Using decision trees and random forests to predict cannabis consumption.

MSc Data Science - Jorge Rodríguez Peña – City University of London
INM431 Machine Learning

Supplementary material

GLOSARY

The dataset in the Project contained 7 different personality traits. The NEO Five Factor Inventory is a test consisting in 41 questions that measure 5 dimensions of personality scored from 12 to 60 [1]:

- Extraversion: Measures the behavior and preferences of a person in social interaction. Higher values are for more energetic people that really like social interaction.
- Agreeableness: Measures the friendliness of the individual and other factors such as loyalty and cooperation skills. Higher values are for friendlier, more loyal and more cooperative people.
- Conscientiousness: People with lower scores tend to be more unorganized, vague and more likely to give up on their goals.
- Neuroticism: People with high scores tend to feel more insecure and are more likely to collapse under stressful situations.
- Openness to experience: Measures the creativity and interest in artistical and cultural experience. People with lower scores are realistic and do not feel connection with art.

The other two tests included in the dataset are:

- Barratt Impulsiveness Scale [2]: Goes from 0 to 9 and reflects the individual's ability to think before making a decision or an action. Higher values mean more tendency to act without previous consideration.
- Sensation Seeking [3]: Goes from 0 to 10. Individuals with a higher value have a tendency to look for novel, risky or intense experiences.

MODELS AND FEATURE SELECTION

Due to the imbalance presented in the data, we presented two possible models, each one trained with a decision tree and a random forest to compare results. Figures 1 and 2 show the feature

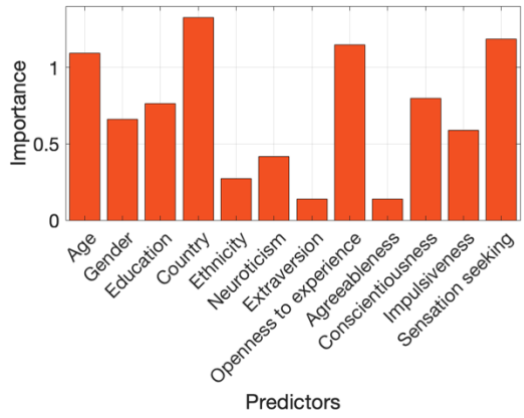

Figure 1. Feature importance for the first model using random forest.

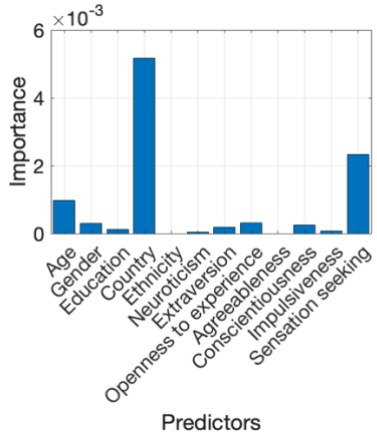

Figure 2. Feature importance for the second model using decision trees.

importance for the random forest and the decision tree. We see that country tends to influence the models highly, while about 90% of the samples are either UK or USA. This sometimes led models to classify a UK citizen as user/non-user without other considerations.

As shown in Figures 1 and 2 Ethnicity however did not have such a great impact. But considering that about 95% of the data represent white people, other ethnicities were misrepresented. Although our final decision tree did not use ethnicity at all (see Figure 2), this feature had some relevance in the first random forest. Looking at the data and plotting the percentage of users in each group, it was discovered that the Black-Asian ethnicity only had users.

The conclusion was that these two variables were very biased and not representative of the population, therefore the implemented model could not be generalized and used for further predictions. Following the original paper [4], those variables (Country and Ethnicity) were removed for the second model.

CLEANING THE ORIGINAL DATASET

The original dataset was collected by Fehrman, E. et al. (2017) [4] and the data was transformed using categorical principal component analysis (CatPCA) [5]. For simplicity and better understanding of the data the original dataset was cleaned to have categorical values for each of the groups in Age, Gender, Education, Ethnicity and Country and the test scores for Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness to experience, Impulsiveness and Sensation Seeking.

This process was performed in the "cleaning.m" MatLAB file.

BAYESIAN OPTIMIZATION

Bayesian optimization seeks to maximize an objective function (in this case, the AUC of the model). It first selects random values inside the hyperparameter grid and evaluates the objective function. Then, seeks to maximize the acquisition function to select next point to evaluate the objective function. Then algorithm updates the objective function using Bayesian inference [6]. MatLAB's in-built function *bayesop* automatically does this process with 30 iterations.

