Who has used Heroin? An Application of Machine Learning in Drug Use Prediction

Isaac Lee | Imperial College London

Introduction

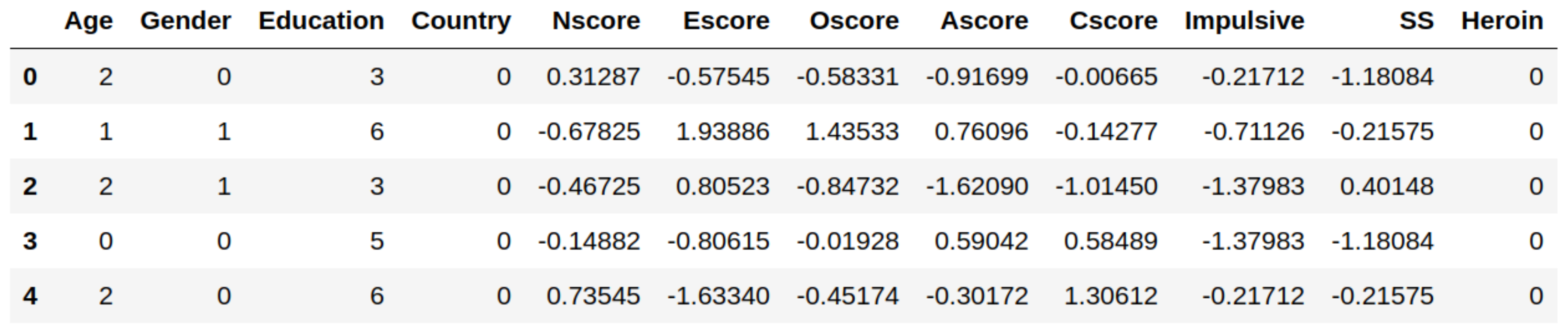
In this project we look to answer the question: is it possible to distinguish between people who have used Heroin Vs people who have never used? We also look to answer whether it is possible to build a model which predicts whether a given subject has taken Heroin at some point in their life.

The Data

This data set contains information for 1885 subjects on previous drug use for a vast array of drugs. (Note that in this project we will only be focusing on Heroin use). This data set comes from the UCI Machine Learning Repository[1] In terms of the method, "an online survery methodology was employed to collect data" [2]. For each person we have 12 features, which include a mix of categorical information and personality test score results. The Heroin column is the target vector, with binary values:

- **0** : Never used Heroin at any point in the past
- **1** : Used Heroin at least once

In terms of the balance of data, we have 1605 non-users and 280 who have used. This is not a massive number of participants so it will remain to be seen whether this affects the ceiling of our models accuracy/variation.



	Age	Gender	Education	Country	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Heroin
0	2	0	3	0	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084	0
1	1	1	6	0	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	0
2	2	1	3	0	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148	0
3	0	0	5	0	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	0
4	2	0	6	0	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	0

Figure 1:The head of the data set.

- **Age :** 0 : 18-24 | 1 : 25-34 | 2 : 35-44 | 3 : 45-54 | 4 : 55-64 | 5 : 65+
- **Gender :** 0 : Female | 1 : Male
- **Education :** 0 : Left school before 18 years | 1 : Left school at 18 years | 2 : Some college or university but no certificate or degree | 3 : Professional certificate/ diploma | 4 : University degree | 5 : Masters degree | 6 : Doctorate degree
- **Country :** 0 : UK & Ireland | 1 : New Zealand & Australia | 2 : USA & Canada | 3 : Other

For the numerical features, we have information on "the Big Five" personality traits which are tested by the NEO-FFI-R [3]
Note: SS stands for "Sensation Seeking".

Feature Engineering/Encoding

- **Encoding :** The original data set did not use integer values for the categorical features, but in order to use OneHot Encoding integers are required. Then after conversion to integer values we use the pandas get dummies function to encode the Country and Gender features, since they are nominal and so therefore integer encoding will add unwanted ordinal information.
- **Binarization :** The original data set contains four classes, such as: "used in the past day", "used in the past month" e.t.c. In order to simplify the problem we convert to: "never used" and "used at some point in the past".
- **Country Combinations :** Several countries have very few participants and so we combine geographically/culturally similar countries as seen above.
- **Education Combinations :** Originally there existed several classes of people who left school before 18. To simplify things we combine into the 0 : Left school before 18 years education class.

