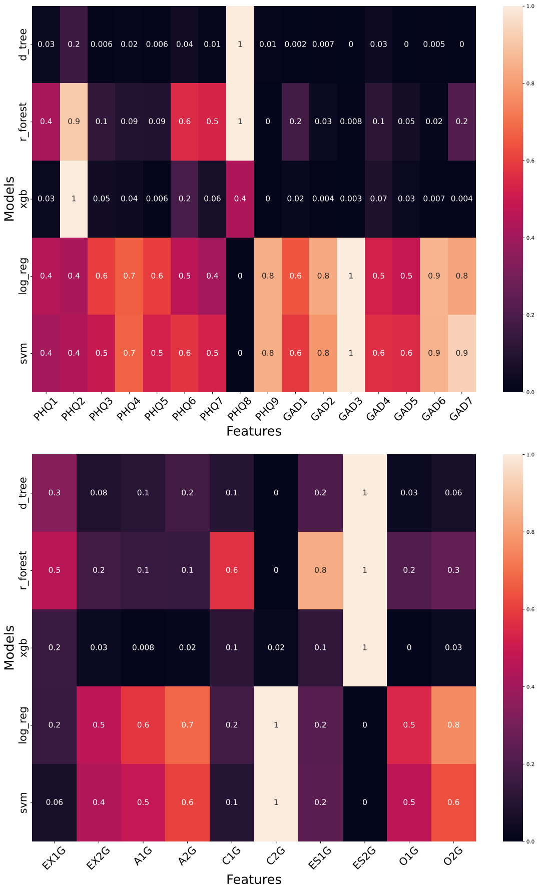


Figure 2: Normalized importance for each feature as calculated by the various machine learning models. Datasets 5 (top) and 15 (bottom).

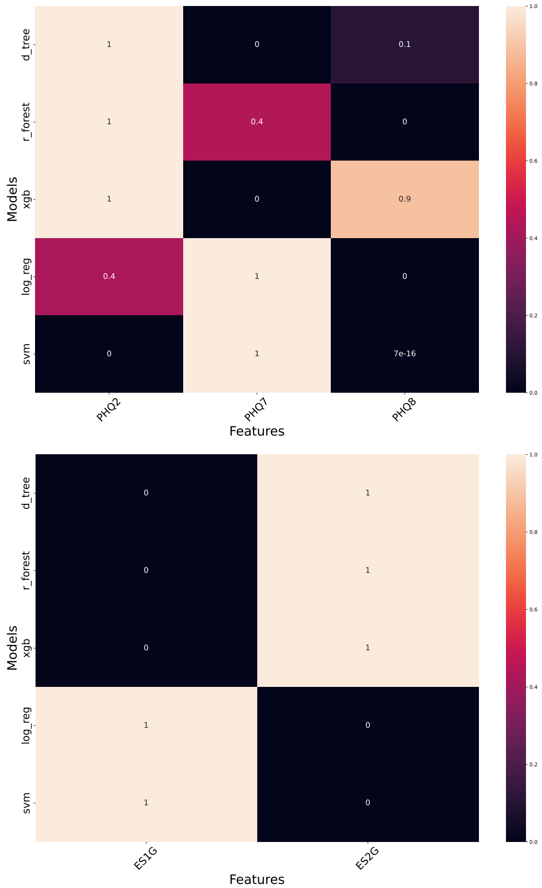


Figure 3: Normalized importance that machine learning models give to the features that statistic based feature selection methods pick. Datasets 5 (top) and 15 (bottom).

methods as well.

The results were compared and analyzed depending on the stability measures defined in section 5.3.

6.3. Jaccard Similarity

Jaccard similarity was used to assess the stability of the selected features by different selection techniques, since it is a measure of similarity between two sets. The Jaccard similarity coefficient ranges from 0 (no similarity) to 1 (identical sets). Fig. 4 shows the jaccard similarity of the set of features selected by the machine learning models and feature selection methods for datasets 5 and 15.

As it was discussed earlier, tree-based machine learning models behave similarly and tend to select the same sets of features. This is also the case for coefficient-based models. However, the two types of models don't share common relevant features. As per the similarity within feature selection methods, we observe that χ^2, mutual information and ANOVA usually agree on the selected features. There is a partial agreement with permutation importance and no agreement at all with the features chosen by the PCA analysis.

It's also worth noticing that feature selection methods never agree with coefficient-based models, and in general, agree poorly with tree-based models.

6.3.1 Accuracy

Each machine learning model yields an accuracy when it's trained with the full set of features and another accuracy when it's trained with a chosen subset of the features. To quantify the difference in accuracy before and after, we use the relative change in accuracy as shown in Eq. (1)

$$\text{Relative change in accuracy } (\%) = \frac{\text{SA} - \text{FA}}{\text{FA}} \times 100, \quad (1)$$

where FA and SA are the accuracy of the model trained with the full set and a subset of the features respectively.

