Using decision trees and random forests to predict cannabis consumption

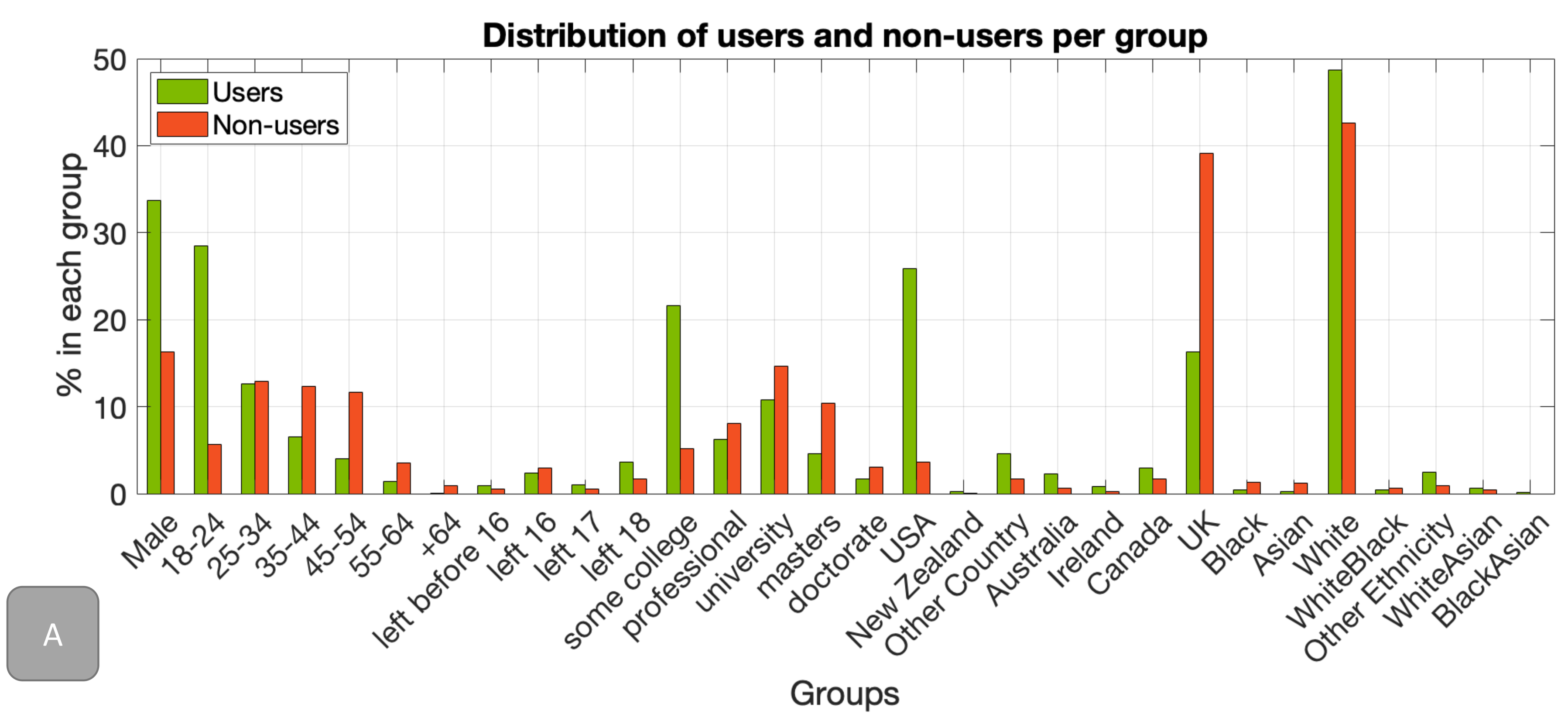
Jorge Rodríguez Peña – INM431 Machine Learning – MSc Data Science – City University of London

