Drug Consumption - Challenge report

Gonçalo Tavares de Bastos nº2020238997
Leonardo Cordeiro Gonçalves nº2020228071

March 2023

1 Introduction to the problem

In this challenge we solve binary classification problem for alcohol, nicotine, and cannabis using a database of 1885 candidates info based on age, gender, education and personality. By merging a subset of the classes into a new class, we determine whether a person is a consumer or not.

2 Database Description

Database contains 1885 respondents' records. Twelve characteristics for every respondent are known: NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), ImpSS (sensation seeking), educational attainment, age, gender, nationality, and ethnicity are all used to measure personality. All input attributes are originally categorical and are quantified, and thats on this data we will be working on. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select if they: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.

2.1 Content

Data base detailed content:

1. ID: is a number of records in an original database. Cannot be related to the participant. It can be used for reference only.
2. Age (Real) is the age of participant
3. Gender: Male or Female
4. Education: level of education of participant
5. Country: country of origin of the participant
6. Ethnicity: ethnicity of participant
7. Nscore (Real) is NEO-FFI-R Neuroticism
8. Escore (Real) is NEO-FFI-R Extraversion
9. Oscore (Real) is NEO-FFI-R Openness to experience.
10. Ascore (Real) is NEO-FFI-R Agreeableness.
11. Cscore (Real) is NEO-FFI-R Conscientiousness.
12. Impulsive (Real) is impulsiveness measured by BIS-11

13. SS (Real) is sensation seeing measured by ImpSS
14. Alcohol: alcohol consumption
15. Amphet: amphetamines consumption
16. Amyl: nitrite consumption
17. Benzos: benzodiazepine consumption
18. Caff: caffeine consumption
19. Cannabis: marijuana consumption
20. Choc: chocolate consumption
21. Coke: cocaine consumption
22. Crack: crack cocaine consumption
23. Ecstasy: ecstasy consumption
24. Heroin: heroin consumption
25. Ketamine: ketamine consumption
26. Legalh: legal highs consumption
27. LSD: LSD consumption
28. Meth: methadone consumption
29. Mushroom: magic mushroom consumption
30. Nicotine: nicotine consumption
31. Semer: class of fictitious drug Semeron consumption (i.e. control)
32. VSA: class of volatile substance abuse consumption

Rating's for Drug Use:

- CL0 Never Used
- CL1 Used over a Decade Ago
- CL2 Used in Last Decade
- CL3 Used in Last Year 59
- CL4 Used in Last Month
- CL5 Used in Last Week
- CL6 Used in Last Day

3 Methodology

Here we have the documentation and detailed explanation of our methodology.

3.1 Clean and prepare Dataset

First, the statement tells us that Semer is a fictional drug, so we have to remove the people who claim to have consumed this drug, as they are not providing accurate information.

Then, for this project, we will only examine only the consumption of alcohol, nicotine and cannabis. To do this, we will clean up the original dataset by removing output items that are not relevant to our study and also drop categorical inputs such as nationality and ethnicity, because we dont find it relevant for our study.

And finally, in order to prepare the dataset, we also need to encode the original quantitative data into more user-friendly data for analysis.

To better understand our results, we have the following classification:

- Age:
  - 0 = 18-24
  - 1 = 25-34
  - 2 = 35-44
  - 3 = 45-54
  - 4 = 55-64
  - 5 = 65+
- Gender:
  - 0 = M
  - 1 = F
- Education:
  - 0 = Left school before 16 years
  - 1 = Left school at 16 years
  - 2 = Left school at 17 years
  - 3 = Left school at 18 years
  - 4 = Some college or university, no degree
  - 5 = Professional Certificate/ Diploma
  - 6 = University degree
  - 7 = Masters degree
  - 8 = Doctorate degree
- Drug Use:
  - 0 = Not drug user
  - 1 = Drug user

3.2 Exploratory Data Analysis

The initial conclusion we take is that the most consumed drugs are Alcohol, Nicotine and Cannabis, then we need to take some conclusions on how the variables are related to the targets, for this we decide to make a Heat map of Variable Correlations, and some conclusions we take where:

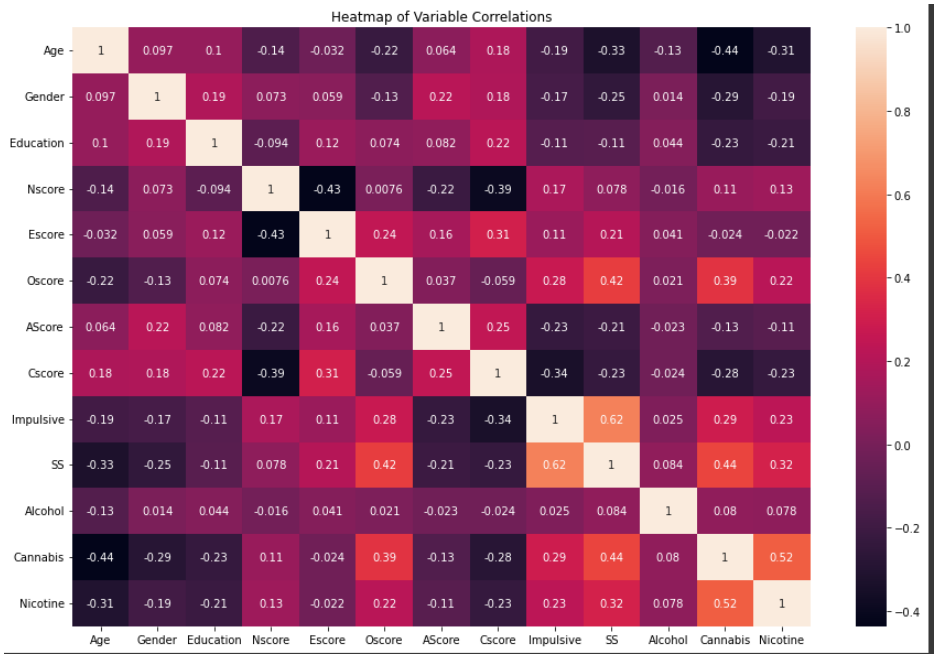

Figure 1: Correlations Heatmap

- The level of education and alcohol intake have a minor positive correlation, with more education being linked to more frequent alcohol use. Alcohol consumers show higher Impulsive Sensation Seeking (SS) scores, such that a majority of non-consumers had an negative SS score.
- Men consumed significantly more nicotine than women, showing a clear gender difference in nicotine use. Also, nicotine consumption was correlated with impulsiveness score and SS (Sensation seeking).

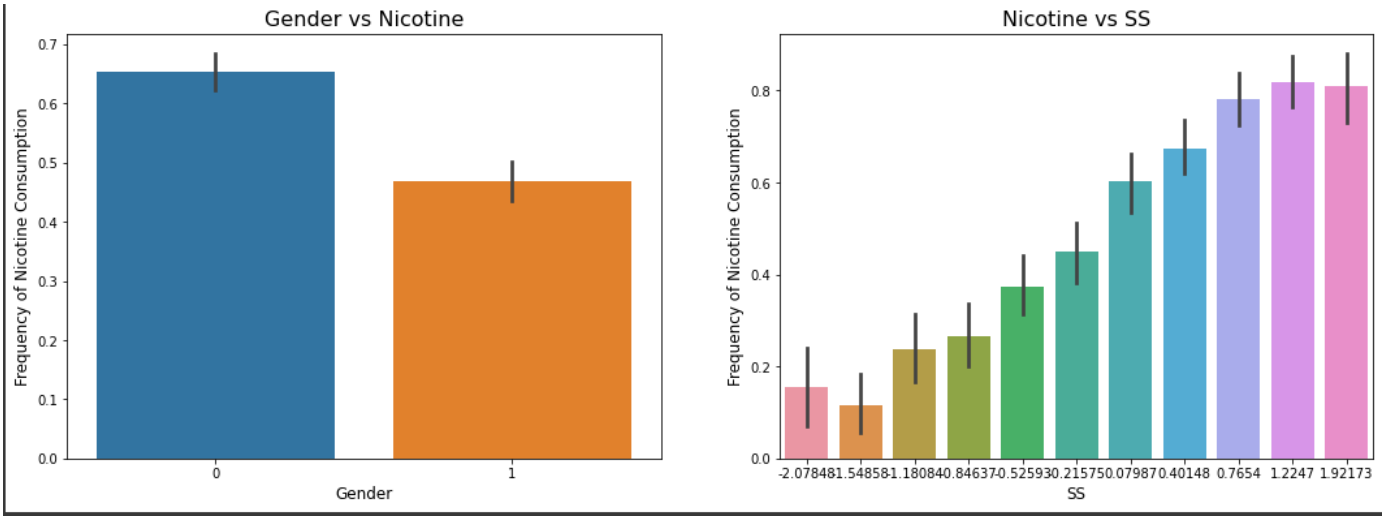

Figure 2: Gender and SS vs Nicotine

- Similar to nicotine, men used cannabis more frequently than women. Also, the frequency of cannabis use was inversely connected with age, with younger people using it more frequently than older people, and an high correlation with SS(Sensation Seeking) and OS(Openness to experience) scores.

