Cognitive, Behavioral and Social Data Project
Dark Triad Dirty Dozen

Nikoloska, Nora
nora.nikoloska@studenti.unipd.it

Reddavide, Matteo
matteo.reddavide@studenti.unipd.it

Aslanova, Ellada
ellada.aslanova@studenti.unipd.it

January 2022

1 Introduction

Dark Triad is a term used to describe a set of three personality traits that has attracted empirical attention in personality, social and clinical psychology. It is composed by:

- Machiavellianism, characterized by strategic, cold, selfish and manipulative behavior in laboratory and real world settings.
- Psychopathy, associated with high impulsivity and reckless antisocial behavior, thrill-seeking along with low empathy and anxiety.
- Narcissism, characterized by an unrealistically positive self-view, feelings of entitlement, and a lack of regard for others.

The Dark Triad Dirty Dozen [1] is a reduced questionnaire derived by the original 91 items, it is composed by 12 questions and each trait is investigated by 4 questions. The advantages of this new measure are both technical, reducing the response bias of each measure, and practical, the participants are less subjected to fatigue due to the length of the questionnaire. In this report we show the results obtained for the classification task that tries to differentiate between the honest and dishonest responses, as well as the reconstruction of the predicted dishonest responses. The code used for the analysis, as well as the modified datasets are available on Github.

1.1 Dataset description

The dataset we have worked on was based on 493 participants and they were asked to answer the Dark Triad Dirty Dozen questionnaire twice, once honestly and once faking good, so lying in order to minimize the Dark Triad's traits. The answers' values are distributed with a Likert scale where the values are between 1 and 5.

2 Exporatory Data Analysis

The first step in the analysis is trying to gain a deeper understanding of the dataset by computing answers distributions per question, comparison between honest and dishonest responses and principal component analysis (PCA).

As it can be seen see from Figure 1, for almost all the questions the dishonest answers have a higher density of answers for lower values (1 or 2) while the honest answers are more sparse and distributed along all the possible values.
This result is clearly visible also in the Figure 2 where the average answer for each answer is plotted separating the honest and the dishonest again. As predicted the dishonest line is always below the honest one.
One other important result we derive from this plot is how close the two lines are. This shows that even the honest answers are subjects of a certain degree of lying, there is a tendency also in the honest case to minimize the Dark Triad traits.

2.1 Principal Component Analysis

In the attempt to reduce the feature space by using PCA, the behavior of the Explained Variance with the number of components is computed. Here in Figure 5 the results for the two and three components cases are plotted, from which it can be seen that the samples are not easily separable with Total Explained Variance of $56,88\%$ and $65,29\%$, respectively. From the plot c). it's clear that in order to achieve an higher value for the TEV we need a higher number of PCA components need to be considered (good results with 10 or more).

The completed correlation matrix shown on Figure 4 d). shows that there is higher correlation between the questions that belong to the same group, particularly in the case of Machiavellianism. The first question shows peculiar low correlation behavior and it is the only one whose responses are reversed. Additional correlation matrices have been computed to show the relationship between honest and dishonest responses. It can be seen from the matrix c). that the correlation between honest and dishonest responses is very low which suggests a the task differentiation is hard.

2.2 Outlier detection

To compute the qualitative outliers - the participants whose answers do not change a lot when giving honest and dishonest answers, the average sum of difference in answers was computed as shown on the histogram on Figure 3 with the threshold of 5 shown as a red line. 9 participants whose sum of difference in answers was less than 5 were removed from the dataset. Their indices are: 34, 78, 138, 188, 214, 312, 347, 413 and 470 along with their dishonest answers.

3 TF-IDF

In this phase an attempt is made in order to normalize and transform the data, so that the accuracy of the model improves. Based on the structure of our questionnaire each subject is represented each subject in 3D space, where each dimension is assigned to a trait of the Dark Triad.

