Università degli Studi di Padova

A.Y. 2022/2023

Department of Mathematics

STUDENTS

Angelica Giangiacomi
Anthony Palmieri

Cognitive, Behavioral and Social Data - Project

A.Y. 2022/2023

Overview

I Database: 13 groups of datasets, totalling 20

II Goal: build models that are able to distinguish between Honest answers and Dishonest ones

→ Find the most relevant features to identify this distinction

III Preprocessing and Data Cleaning

IV Exploratory Data Analysis

V Training Phase

→ Unsupervised Learning: PCA, Sparse PCA, Variance Threshold Feature Selection

→ Supervised Learning: Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost

→ Greedy Backward and Forward Features Selection strategies, based on p-values, Information Criterion and accuracy.

Preprocessing and Data Cleaning

- Check for NA and NaN values

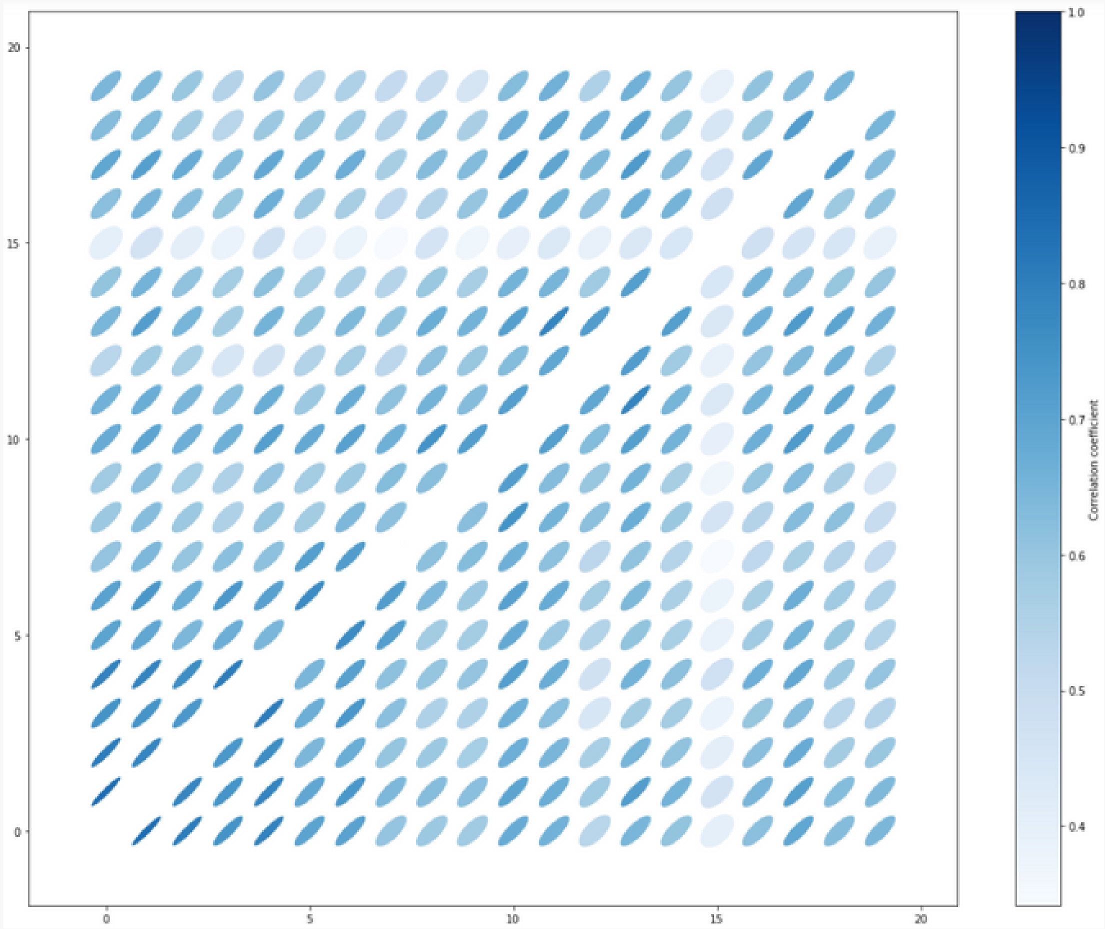
- Check for categorical features and eventually apply one-hot-encoding