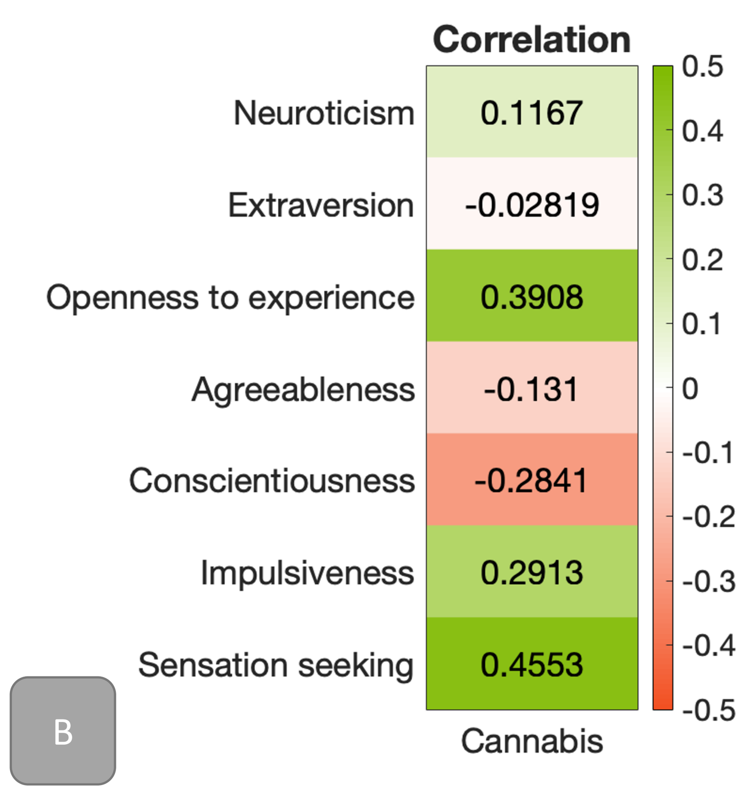
Description and motivation

- The original dataset was used and created by Fehrman E., et al. (2017) [1]. The purpose of their study was to predict drug consumption from individual and personality traits.
- It uses the revised **NEO Five Factor Inventory** [2], the reviewed **Baratt Impulsiveness Scale** [3] and the **Sensation Seeking scale** [4].
- The aim of this project is to **predict** possible **cannabis users** on a yearly basis and the **influence personality traits** might have on it.
- The results will be **compared** with those obtained by Fehrman E., et al. (2017) [1].

Analysis of the dataset

- Dataset containing **1885 samples** with 12 predictors:
  - **5 categorical predictors**: Age (binned), Gender, Country and Ethnicity.
  - **7 numerical predictors** (personality tests): Neuroticism, Extraversion, Openness to experience, Agreeableness, Conscientiousness, Impulsiveness and Sensation seeking.
- **1 binary dependent variable** for yearly basis cannabis user.
- Original dataset was cleaned to make our study simpler.
- **Data imbalance**: Country: ~60% UK ~30% USA and Ethnicity: ~95% White (see figure A).
- **Correlation** with cannabis consumption **was higher** for personality traits: Openness to Experience, Conscientiousness, Impulsiveness and Sensation Seeking (see figure B).





Decision trees vs random forest

- **DECISION TREES:**
  - **Split** the data by making a decision.
  - Use **information gain** to make the decision.
  - Generate **rules** to predict new samples.
- **RANDOM FORESTS:**
  - Use **decision trees** as basis.
  - Train each decision tree **bootstrapping** from the data (bag method) and randomly selecting predictors.
  - Use **majority vote** from each tree and average out the result.

Differences between the models:

	Decision tree	Random forest
Interpretability	✓Readable rules [6]	✗Random rules
Noise handling	✗[7]	✓Averages out the predictions
Do not overfit	✗[7]	✓Results are averaged out
Best performance	Small data [8]	Large data
Time	Shorter	Longer

Hypothesis

- The original paper [1] concluded that the **decision tree** was the **best method** for classifying cannabis users for a decade basis user definition.
- Similar behaviour expected for the yearly basis definition used in this project.
- **Random forest** should take **longer** time to train than decision trees.
- **Openness** to Experience, **Conscientiousness**, **Impulsiveness** and **Sensation Seeking** should have a **higher impact** in the model due to their **correlation** values with cannabis consumption.

Methodology

Due to the **imbalance of the data** with Country and Ethnicity predictors (see figure A) 4 models will be trained: one **decision tree** and one **random forest** per one of the following cases:

1. **All variables** included.
2. **Country** and **Ethnicity removed**.

This will allow us to compared how data imbalance affect the model.

The approach will be the same in the four models:

- **Split** the data in a training (*80%*) and test set (*20%*).
- Perform **bayesian optimization** aiming to **maximize** the **AUC** with:
  - **Parent size** and **leaf size** for *decision trees.*
  - **Leaf size**, **number of predictors** and **number of trees** for *random forest*.
  - Run 25 times and used **the median** as the optimal parameter.
- Evaluate the models using **10-fold cross validation** and **bootstrap validation** (10 bootstraps for decision trees and out of bag prediction for random forests).
- **Train** the model with the *training set* and its *optimal parameters*.
- **Predict** with the *test set* **and compare** the results.