In Figure 6 the results are reported: in (a) a simple average of the answers for each trait, in (b) answers are weighted according to Cronbach's alpha, which measures the reliability of a question, As it can be seen even after these two transformations the two groups are not distinct. The last attempt was done based on a TF-IDF like transformation:

$$Tf = \frac{n_{col}}{\text{column length}} \tag{1}$$

$$Idf = log_{10}\left(\frac{\text{total columns}}{n_{glob}}\right) \tag{2}$$

As Tf we used the ratio between the number of appearances of an answer in a column (i.e. how many times that value was used in the specific question) while the Idf was the logarithm of the ratio between the total number of columns (the number of questions in our case) and the number of appearances of the value in all the questions.

The improvement is visible, there is a distinct cloud for the honest subjects while the dishonest are more sparse. To further investigate the data, the dishonest answers are colored by the number of changed answers from their honest profile. By doing this a predictable effect is spotted: the subjects difficult to classify (i.e. the closest to the honest cloud) are almost all green, and they changed their answers only on a few question.

4 Classification

The models used to differentiate between the honest and dishonest responses are: Logistic Regression, SVMs, Random Forest, Naive Bayes, Multilayer Perceptron and XGBoost. Two different datasets are used for the classification task: the raw data scaled accordingly and the adjusted TF-IDF data. First, the benchmark accuracy is computed for both datasets and it is 75% for the original data and 91% for the TF-IDF adjusted data. The benchmark is computed by calculating the mean vector between all samples in a training subset and then classifying a given sample as honest if it is on average higher than the mean and as dishonest otherwise.

4.1 Logistic Regression Classifiers

To detect the optimal number of variables for a logistic regression model, experiments with different number of variables are shown on Figure 7 a). For each number of features, the optimal features were selected. The variance comparison for these models is shown on Figure b). The increase of variables does not significantly increase the accuracy, but it does increase the ROC score. For this reason all variables are used in experiments with other models. Another experiment is the removal of all Machiavellianism features becase of the high correlation. The obtained results and variance from all of the classifiers are reported on figure 8. Since no significant increase in accuracy has been detected for the classifiers influenced by correlated data, all features are used in the further analysis.

4.2 Comparison between classifiers

All classifiers have been tested both the original scaled data and the modified TF-IDF data. The the results of 10-fold cross validation (for the original data) are shown on Figure 9 a) and the variance is shown on Figure b). The best model for the classification with the original data is the Random Forest classifier with 10-fold cross validation accuracy of 82%. The chosen model has been used on the test data with accuracy of 84% to finally generate the predicted samples. All of the models trained on the adjusted TF-IDF data reach very high accuracy, but the XGBoost correctly classified all of the test samples.

In order to understand the behavior of the misclassified items, the mean responses for the correctly and incorrectly classified items are shown on Figure 11. Although the distributions are very similar, the difference can be seen in the responses of questions M4 [I have used flattery to get my way] and M11 [I tend to exploit others towards my own end].

5 Malingering Remover

The next step is to reconstruct the honest answers starting from the ones classified as fake in the previous section. For this purpose two methods were tested: multi-output regression and denoising autoencoder. This last option was then discarded since the dimension of the dataset is small and the error of the process was then too high.

The test set samples are the subjects flagged as dishonest in the classification phase (95 out of 493) while the training set was composed by all the other participants. In one case (subject 278) both the honest and the dishonest answers were classified as dishonest so they were both considered in the reconstruction, once starting from the honest and once from the dishonest.

First of all a grid search with 5 fold cross validation is set up to tune the best hyperparameters for each of the following regressors: Linear Regression, K Neighbors, Decision Tree, Random Forest, Support Vector, Multi Layer Perceptron and Ridge.

The metrics used to evaluate the models are the Root Mean Squared Error, the Mean Squared Error and the Mean Absolute Error. The results are reported in Table 13, as it can be seen the regressors perform in a similar way, with values of the RMSE around 0.26. The best regressor is the Multi Layer Perceptron with a value of 0.260301 for the RMSE.

