Drug Consumption Classification

Group: Group_A1_8
Bruno Fernandes (up202108871)
Catarina Canelas (up202103628)
Vasco Oliveira (up202108881)

Work to be performed

In this project, our objective is to develop and apply machine learning models capable of **predicting an individual's cannabis use consumption** status based on a range of demographic and psychological variables. The predictors include: Age, Gender, Education, Country, Ethnicity, Nscore (Neuroticism), Escore (Extraversion), Oscore (Openness), Ascore (Agreeableness), Cscore (Consciousness), Impulsive and SS (Sensation Seeking).

The aforementioned scores represent the Big Five personality traits, which are widely recognized in the field of psychology. These traits are used to describe human personality and are believed to be fundamental to individual differences.

The target variable, **cannabis use**, is categorized into seven classes, ranging from CL0 to CL6. These classes represent the frequency of drug use, with CL0 indicating 'Never Used' and CL6 signifying 'Used in the Last Day'.

By utilizing these features, we aim to develop predictive models that not only enhance our understanding of the relationship between personality traits and cannabis use but also assist in creating targeted interventions for drug abuse prevention and treatment.

Data pre-processing

In our project, we undertook several essential steps to preprocess the dataset for **predicting cannabis use**, so that the data would be clean and suitable for building effective machine learning models.

After loading the dataset from a CSV file, we ensured that there were no **missing values**, as **missing data** can lead to **inaccuracies in model predictions**. Taking that into account, we checked for any missing values in each column of the dataset. The output received showed that there were no missing values, indicating that our dataset was complete and ready for further processing.

Since our dataset included a column with a fictitious drug name called Semer, which was used to detect over-claimers or outliers, we **removed these rows** from the dataset to **ensure the data's integrity**. The ID column also seemed to be unnecessary for our analysis, as it did not provide any relevant information, therefore it was also removed.

This dataset contained **normalized values** for various features. To enhance interpretability and perform our own normalization later, we replaced these normalized values with their **actual categorical values**. For example: in the gender column, the value 0.48246 was translated to ‘Female’ and -0.48246 to ‘Male’. We also replaced the categorical labels in the drug usage columns with numerical values to facilitate modeling.

Data pre-processing

Encoding categorical variables is an essential step in data preprocessing for Machine Learning, since most machine learning algorithms require numerical input data. **One-hot encoding** was the method we used to convert categorical variables into a format that can be provided to machine learning algorithms to improve predictions. In one-hot encoding, each **category** is converted into a **new binary column**.

For example, considering the categorical feature Country, each unique value of that column is now a new column, with the value 0 if the person is not from that country, or 1 if it is.

In this dataset, the categorical features that needed to be encoded were ‘Age’, ‘Gender’, ‘Education’, ‘Country’ and ‘Ethnicity’.