- Split the chosen dataset into Training (80%) and Test (20%) sets

Exploratory Data Analysis

- Check for possible correlations and multicollinearity

Correlation ellipses



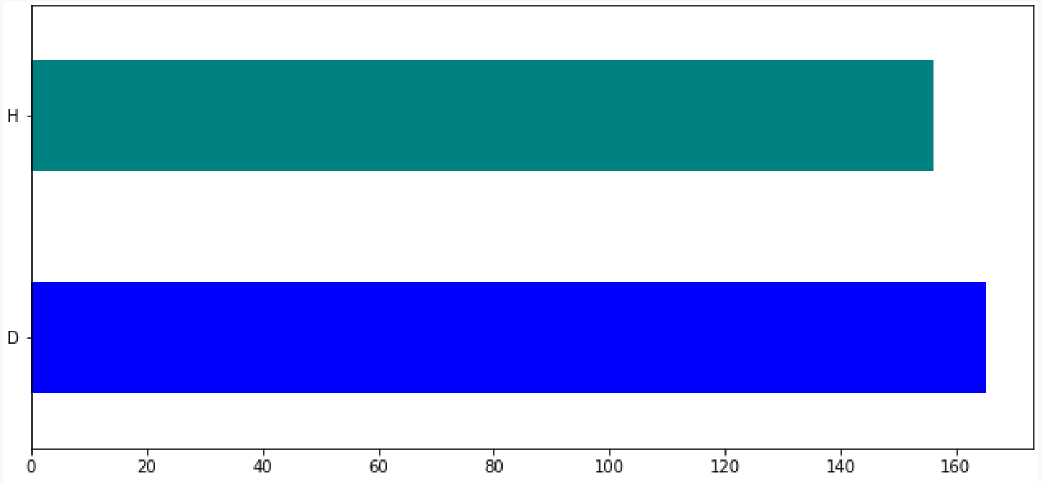
Eventually drop highly correlated predictors