In Figure 12 the average performance of each regressor in the reconstruction task is reported and it is compared with the benchmark method as a trivial method of reconstruction. First, a vector of average differences between honest and dishonest answers is computed from 90% of the samples. This vector is then added to the dishonest answers of the remaining 10% of samples to reconstruct the honest response and compute the average MSE which is 0.07.

Again it can be seen that the regressors performs almost the same, with the exception of the Support Vectors. This difference is clear also in the reconstruction similarity plot (Figure 14) where the mean similarity with the honest responses for each question is reported: the SVR achieves a higher value only for two questions (P3 and P10) while in the other 10 it is far worse. Based on tis plot the best model is the K Neighbors which achieves a value of 87.84%. Since the computed errors are similar, but K Neighbors regressor is more interpretable, it is suggested as the chosen regressor for the reconstruction of the honest answer.

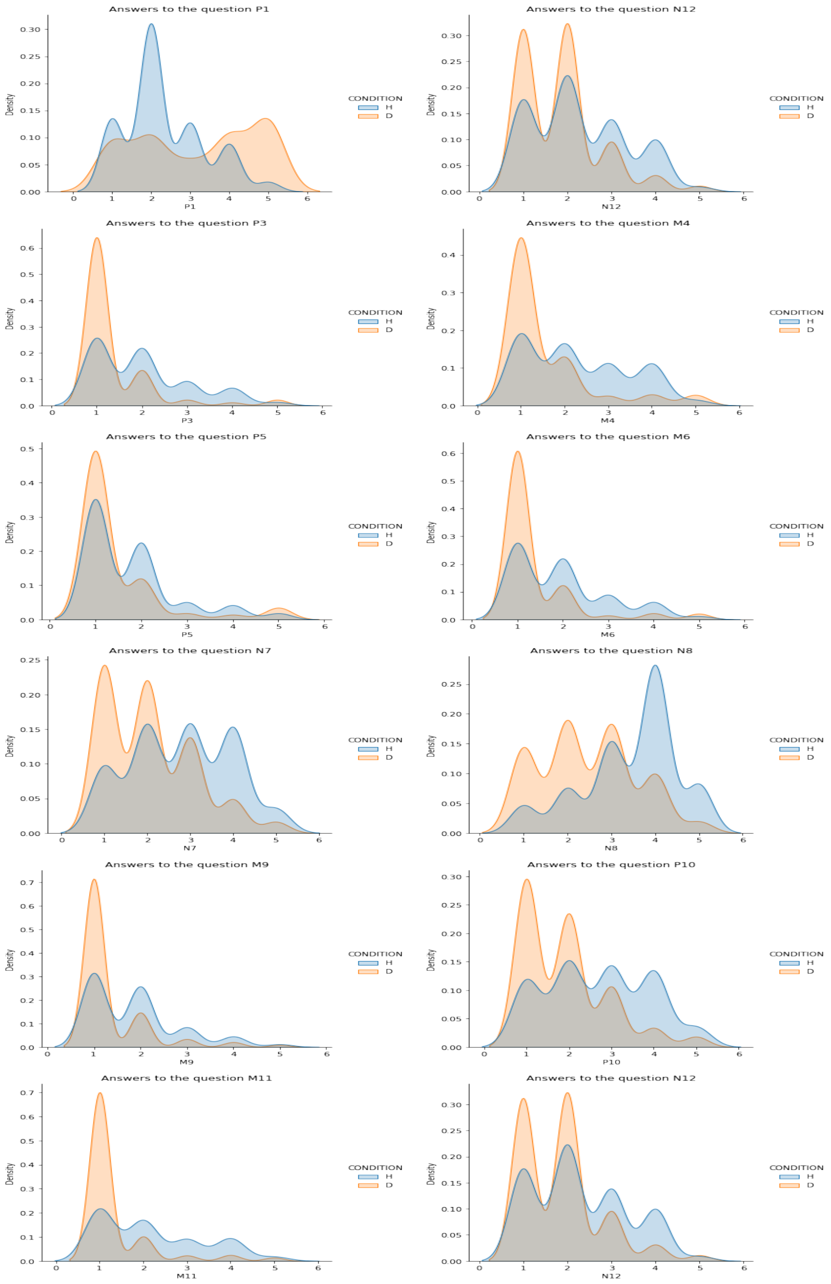
6 Conclusion

In this project we show the results from the classification between honest and dishonest dark triad questionnaire responses as well as reconstruction attempts to remove malingering. From the analysis of the data, it can be seen that the even though the honest responses are on average higher than the dishonest, the answers overlap and the behavior is similar which makes the differentiation