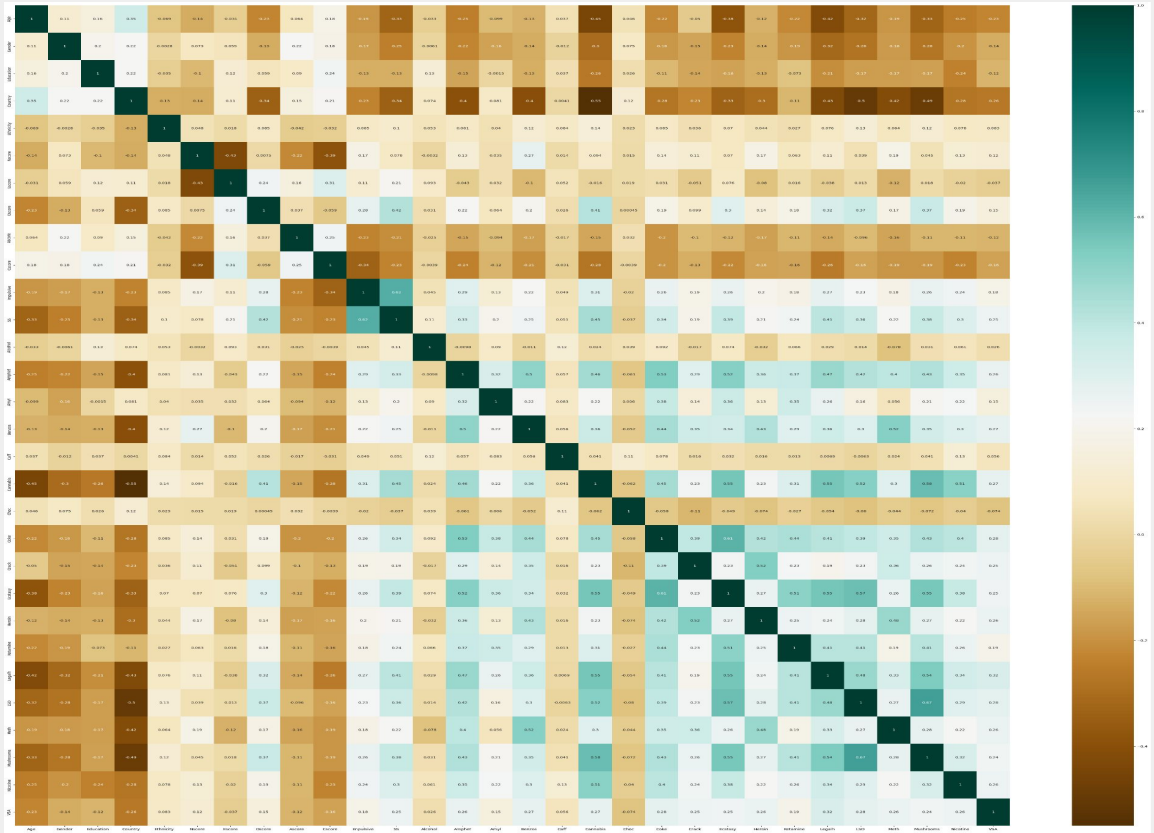
By performing one-hot encoding, we transform categorical features into a suitable numerical format.

All these preprocessing steps helped ensure that the data was **clean** and in the **right format** for building **effective machine learning models**.

Data pre-processing

To know if any columns should be removed, we drew the **correlation matrix** to see if any **correlation values** were close to 1 or -1.

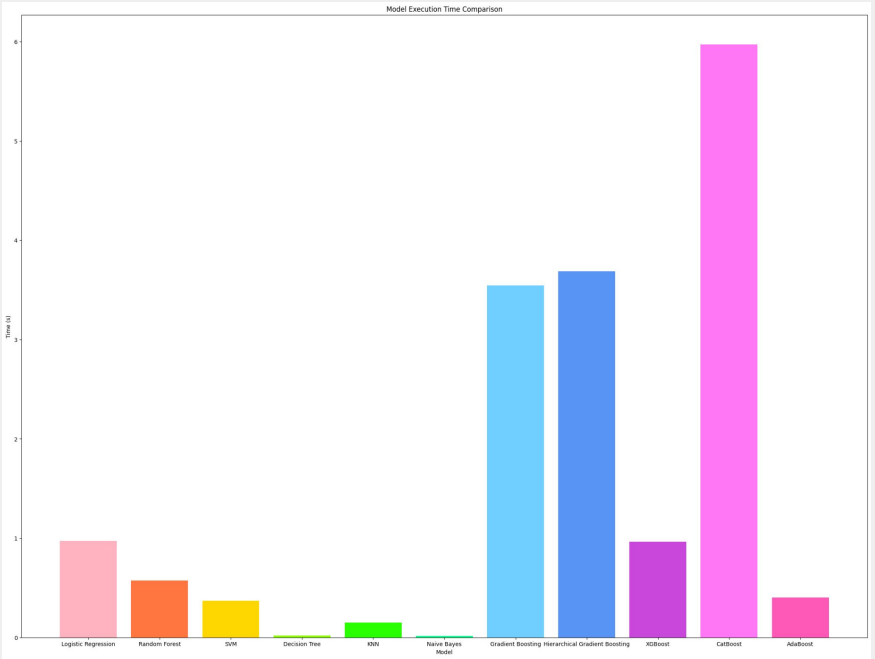
As seen in the matrix showed, no value was close enough to 1 or -1 so that it justified deleting that column, so no changes were made.