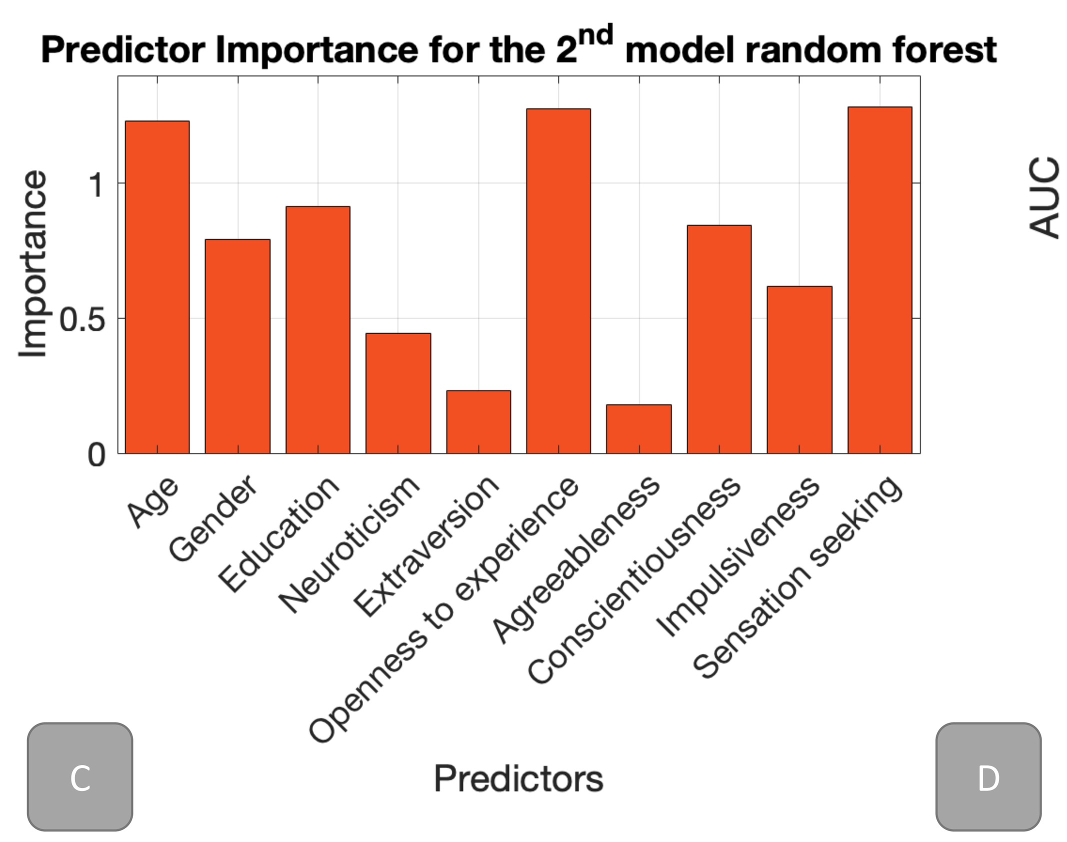
Main results

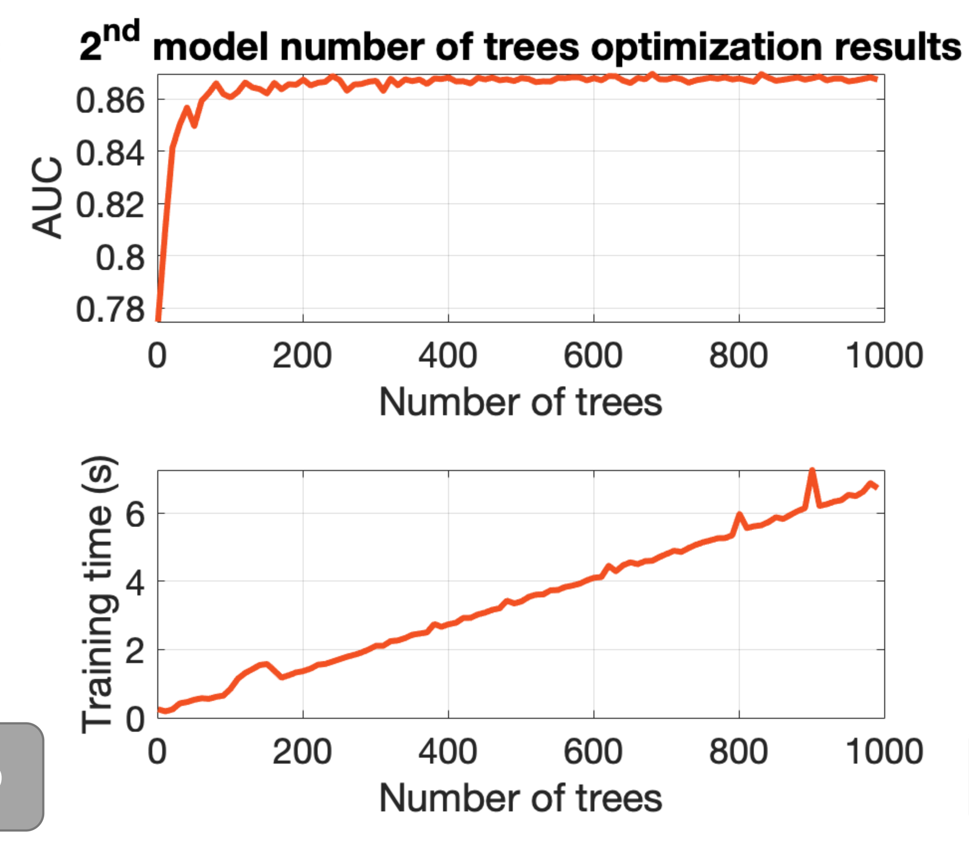
The AUC obtained for each of the models is shown in the table bellow.