task harder. The classification results obtained on the test set for the scaled data reach 84% accuracy, while the computed benchmark is 75%. Better results are obtained when using the adjusted TF-IDF dataset where the classes become more easily separable with accuracy values reaching 100% for some models, while the computed benchmark id 91%. Other metrics, such as the variance were computed for more accurate comparison between the models.

In the reconstruction phase, several models have been tested, from which some managed to generate reconstructed samples with MSE close to 0.067, whereas the computed benchmark error is 0.07. More sophisticated methods for reconstruction like the autoencoder did not generate satisfactory results.

We hope that these results can be used as a starting point for more advanced methods or for similar types of analysis for different datasets.

7 Figures


Answers to the question P1

Answers to the question N12

Answers to the question P3

Answers to the question M4

Answers to the question P5

Answers to the question M6

Answers to the question N7

Answers to the question N8

Answers to the question M9

Answers to the question P10

Answers to the question M11

Answers to the question N12

Figure 1: Answers to each question for honest and dishonest condition

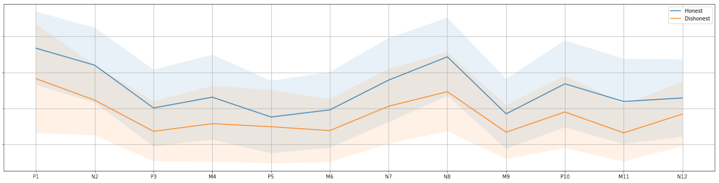


Figure 2: Average answers of both groups for each question

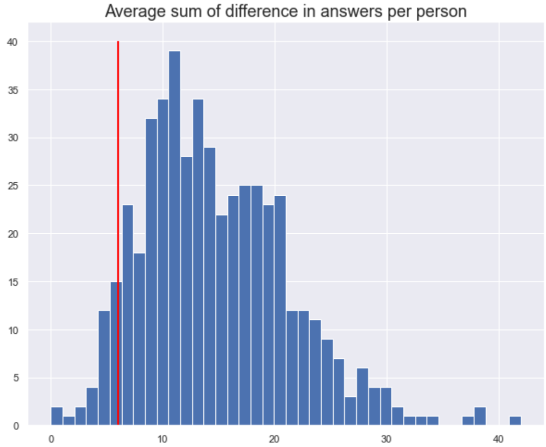
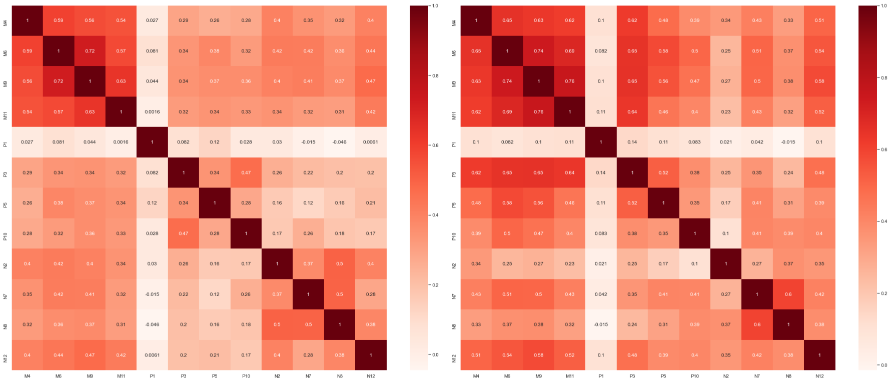
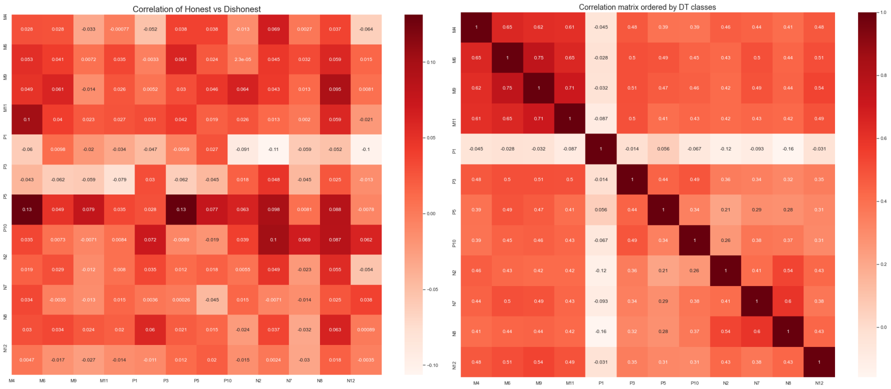