Developed Models and Evaluation and Comparison

Initially, we ran multiple algorithms to see which ones would have better results. The algorithms we used were the following:

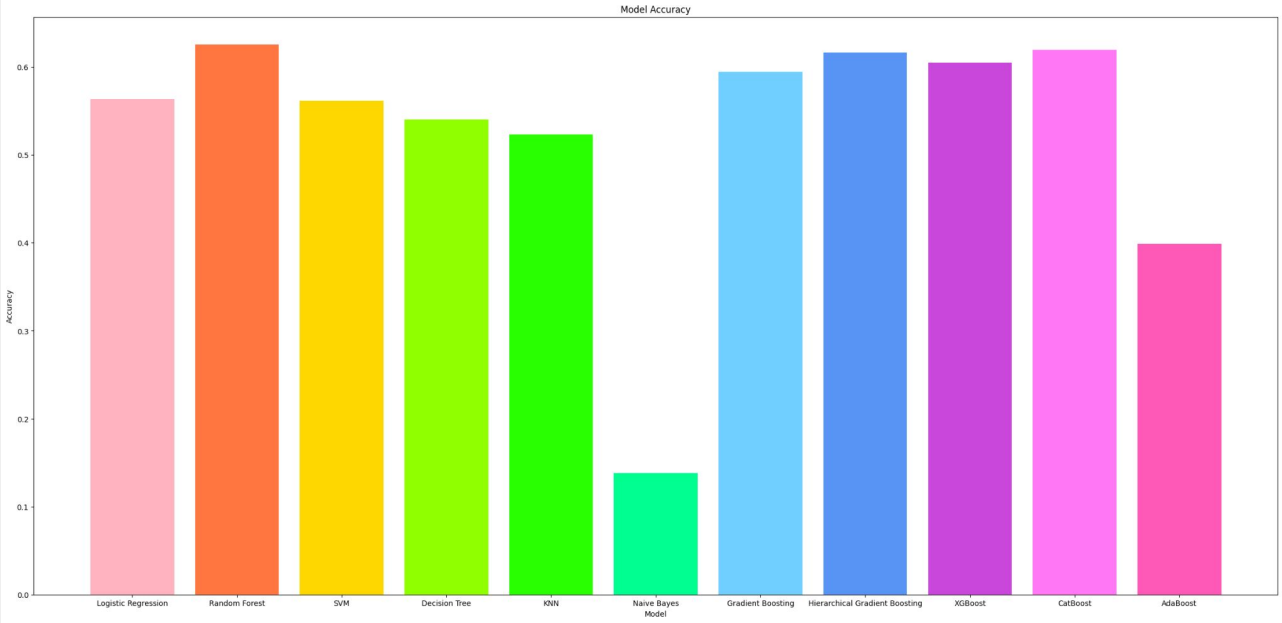
- Logistic Regression;
- Random Forest;
- SVM;
- Decision Tree;
- KNN;
- Naive Bayes;
- Gradient Boosting;
- Hierarchical Gradient Boosting;
- XGBoost;
- CatBoost;
- AdaBoost.



In the plot, we can see the execution time comparisons.

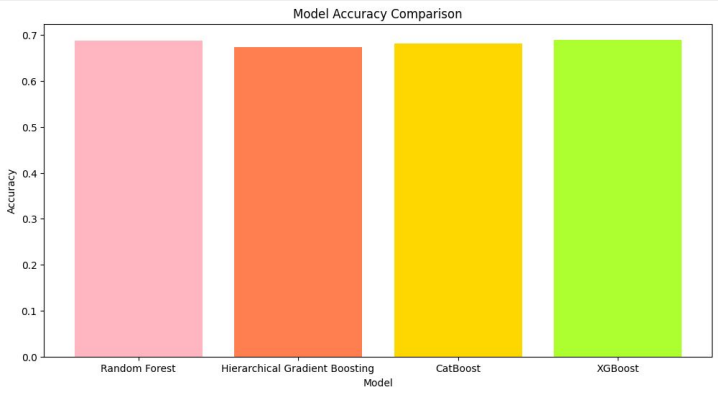
Developed Models and Evaluation and Comparison