Compute VIFs

Eventually drop predictors with a value greater than ~~5~~ 4 ✓

- Scale the data

- Check for Class Balancement



Random Undersampler ✓

Eventually apply ~~SMOTE~~

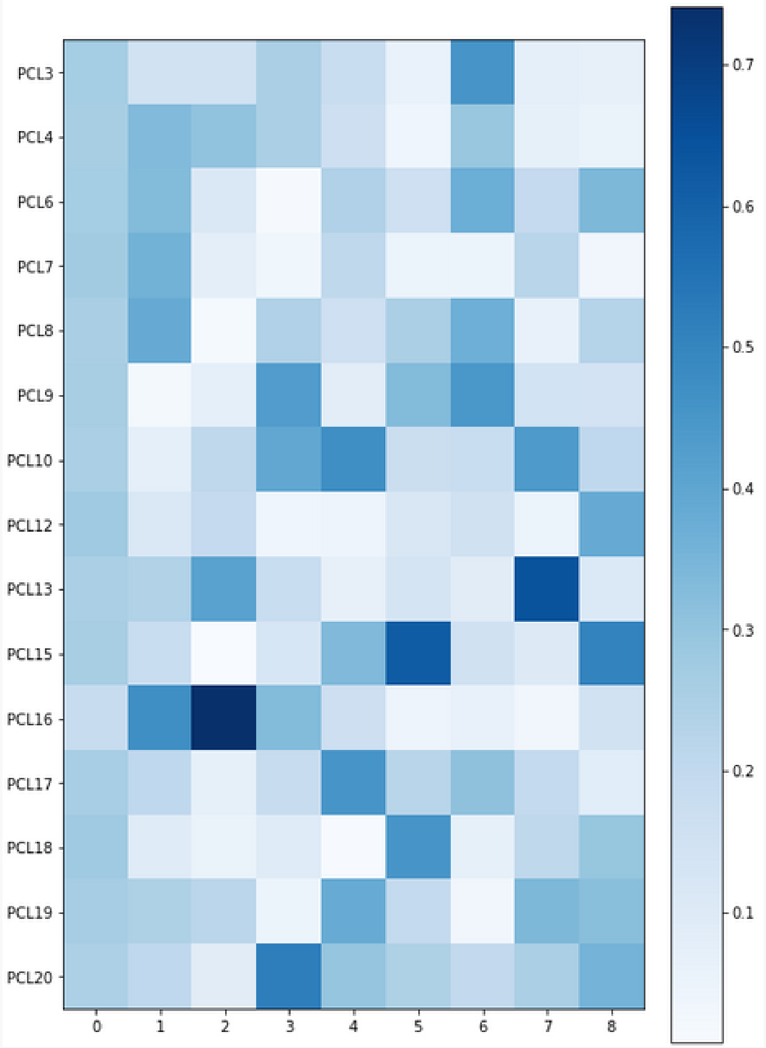
Training Phase: Unsupervised Learning

We apply all the three models first on the whole dataset, then keeping Honest and Dishonest separate; a set of the most relevant features is selected by each model.

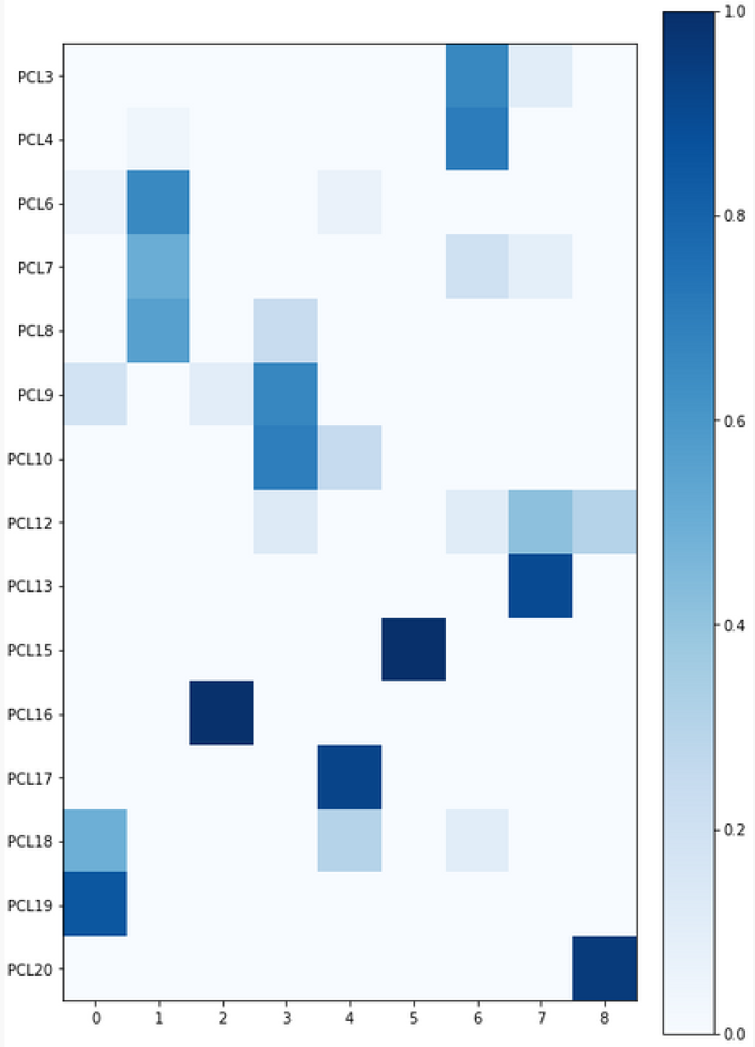
Then, for each of the cases (whole, H and D) we take the union of the sets of the selected predictors.

Finally, we apply a Logistic Regression to see the behavior of the features selected among the models.