Figure 3: Detecting qualitative outliers



(a) Honest responses

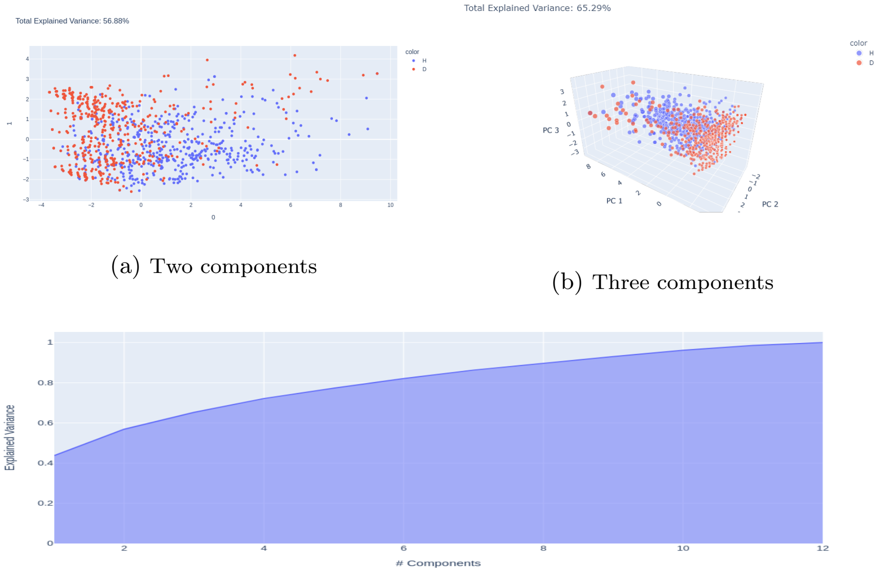
(b) Dishonest responses



(c) Honest vs dishonest responses

(d) Complete correlation matrix

Figure 4: Correlation matrices

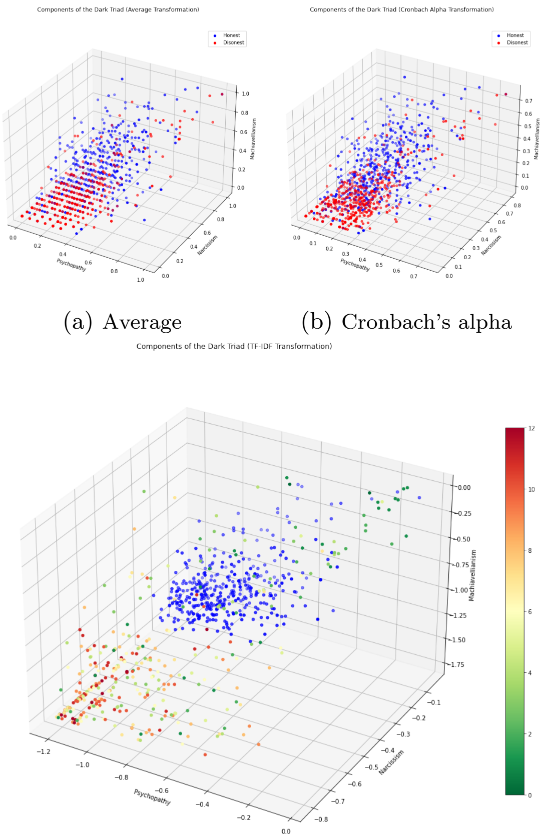


(a) Two components

(b) Three components

(c) Explained Variance against number of components

Figure 5: PCA plots

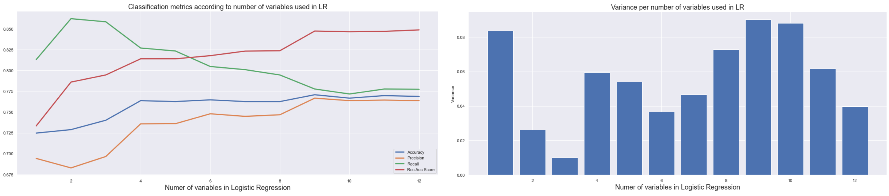