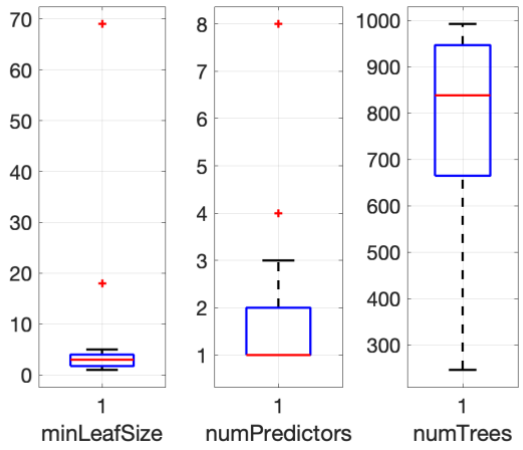

Figure 3. Results obtained running 25 bayesian optimization for the first model of random forest.

Due to the randomness of the process, we performed 25 Bayesian optimizations for each model. It still worked faster than grid search and provided very efficient results.

Figure 1 on the left shows the boxplots obtained after running Bayesian optimization 25 times for the first random forest model.

For all model, the median value of the resulted optimized hyperparameters distribution was chosen as the best hyperparameter. It was done this way to

avoid outliers such as those that can be seen in Figure 3 in minLeafSize or numPredictors. The same process was done for decision trees.

The number of trees used in the model had a very high variation, as shown in Figure 3 in numTrees. To avoid assigning an incorrect number of trees, minLeafSize and numPredictors were determined first and after that, different models with different number of trees were trained to evaluate the time and AUC obtained.

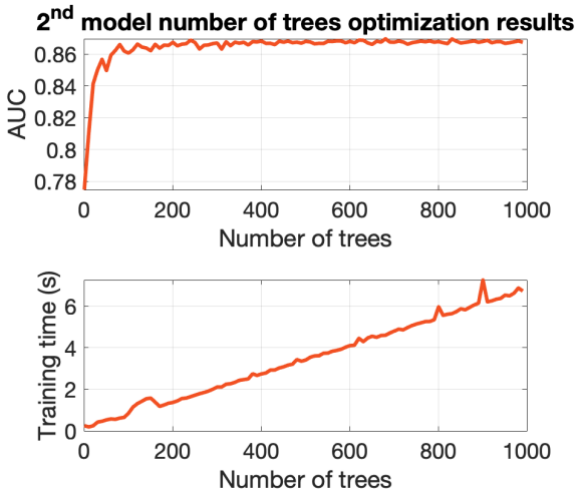

Figure 4. Number of trees and the AUC obtained and the time it took to train the model.

Results for both models (with and without Country and Ethnicity as variables) showed that around 600 trees the AUC did not vary although the training time increased. So 600 trees were used to trained both random forest models, in order to save time and still get good values of AUC.

For decision trees, parameter optimization using Bayesian optimization had very low variance with almost all 25 runs ending in the same optimal parameters that were then used in the models.

AUC scores and plots

ROC plots were not included in the poster as they do not provide useful enough information to our analysis. Instead, the analysis focused in the AUC as method of evaluation the performance of the model.

Figure 5 shows the ROC plots for each model obtained from the test set, with the AUC colored in blue. Here we can see how the area slightly decreases from the first model to the second, as explained in the poster. But these ROCs show that our models perform better than a random one, with the curve being above the straight line that goes from (0,0) to (1,1).

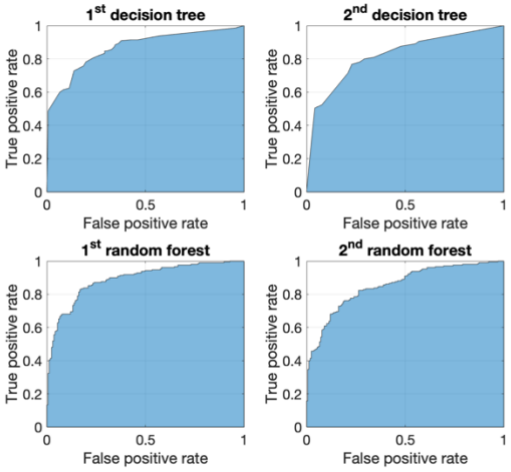

Figure 5. ROC and AUC for each of the models.

Confusion matrix

Other evaluation tool that could have been used in this project are confusion matrices. Multiple parameters can be extracted from them.