Fig. 5 shows the relative change in accuracy of the machine learning models when the 20% more important fea-

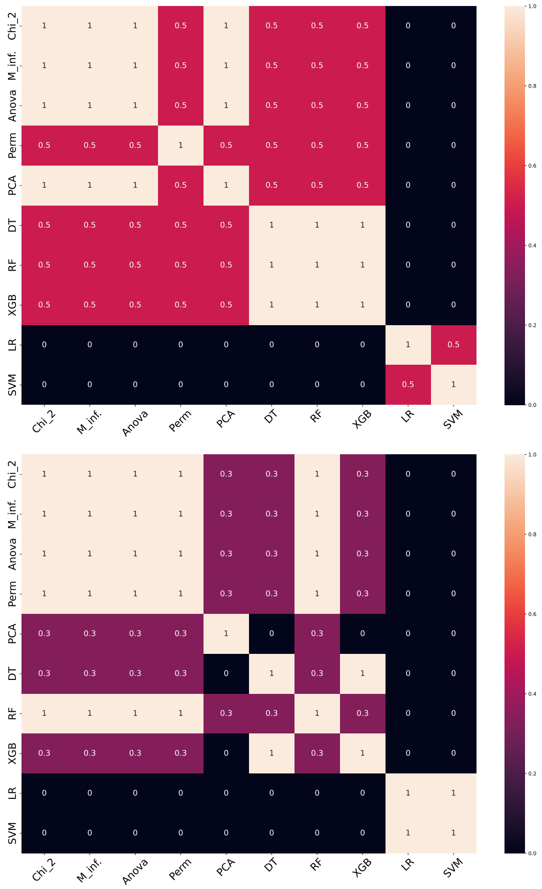


Figure 4: Jaccard similarity for the subset of features selected with feature selection methods and machine learning models. Datasets 5 (top) and 15 (bottom).

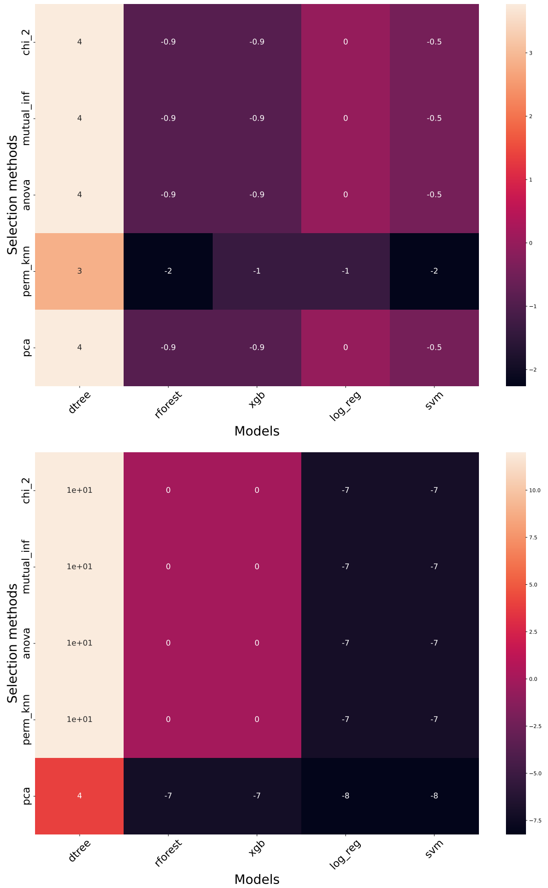


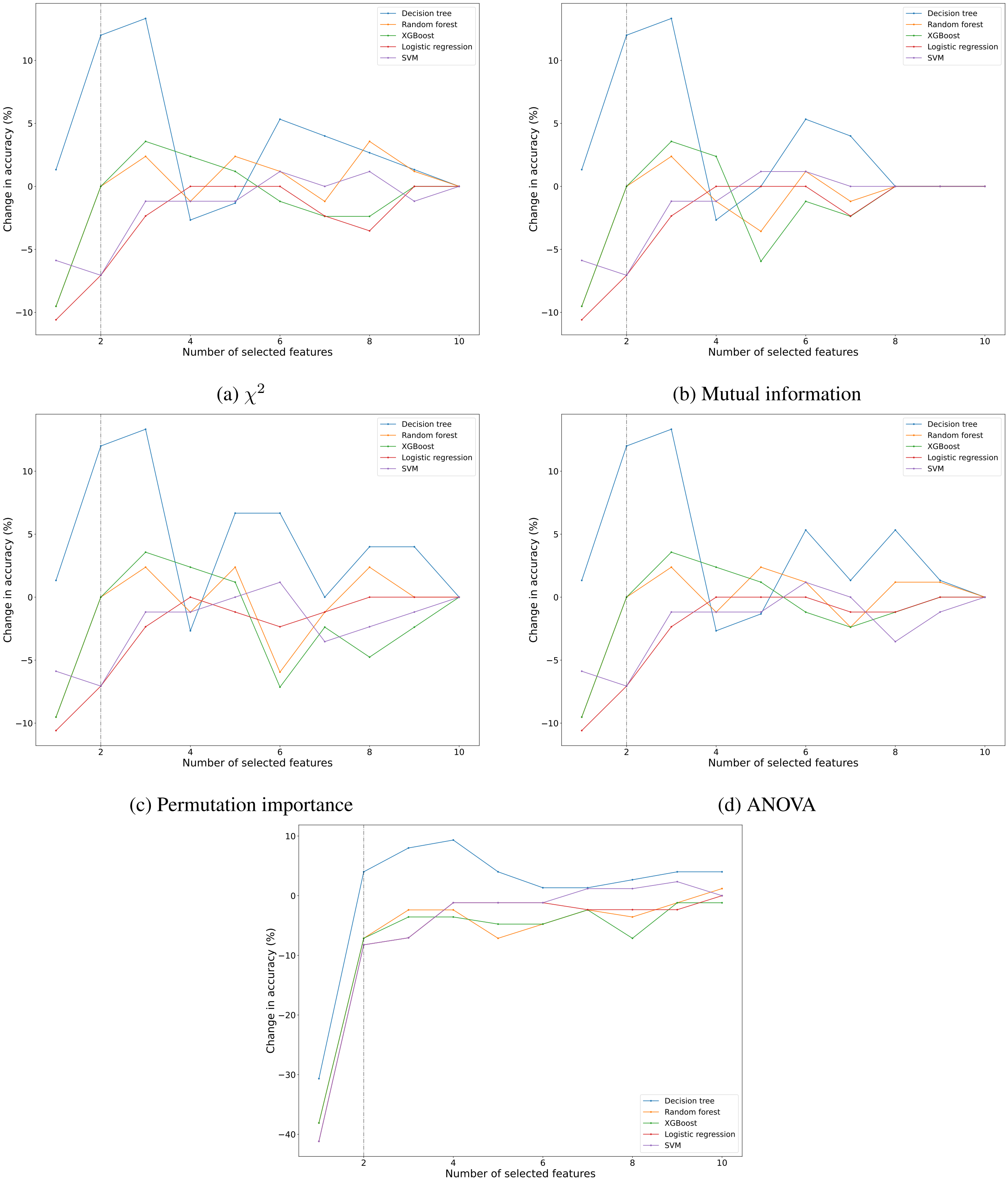
Figure 5: Relative change in accuracy of the machine learning models when the 20% more important features are selected with the feature selection methods described in section 5.2. Datasets 5 (top) and 15 (bottom).