(a) Average

(b) Cronbach's alpha

(c) TF-IDF

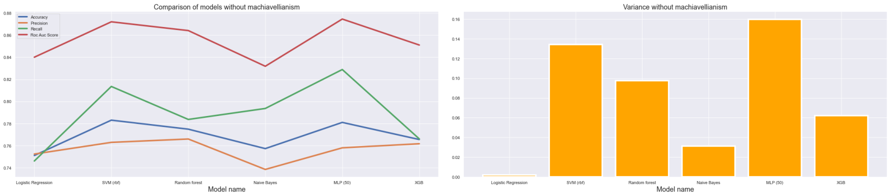
Figure 6: TF-IDF adjusted data



(a) Comparison of results

(b) Comparison of variance

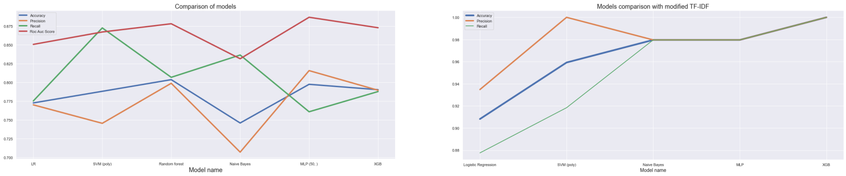
Figure 7: Learning regression models with different number of variables



(a) Comparison of results

(b) Comparison of variance

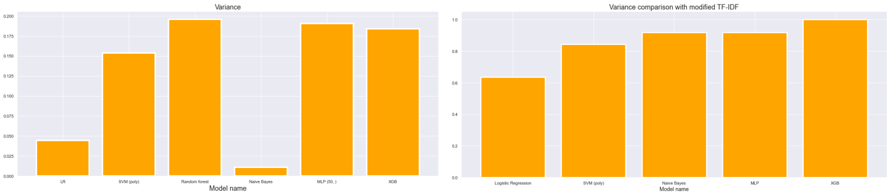
Figure 8: Models without machiavellianism