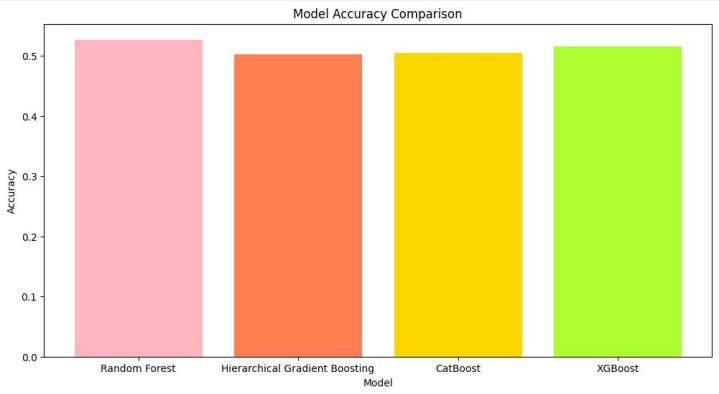
The best algorithms were **Random Forest** and the ones with **boosting**, except for AdaBoost. This is shown in the plots below, that represent the mean of each algorithms' **accuracy** for predicting **Cannabis**, **LSD** and **Mushrooms** consumption:



Developed Models and Evaluation and Comparison

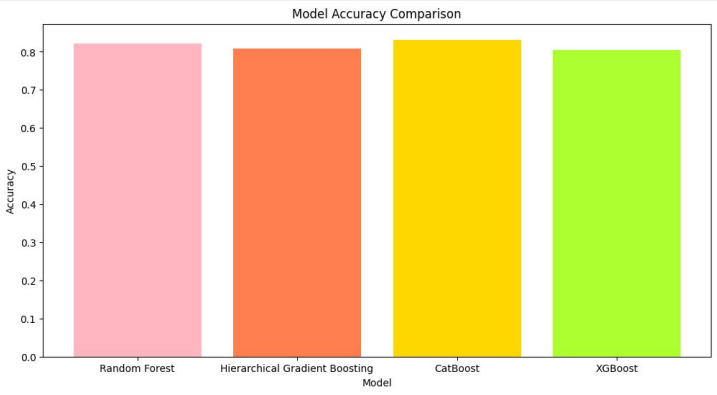
Using only non-drug columns to predict the frequency of drug consumption of an individual, we trained and tested using the best models described in the previous slide. We concluded that all models have a lower **accuracy** and a higher **mean squared error**. The best model was the **CatBoost**. The plot on the left represents the results using **all columns**, and the other using only **non-drug columns**. We concluded that the consumption of different drugs is related.

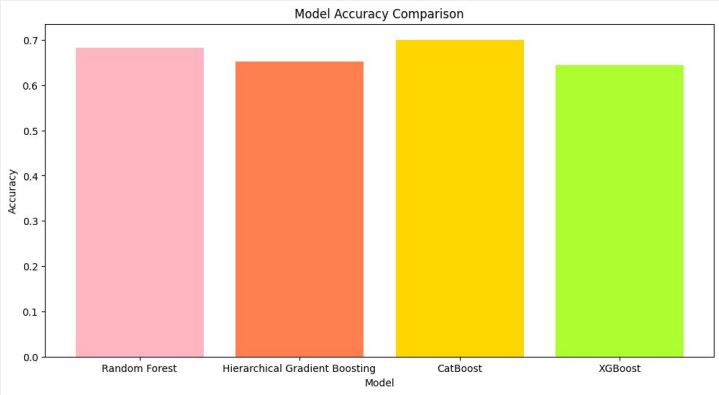




Developed Models and Evaluation and Comparison

After this, we trained and tested the best models to predict whether an individual has ever consumed a specific drug, therefore reducing the problem to a binary one. As expected, the **accuracy** scores of all the models were much **higher** and the **mean squared errors** were much **lower** than in the previous evaluations. The models' output for the **binary** problem is shown in the plots below, being the one on the left using **all columns**, and the other using **only non-drug** ones.





Related work

- Drug Consumption Classification
- Starter Notebook (converting column values)