tures were selected with the feature selection methods described in section 5.2. As a result of feature selection methods agreeing on the chosen features, most of the values within one column (machine learning model) are the same. Decision tree seems to perform very well in general with the selected set of features. In contrast, other machine learning models perform poorly depending on the dataset and the model. A relative accuracy change of less than -5% is considered bad in general.

Next, the behavior of the relative change in accuracy was tested for different feature selection methods as a function of the number of selected features. Fig. 6 provides information about the significant amount of features required to achieve better generalization on the classification task. For machine learning models and feature selection methods 20% of features is a good compromise between accuracy and number of features. However, taking 30% of the features would considerably increase the relative change in accuracy in some cases. Worth noticing also is that PCA shows the most stable behavior and that the tree classifier is the best performing model for a wide range of the number of selected features, independently of the feature selection method used.

To compare the different feature selection methods, the mean relative change in accuracy for a varying number of selected features was plotted (Fig. 7). Mean here refers to the average of the relative change in accuracy for each machine learning model when features were selected with a specific feature selection method. As we can see from Fig. 7, 20% of features seems to be a good compromise between the number of features selected and the accuracy performance we are willing to sacrifice. Once this reference point of 20-30% of features is reached, all features selection methods behave quite similarly.

Finally, a study of the selection of features with PCA was carried out. As previously explained, the most significant features are chosen based on the correlation with the principal components extracted by PCA. The aim was to test whether extracting the principal components with only the



(a) χ^2

(b) Mutual information

(c) Permutation importance

(d) ANOVA

(e) PCA

Figure 6: Relative change in accuracy for various feature selection methods as a function of the number of selected features. Dataset 15. Vertical line indicates 20% of features.

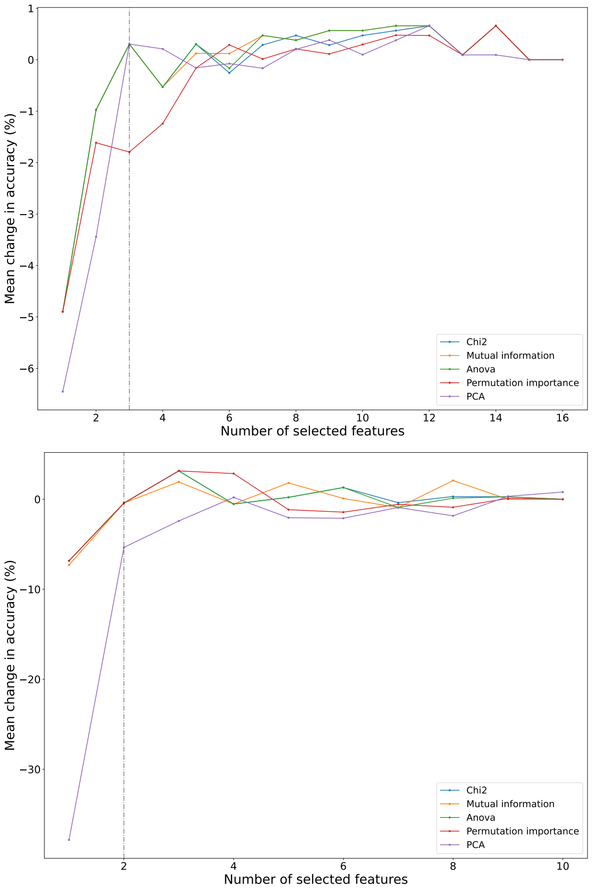


Figure 7: Mean relative change in accuracy for a varying number of selected features. Datasets 5 (top) and 15 (bottom). Vertical line indicates 20% of features.

data corresponding to honest responses would yield more significant features for our classification task. The results are depicted in Fig. 8. Depending on the dataset, both PCA methods produce similar mean changes in accuracy. However, what seems to be the norm is that the best results come from considering both honest and dishonest answers.

7. Conclusions

Machine learning models regard different features as important for classification. Therefore they are not suitable to identify a small subset of items that allow to correctly recognize faking. It was observed that the tree-based models behave similarly as well as the coefficient-based models. However, they don't agree on the most important features. This fact is independent of the behavior or purpose of the respondent, who might either fake good or bad by giving lower or higher marks for the questions. This holds true for all psychological tests in our analysis.

Since it was observed that machine learning models did not agree on the most important features, model-agnostic methods were employed. They tend to agree more with each other regarding the features chosen, and even with some of the tree-based machine learning models. However, this agreement is not complete.