(a) Original data

(b) Adjusted TF-IDF data

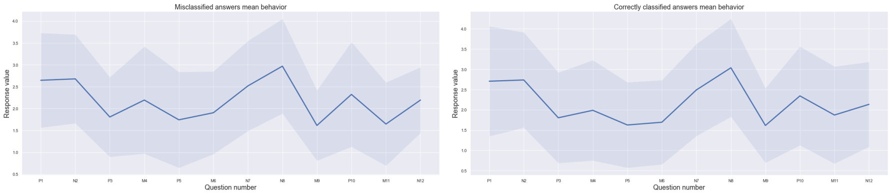
Figure 9: Comparison between classification models



(a) Original data models variance

(b) TF-IDF models variance

Figure 10: Variance comparison between models



(a) Misclassified responses

(b) Correctly classified responses

Figure 11: Comparison of the average response behavior

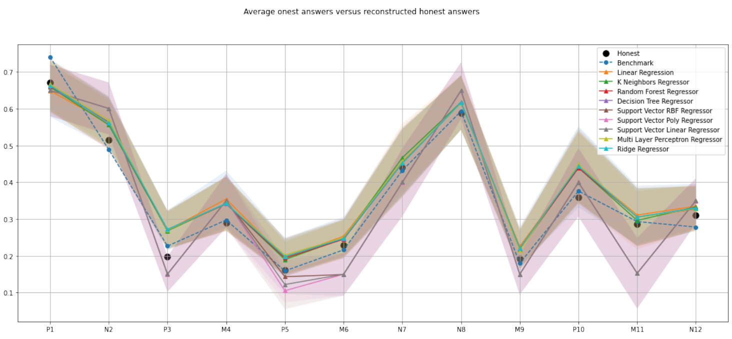
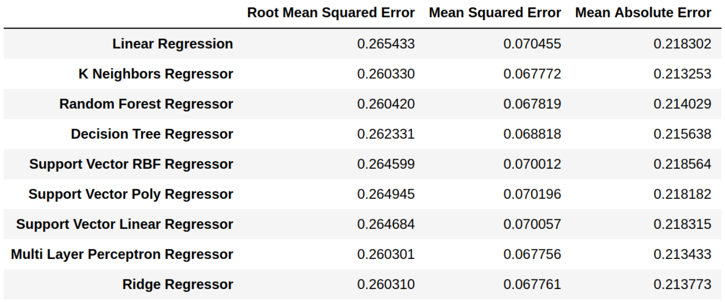


Figure 12: Reconstruction of the honest profile



	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error
Linear Regression	0.265433	0.070455	0.218302
K Neighbors Regressor	0.260330	0.067772	0.213253
Random Forest Regressor	0.260420	0.067819	0.214029
Decision Tree Regressor	0.262331	0.068818	0.215638
Support Vector RBF Regressor	0.264599	0.070012	0.218564
Support Vector Poly Regressor	0.264945	0.070196	0.218182
Support Vector Linear Regressor	0.264684	0.070057	0.218315
Multi Layer Perceptron Regressor	0.260301	0.067756	0.213433
Ridge Regressor	0.260310	0.067761	0.213773

Figure 13: Results of each regressor

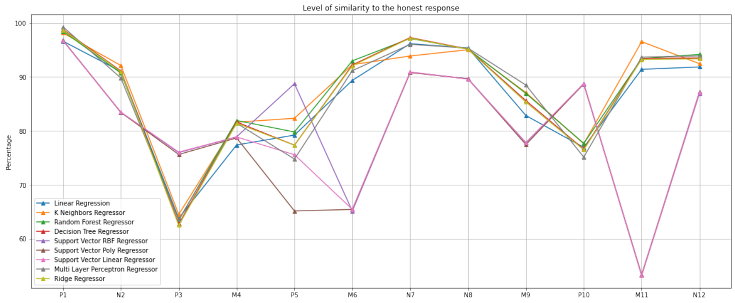


Figure 14: Performance of each regressor in the reconstruction

References

[1] Peter K. Jonason. "The Dirty Dozen: A Concise Measure of the Dark Triad". In: *Psychological Assessment* 22.2 (2010), pp. 420–432. DOI: http://dx.doi.org/10.1037/a0019265.