Figure 6 shows the confusion charts for each of the models. Again, we see how the percentage of accurately classified classes descends from the first model to the second, suggesting again that the dataset is very biased and not representative of a real-world situation.

Despite all that, the classification statistics extracted from the second model are still very good and show that the chosen personality traits can help predicting cannabis consumption.

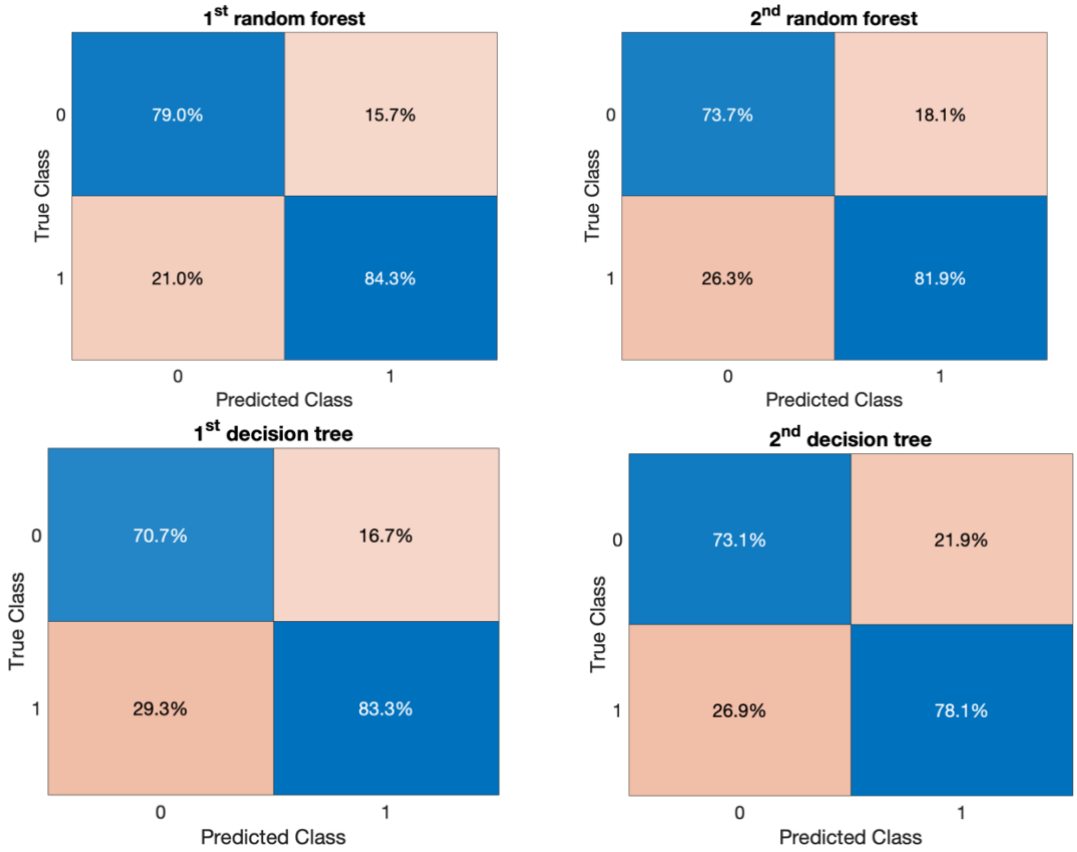

Figure 6. Confusion charts for all the models

REFERENCES

[1] Robert R. McCrae, Paul T. Costa, 'A contemplated revision of the NEO Five-Factor Inventory'. Personality and Individual Differences, Volume 36, Issue 3, 2004, Pages 587-596, ISSN 0191-8869, https://doi.org/10.1016/S0191-8869(03)00118-1.

[2] Stanford MS, Mathias CW, Dougherty DM, Lake SL, Anderson NE, Patton JH. Fifty years of the Barratt Impulsiveness Scale: An update and review. Personality and Individual Differences. 2009; 47(5):385–395.

[3] Zuckerman M. Behavioral expressions and biosocial bases of sensation seeking. New York: Cambridge University Press; 1994.

[4] Fehrman, E. et al. (2017). The Five Factor Model of Personality and Evaluation of Drug Consumption Risk. 10.1007/978-3-319-55723-6_18.

[5] Linting M, van der Kooij A. Nonlinear Principal Components Analysis with CATPCA: A tutorial. Journal of Personality Assessment. 2012; 94(1):12–25.

[6] Snoek, J., H. Larochelle, R. P. Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*. https://arxiv.org/abs/1206.2944, 2012.