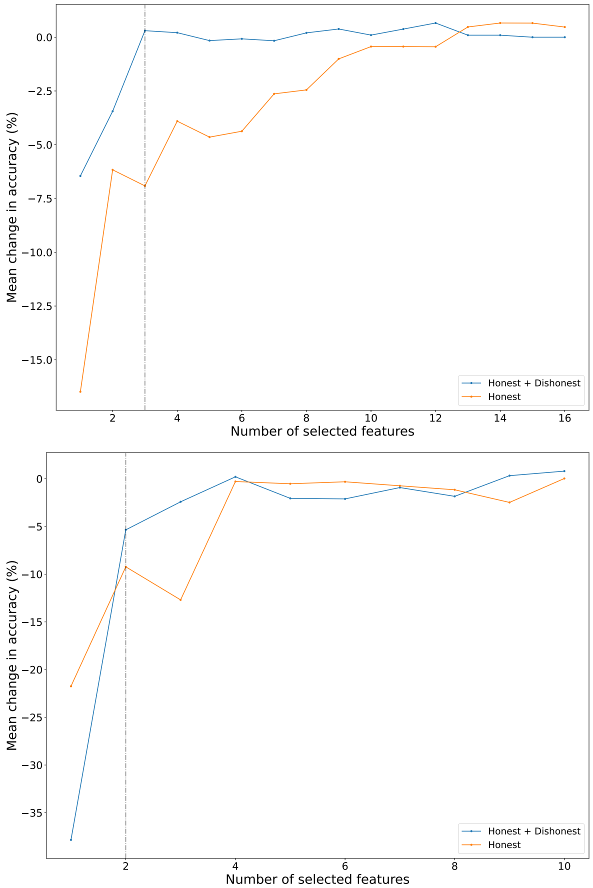


Figure 8: Mean relative change in accuracy for a varying number of selected features. Comparison of the use of PCA on all the dataset and on the honest answers. Datasets 5 (top) and 15 (bottom). Vertical line indicates 20% of features.

As per the relative change in accuracy after reducing the number of features, we found in general good results. Trees usually yield an improvement in the accuracy, but other machine learning models' results are within our acceptance threshold (higher than -5%) for a wide range of selected features, regardless of the selection method that was used. In most datasets, 20% of the features yield a good compromise between number of items and accuracy that we are willing to sacrifice. However, we suggest taking 30%, since the results will be mostly better.

Finally, our analysis revealed that utilizing principal component analysis on the combined dataset of both honest and dishonest responses resulted in a more effective feature selection. This, in turn, resulted in improved performance of the machine learning models.

References

[1] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

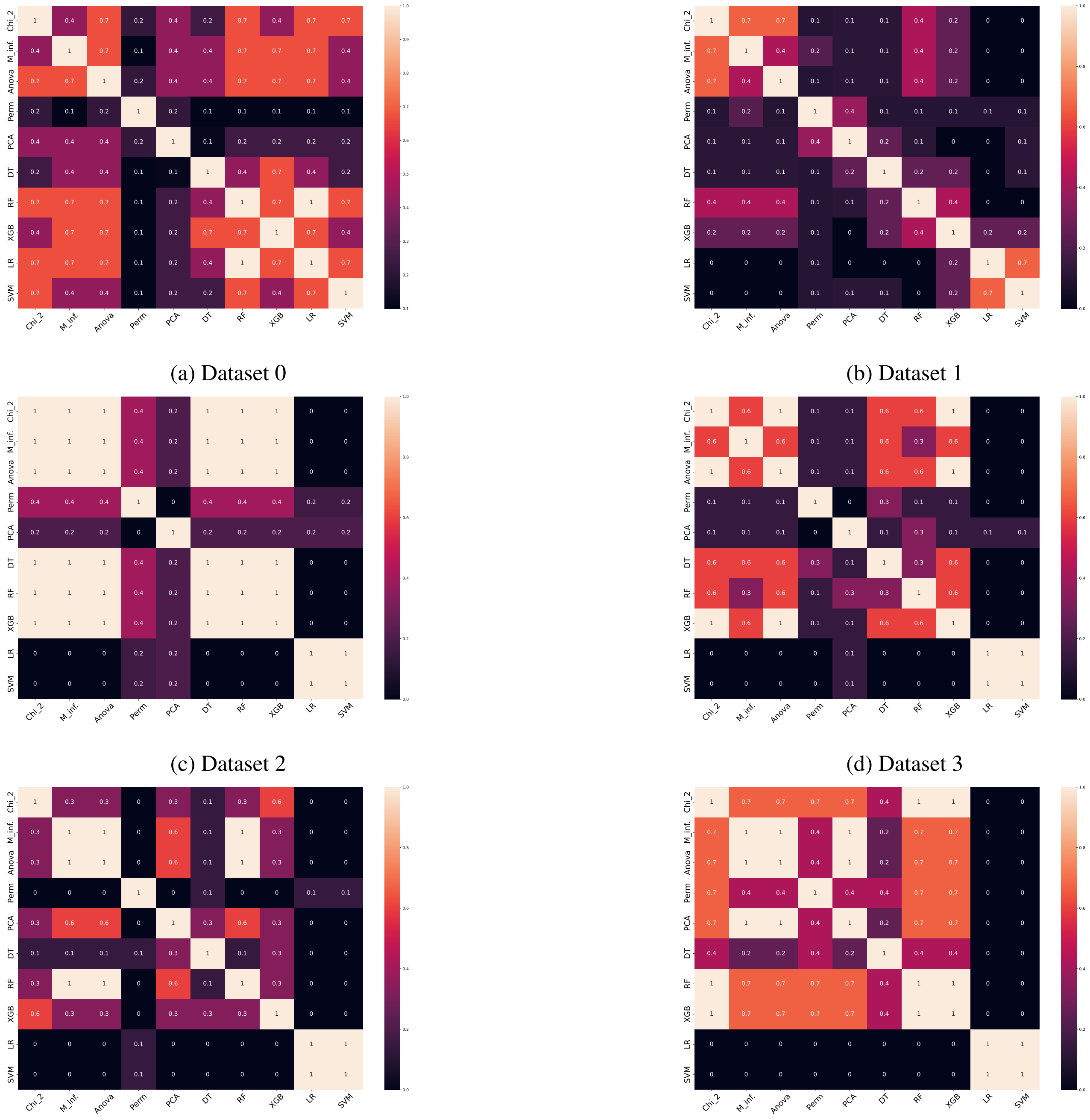
[2] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[3] Jason Brownlee. "How to choose a feature selection method for machine learning". In: *Machine Learning Mastery* 10 (2019).