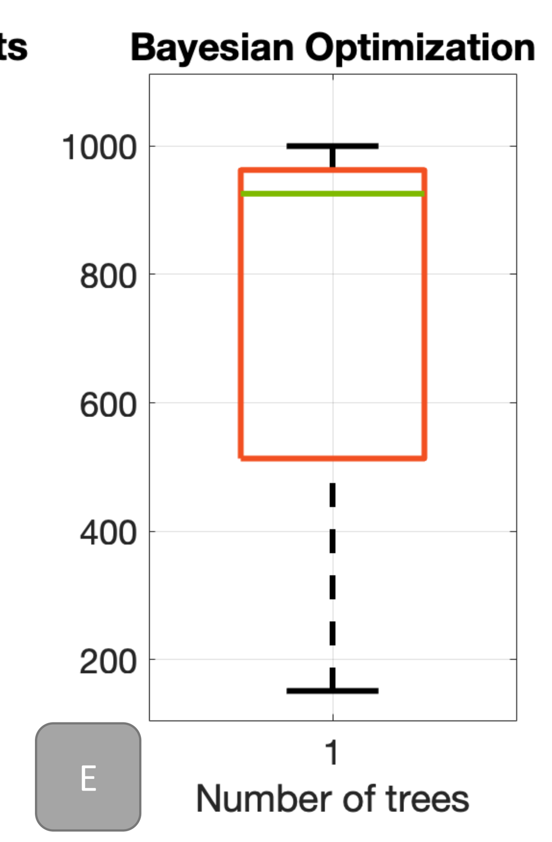
- 1st model includes all variables
- 2nd model removes Country and Ethnicity.
- Bootstrap means:
  - Bootstrap validation for decision trees
  - Out of bag validation for random forests

We see lower AUC scores for the 2nd model.

	AUC estimation	Decision tree	Random forest
1st model	CV	0.86 ± 0.03	0.89 ± 0.02
	Bootstrap	0.84 ± 0.02	0.89
	Test	0.87	0.89
2nd model	CV	0.83 ± 0.04	0.87 ± 0.03
	Bootstrap	0.80 ± 0.03	0.87
	Test	0.82	0.85