3.3 Build and Train Models

We construct two simple machine learning models, decision trees and K-Nearest Neighbors and to evaluate the models we chose Confusion Matrix because it provides a comprehensive overview of model performance and interpretation, makes it easy to identify errors and improve model performance, useful for evaluating imbalanced datasets morelike in binary classification problems and enable the comparasion beatween the two models we used. Next we show the results we obtained only for cannabis, but we also can test for other drugs by changing the parameters on the functions:

3.3.1 Decision Tree

Decision tree classification is a type of machine learning algorithm for classification tasks. It works by creating a tree-like model of decisions and possible outcomes. The tree consists of nodes representing features or attributes of the data, and branches representing possible outcomes or decisions that can be made based on those features. The decision tree is built using a training dataset, which is used to determine the best features to partition the data at each node, in order to maximize the accuracy of the model. Once the tree is built, it can be used to classify new data samples by following the path of the tree based on the values of their attributes

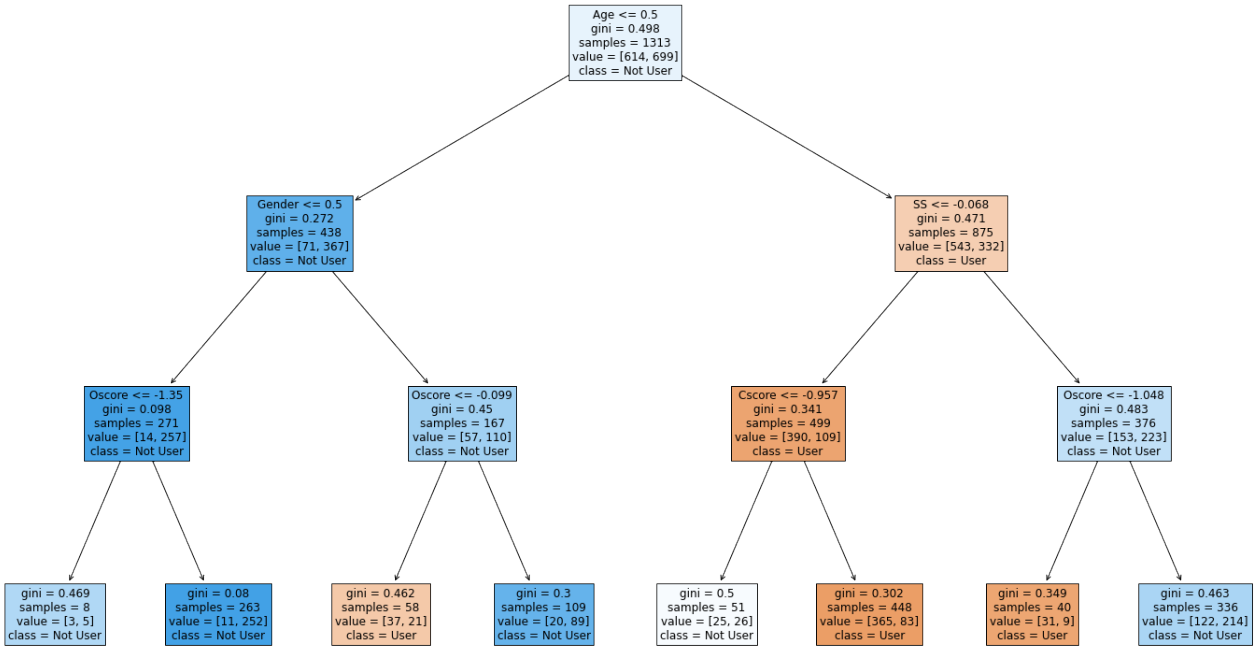

Figure 3: Decision Tree

First the results using Entropy: Accuracy: 0.746; Precision: 0.75; F1-Score: 0.75