[4] J Brownlee. "How to perform feature selection for regression data". In: *URI= https://machinelearningmastery. com/feature-selection-for-regression-data* (2020).

[5] Jason Brownlee. "How to calculate feature importance with python". In: *Machine Learning Mastery. https://machinelearningmastery. com/calculate-feature-importance-with-python* (2020).

A. Appendix - Jaccard similarity



(a) Dataset 0

(b) Dataset 1

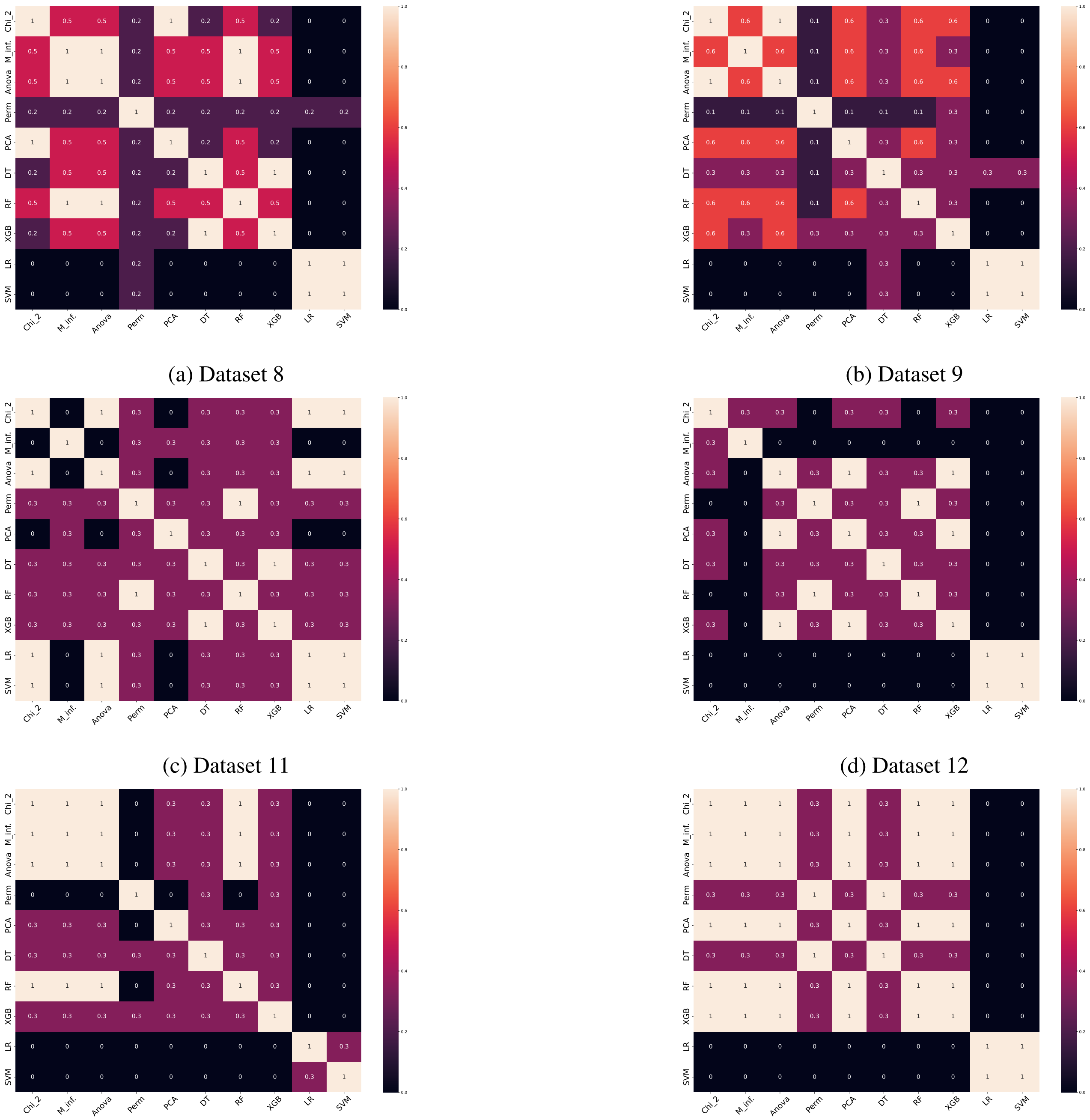
(c) Dataset 2

(d) Dataset 3

(e) Dataset 4

(e) Dataset 7

Figure 9: Jaccard similarity for the subset of features selected with feature selection methods and machine learning models.