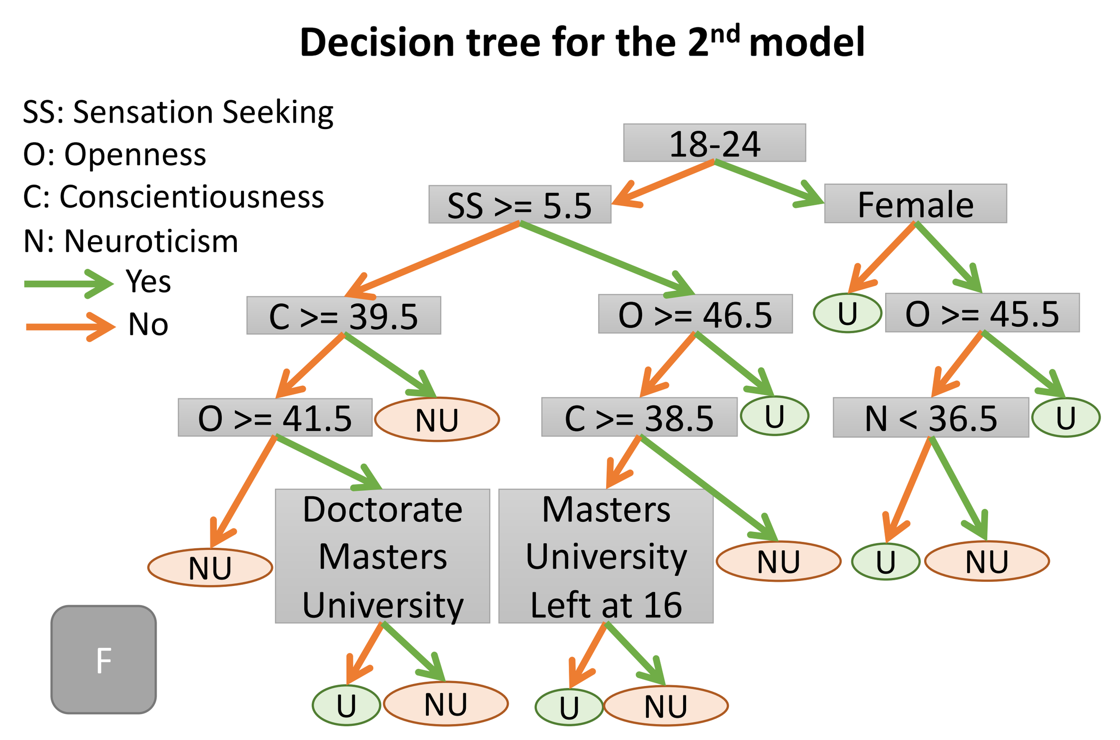






Discussion of results

- **MODEL PERFORMANCE:**
  - **Random forests performed better** than decision trees in all of the models.
  - The AUC obtained for decision trees is sometimes inside the error range, suggesting the **decision trees may outperform random forests** in some cases as shown by Fehrman E., et al. (2017) [1].
- **VALIDATION RESULTS:**
  - **Out of bag validation** is **consistent** with cross validation in random forests.
  - **Bootstrap validation** tends to **underestimate** the AUC value. Bootstrap validation has been studied as low variance but higher bias validation method [8].
- **TRAINING TIME:**
  - Random forests take around **4 seconds** to train (see figure D), decision trees around **0.03 seconds**.
- **DIFFERENCES BETWEEN MODELS:**
  - Models including **all variables** performed *slightly better* than those **excluding Country and Ethnicity**.
  - **Country** had the highest predictor importance in the 1st model. This suggests a **very biased dataset** and not representative of a real-world situation.
  - The **2nd model proves** to be very accurate in predicting and **generalising** results with AUC values over 0.8.
  - **Openness to Experience**, **Conscientiousness** and **Sensation Seeking** were the **most relevant** personality traits in our 2nd model (see figure C).
- **HYPERPARAMETER TUNING:**
  - **Decision trees** tended to have **bigger leaf sizes** (14, 32) as compared to random forests (3, 3).
  - **Random forests** performed better with a **very low** number of **sampled predictors** (2, 1).
  - **Number of trees** presented very **high variance** in *Bayesian optimization* (see figure E)
  - Evaluation of this variable against time and against AUC proved that *incrementing* the **number of trees did not improve** by much *the model* from around 600 trees, but it *added* **extra training** time (see figure D). Models were trained with **600 trees**.
  - The best decision tree for the 2nd model had **8 pruning levels** (see figure F) and 11 for the 1st model.



Conclusions and future work

- **Decision trees** are **less computationally expensive** than random forests and, given the size of this dataset (1885 samples), **perform almost as good**.
- For quickier results, decision trees *are a good option* in this case.
- Imbalanced data meant higher values of the AUC, suggesting that the **dataset is not representative of** the real world.
  - Data collection used snowball method [1], *other approaches* **could improve** the accuracy with reality.
- Most **relevant personality traits** to determine consumption: **Sensation Seeking, Openes to experience** and **Conscientiousness**. Similar to Fehrman, E. et al. (2017) [1].

1. 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.
2. 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.
3. 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.
4. Zuckerman M. Behavioral expressions and biosocial bases of sensation seeking. New York: Cambridge University Press; 1994.
5. T.M. Mitchell, Machine Learning. McGraw-Hill, 1997.
6. Yang, P., Hwa Yang, Y., Zhou, B., Zomaya, Y., et al.: "A review of ensemble methods in bioinformatics". Current Bioinformatics 5(4), 296–308 (2010).
7. Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.
8. Kohavi, Ron. (2001). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 14.