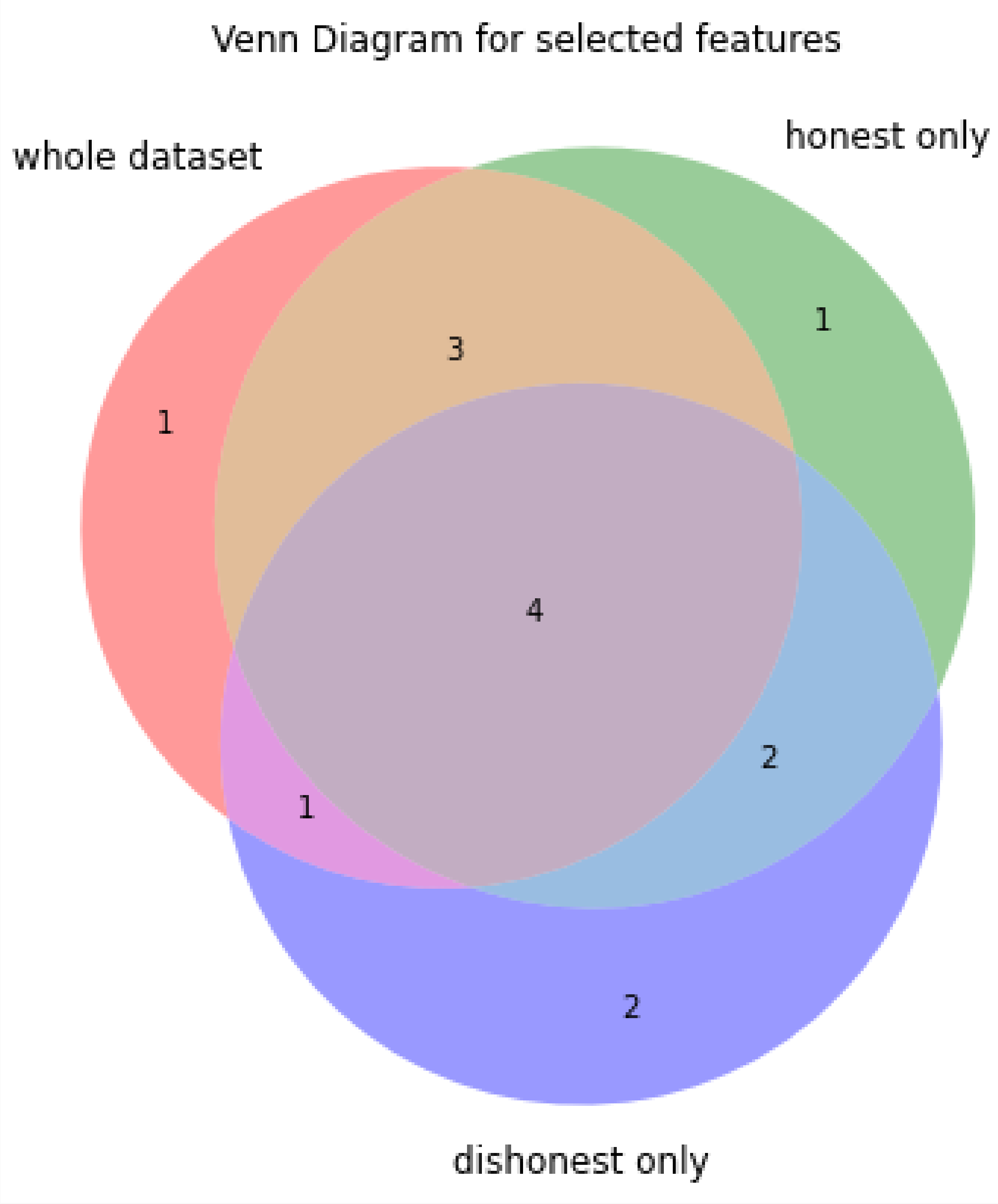
Example of some results



Importance of the features for each PCA component (whole)



Importance of the features for each Sparse PCA component (whole)



→ Features selected in all three cases:

- PCL4
- PCL7
- PCL12
- PCL18

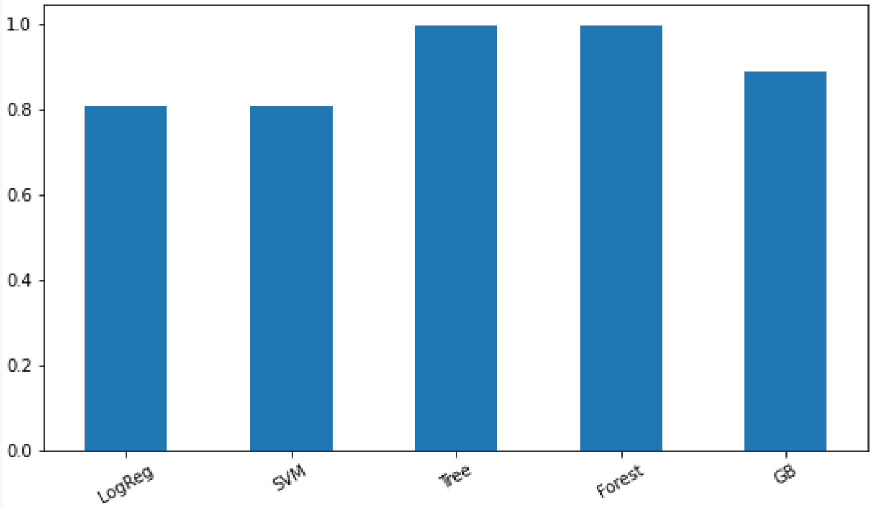
Training Phase: Supervised Learning

We apply each model by starting to fit it on the Training set and then evaluate it on the Test set.

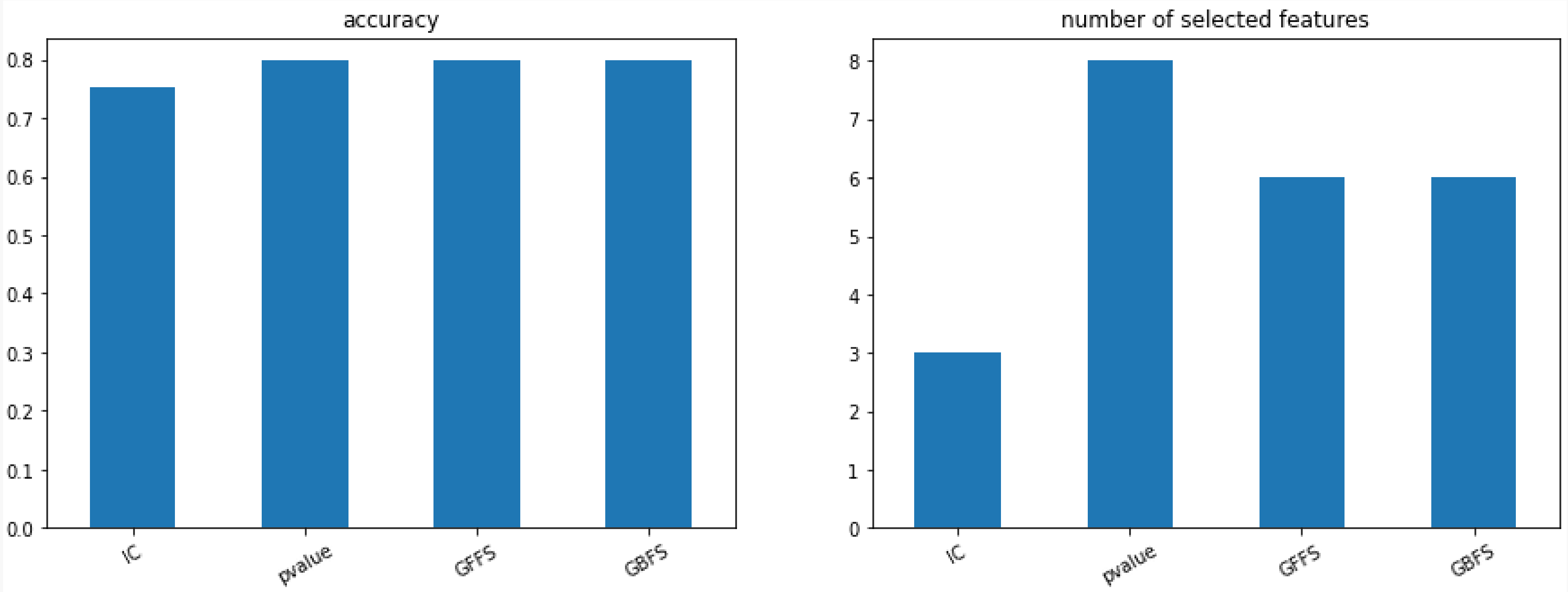
After that, we apply the Permutation Importance and we select the set of the most relevant features as those ones having:

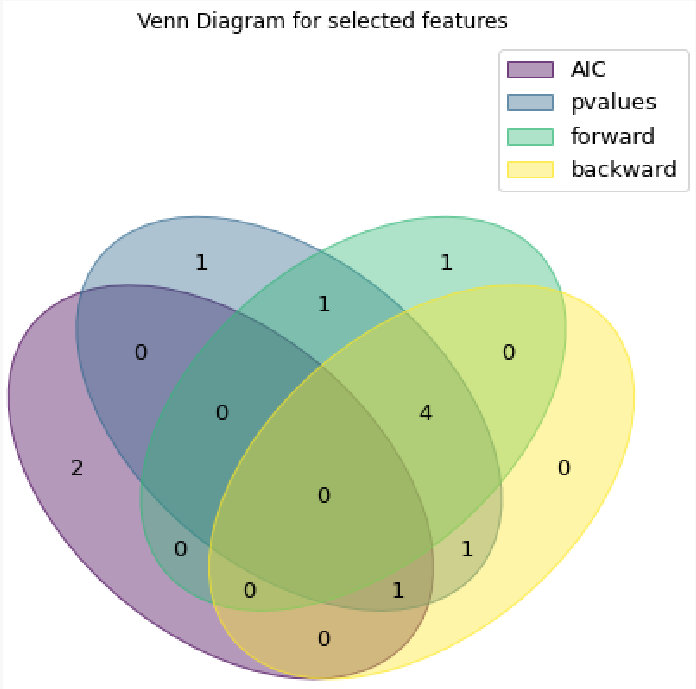
- both |Coeff.|>0 and $P.I.Mean > 0$, in the case of Logistic Regression and SVM;
- both the highest Importance features value and $P.I.Mean > 0$, in the case of the Tree-based algorithms.

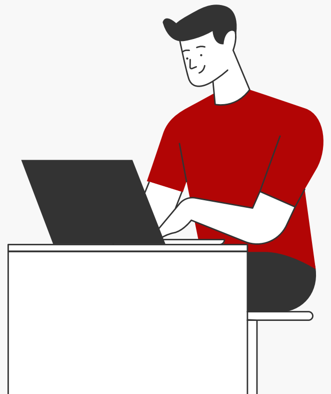
Accuracy obtained through each model



As in the case of Unsupervised models, we take the union of the features selected by each model and then we apply some Greedy Backward and Forward Selection techniques, trying to filter out other predictors.