(a) Dataset 8

(b) Dataset 9

(c) Dataset 11

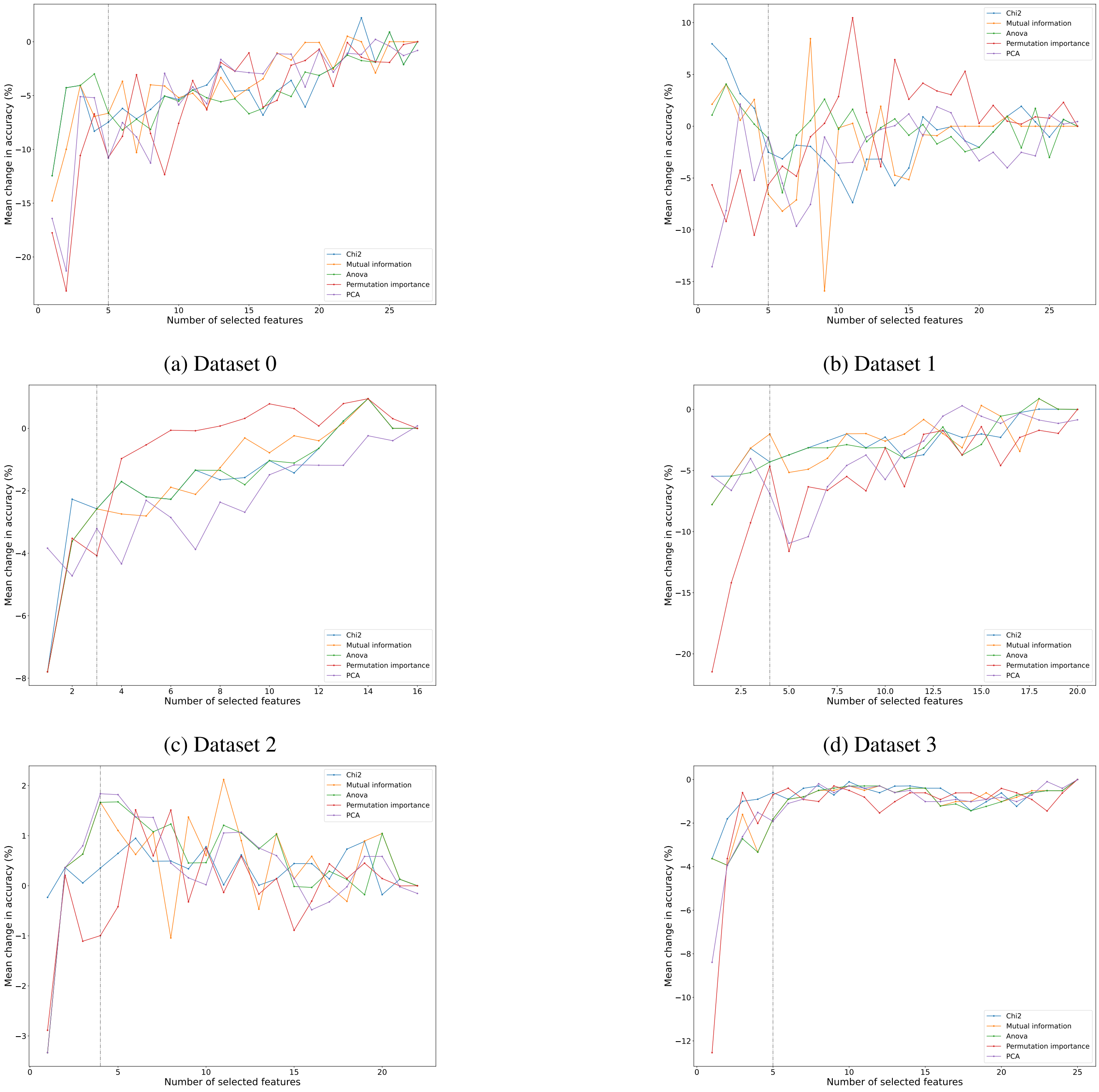
(d) Dataset 12

(e) Dataset 13

(e) Dataset 14

Figure 10: Jaccard similarity for the subset of features selected with feature selection methods and machine learning models.