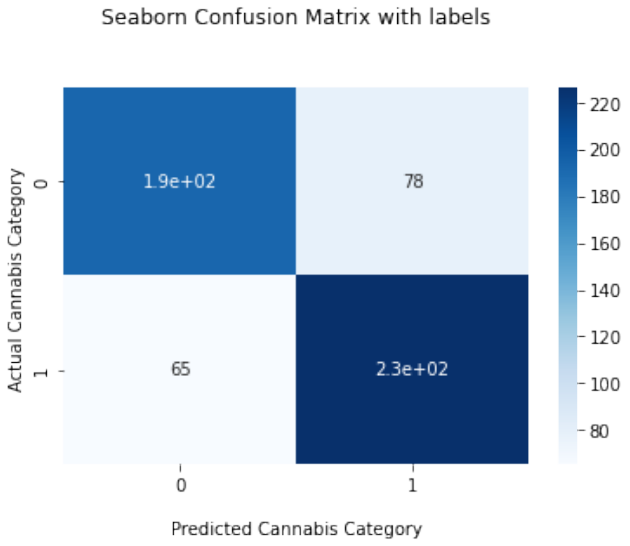

Figure 4: Confusion Matrix using Entropy

And the results using Gini: Accuracy: 0.742; Precision: 0.74; F1-Score: 0.74

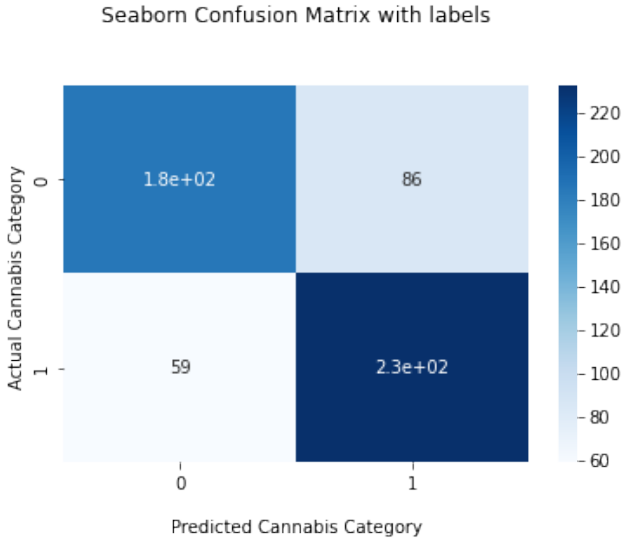

Figure 5: Confusion Matrix using Gini

3.3.2 KNN

KNN works by finding k-nearest neighbors in a new sample based on the distance metric (such as the Euclidean distance) between their objects. The class of the new instance is then based on the class majority of its nearest neighbors. KNN is a simple and efficient algorithm that can be used for both binary and multi-class classification tasks. It is easy to use and requires no training or parameter tuning. However, KNN can be sensitive to the choice of distance metric and k value, and may not perform well on datasets with high or imbalanced features.

We evaluate the performance of the algorithm with different numbers of K-neighbours looping over k values and we find that for Cannabis K=22 will give us better results: Accuracy: 0.802; Precision :0.80; F1-Score: 0.80

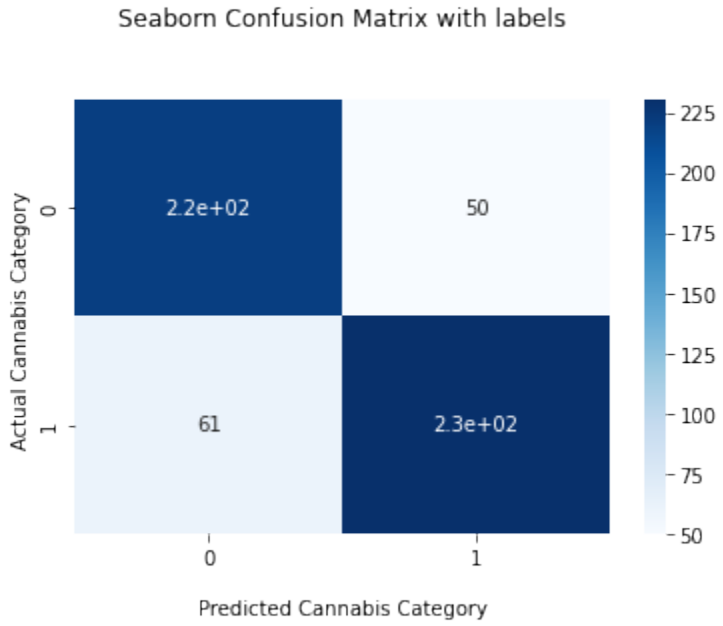

Figure 6: Confusion Matrix using KNN with n=22

3.3.3 Conclusions

In conclusion, decision tree and KNN are two simple and effective machine learning algorithms for classification tasks. Decision tree models provide a simple and interpretable description of the decision process, while KNN models are simple and easy to implement. Both algorithms have their strengths and weaknesses. On our study case KNN performs better than Decision tree with a F1-Score of 80