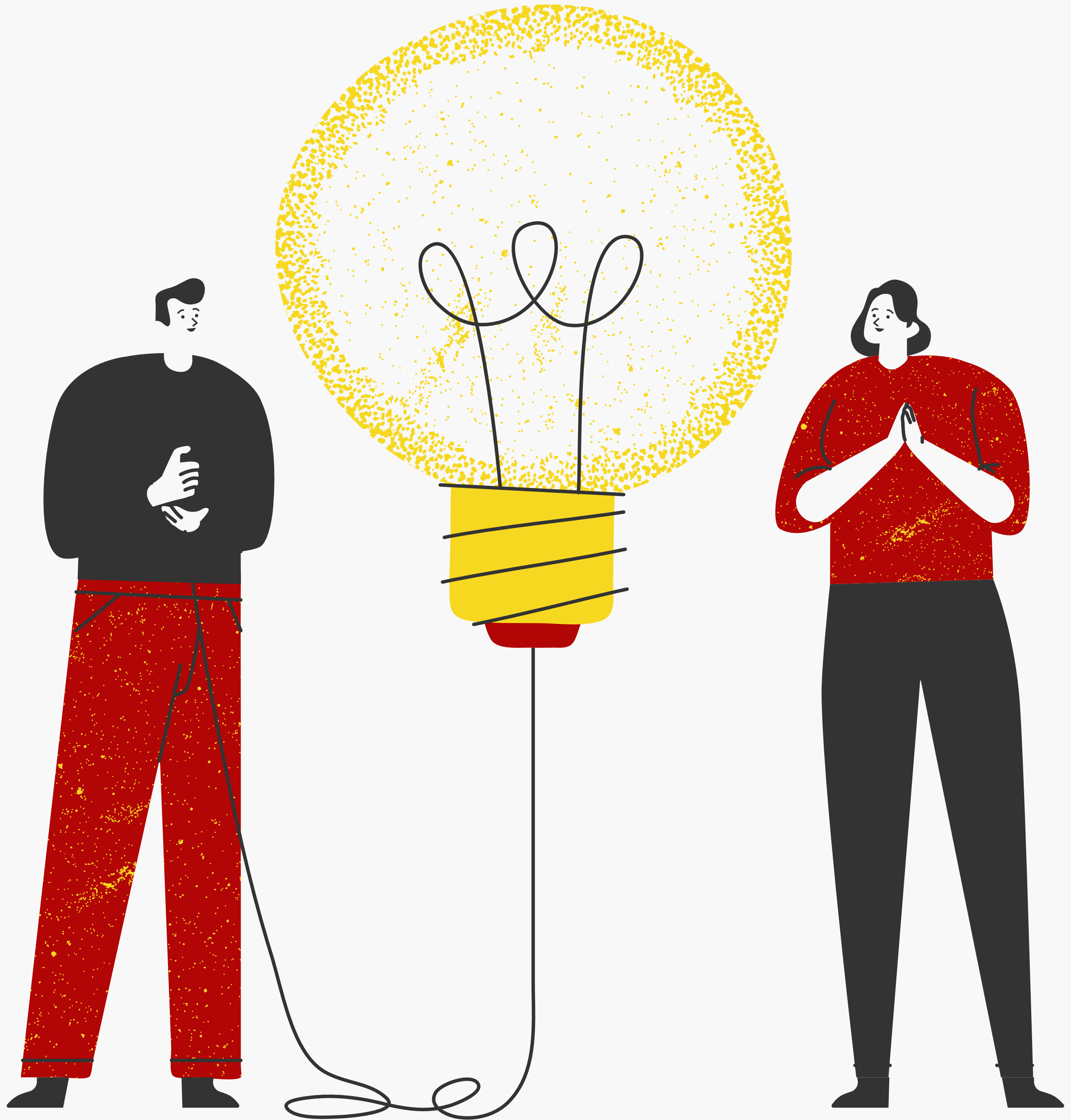


Further Steps

- Analyze whether the implemented models perform well with all the datasets or try to understand the reason why they fail in some contexts.

- Study the differences when using the whole Dataset or keep Honest and Dishonest separate

- Try to explain the reason why some features are more relevant then others.



Conclusions

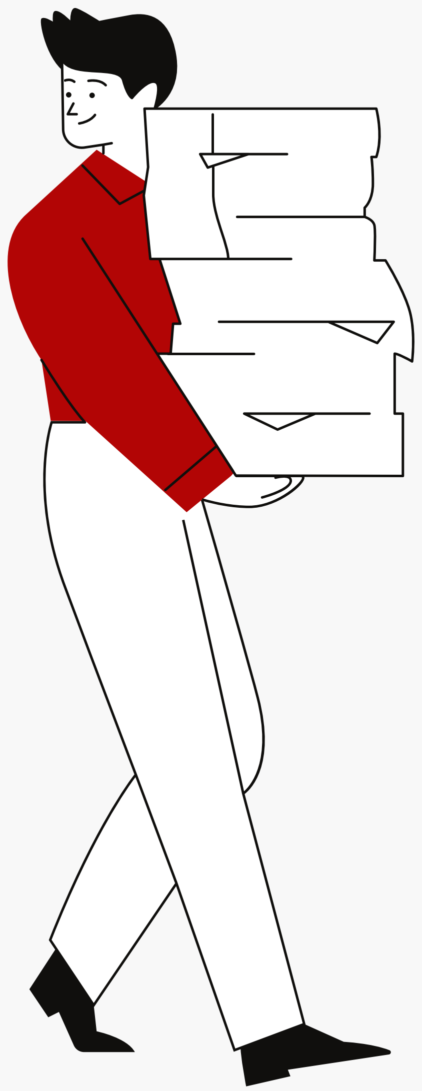
- The results obtained on each of the datasets are reported in the txt files created and updated while running the notebook.

- From what we could see, in general, our feature selection strategies seem to behave quite well, managing in some cases to determine a group of predictors that constitutes about 90% of the accuracy on the Test set.

- Other datasets on the other hand proved to be a bit more problematic, nonetheless our feature extraction strategy still manages to determine small sets of predictors that contribute up to 75% - 80% of the accuracy on Test set.

- It would also be interesting to focus more on each one of the dataset and implement some feature extraction strategies specific to faking good and faking bad separately.

- To conclude, this notebook represents just a starting point and there's a lot of space for improvement and new ideas.



Thanks for the Attention!