B. Appendix - Mean relative change in accuracy



(a) Dataset 0

(b) Dataset 1

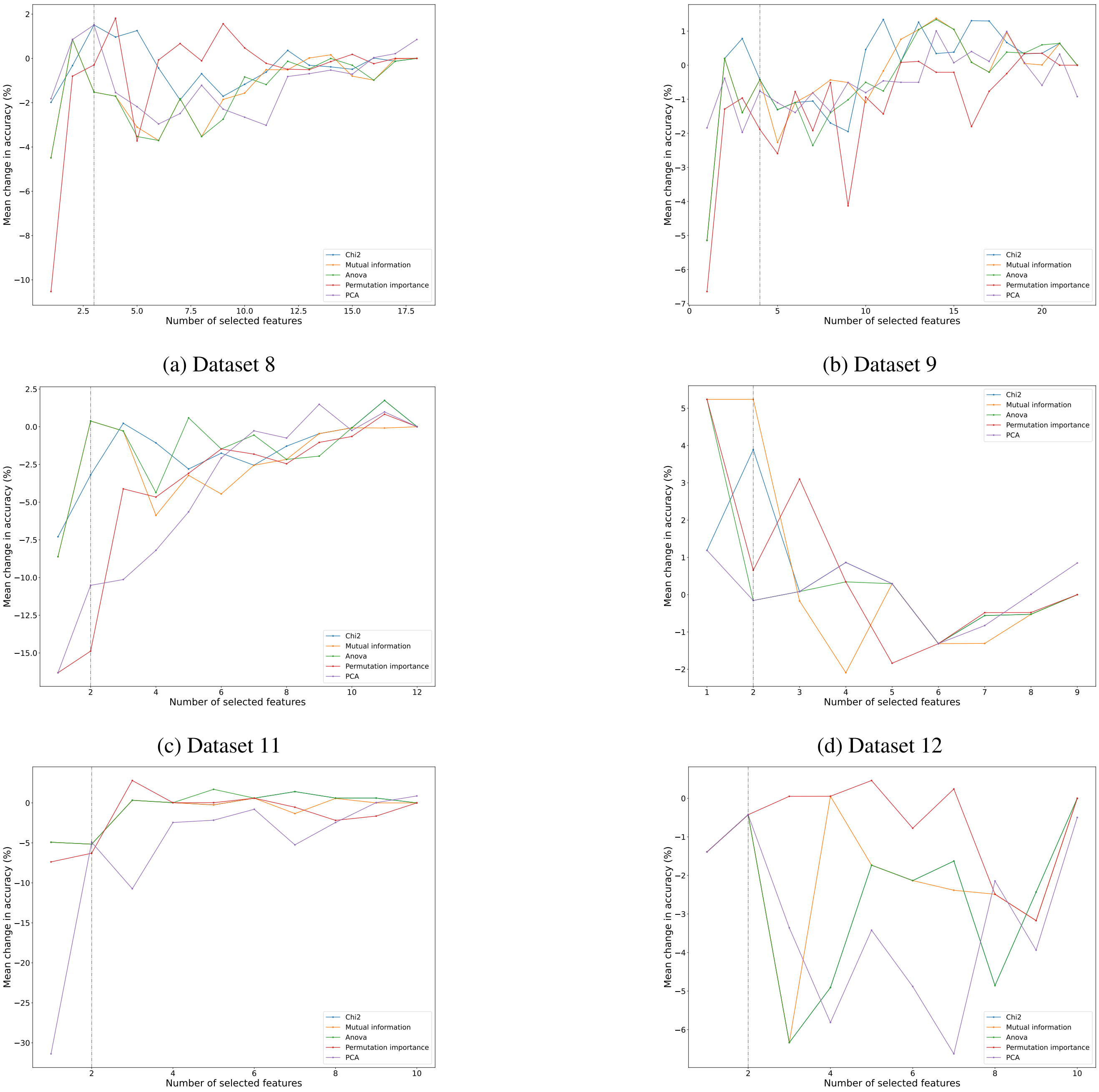
(c) Dataset 2

(d) Dataset 3

(e) Dataset 4

(e) Dataset 7

Figure 11: Mean relative change in accuracy for a varying number of selected features. Vertical line indicates 20% of features.



(a) Dataset 8

(b) Dataset 9

(c) Dataset 11

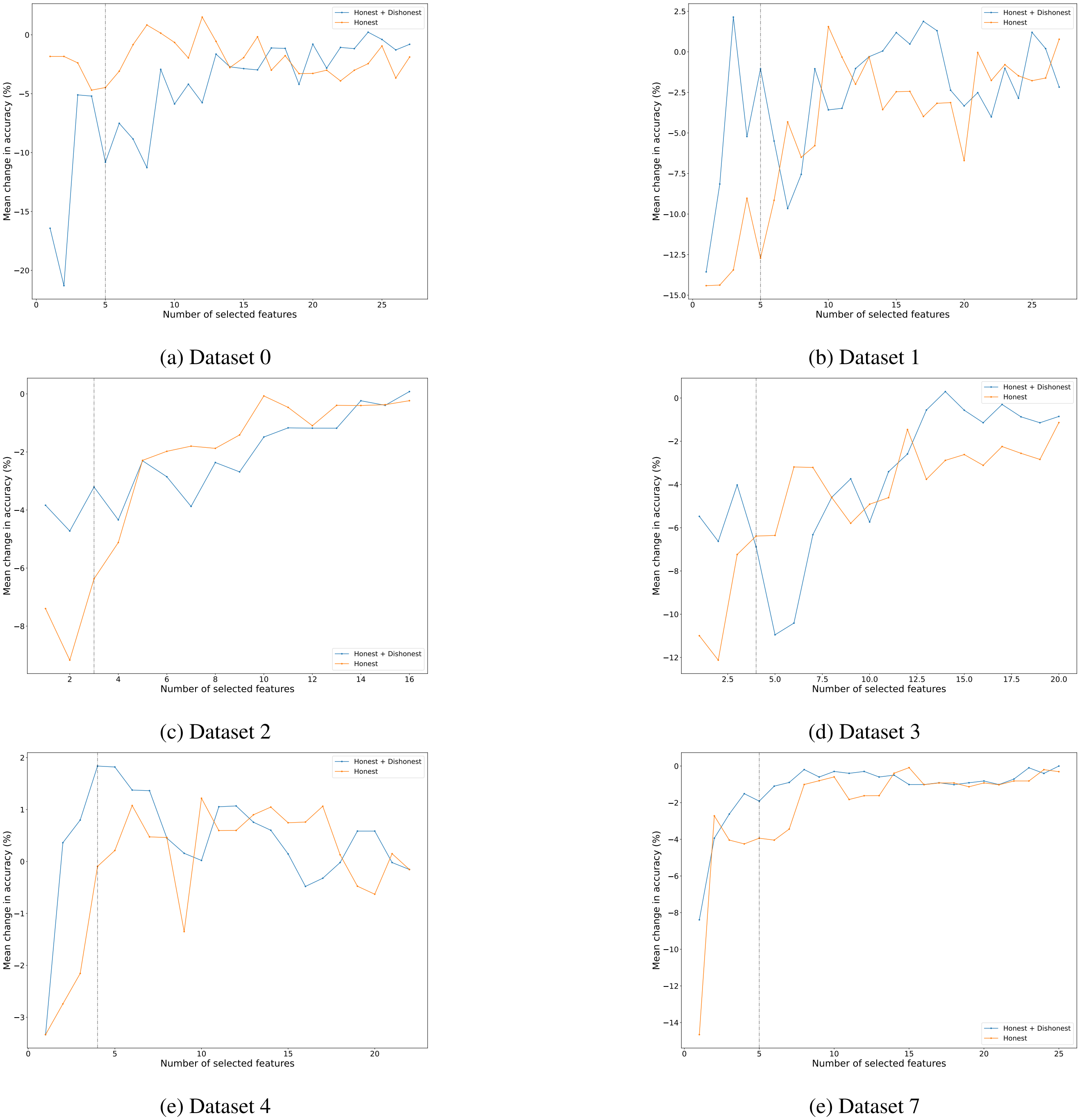
(d) Dataset 12

(e) Dataset 13

(e) Dataset 14

Figure 12: Mean relative change in accuracy for a varying number of selected features. Vertical line indicates 20% of features.

C. Appendix - PCA comparison



(a) Dataset 0

(b) Dataset 1

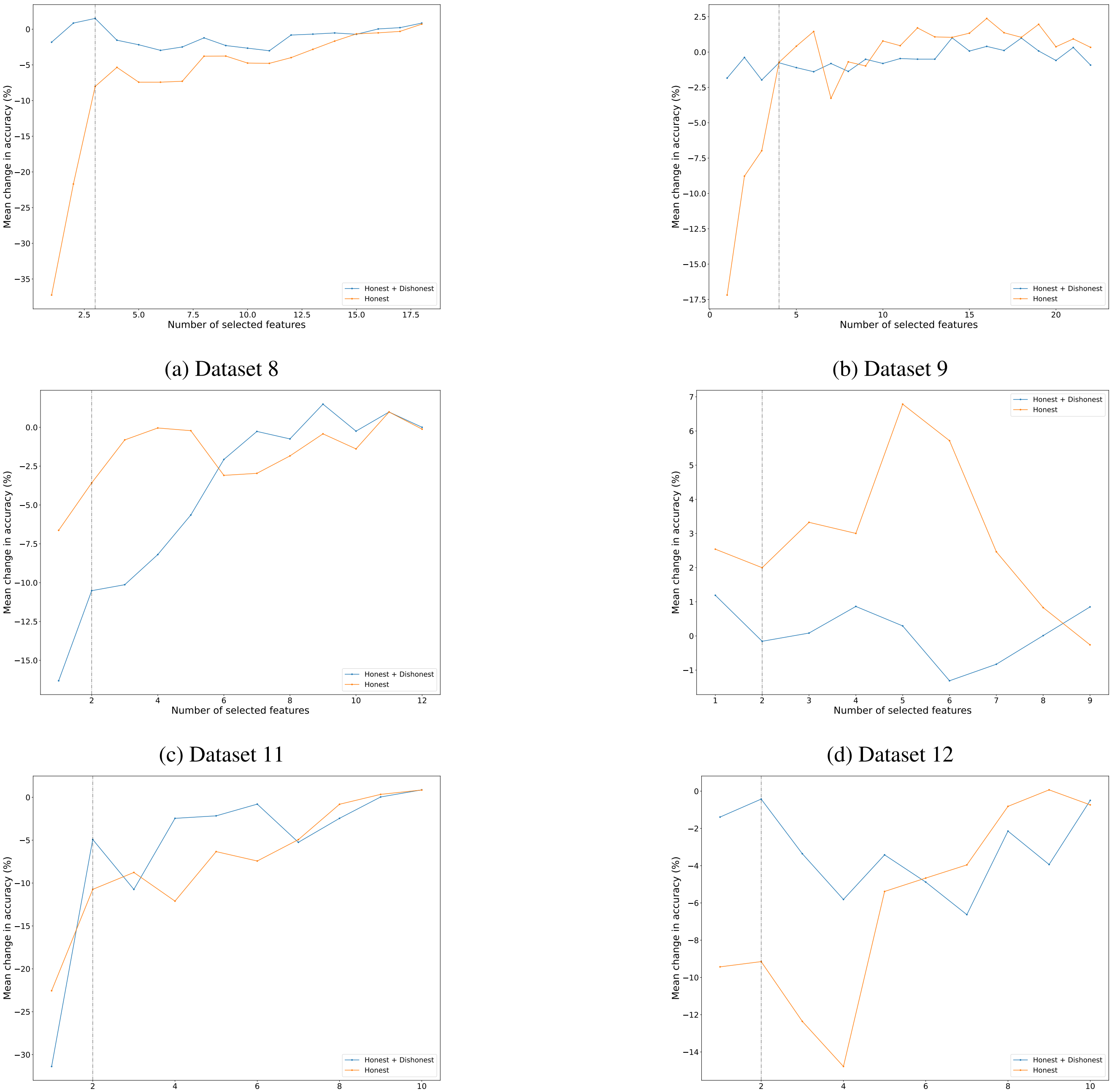
(c) Dataset 2

(d) Dataset 3

(e) Dataset 4

(e) Dataset 7

Figure 13: Mean relative change in accuracy for a varying number of selected features. Comparison of the use of PCA on all the dataset and on the honest answers. Vertical line indicates 20% of features.



(a) Dataset 8

(b) Dataset 9

(c) Dataset 11

(d) Dataset 12

(e) Dataset 13

(e) Dataset 14

Figure 14: Mean relative change in accuracy for a varying number of selected features. Comparison of the use of PCA on all the dataset and on the honest answers. Vertical line indicates 20% of features.