Feature Exploration

We now explore the distributions/correlations for each feature with count, KDE and correlation plots:

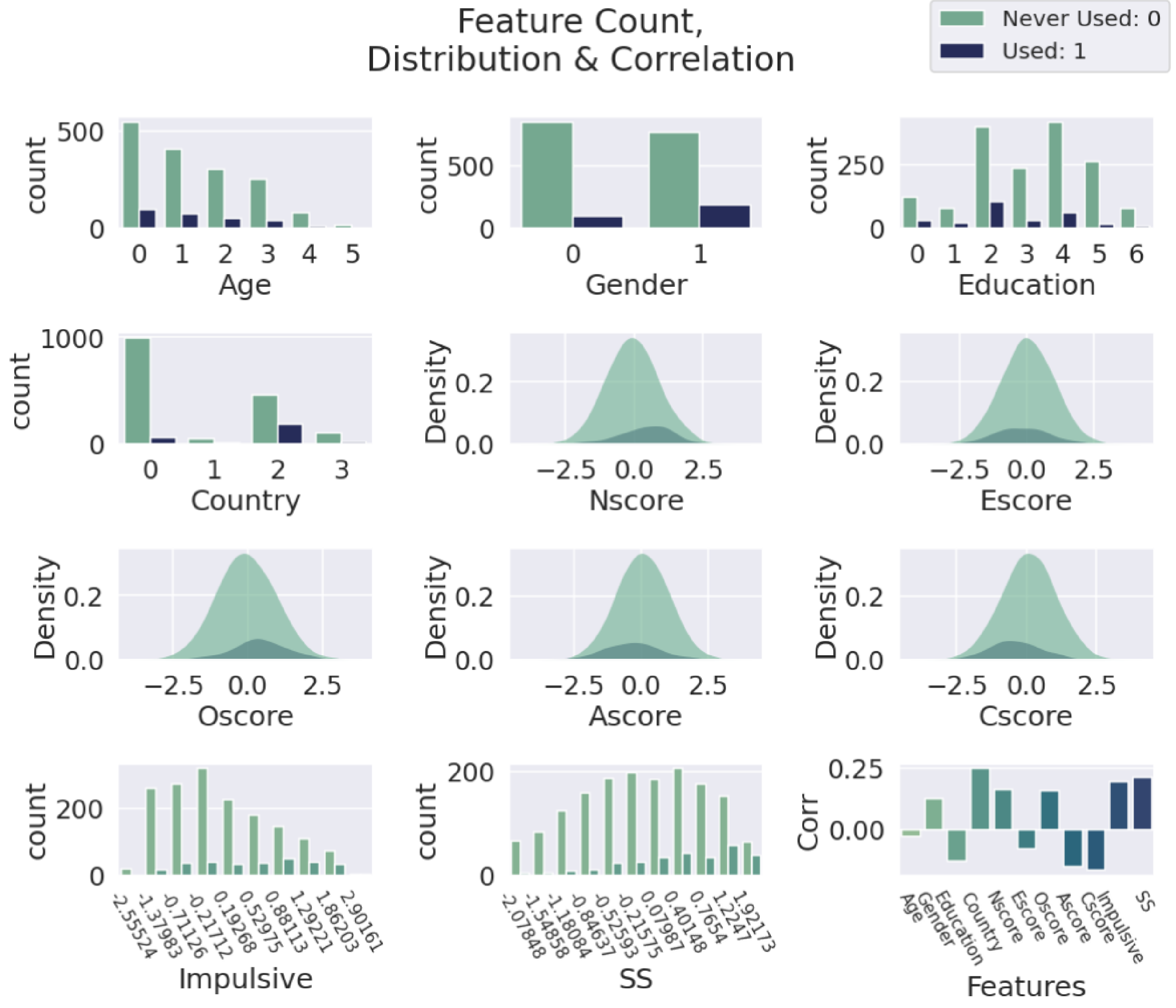


Figure 2:Exploring the distribution and correlation of each feature.

Interpretations

Clearly from these plots it appears that Nscore, Escore, Oscore, Ascore and Cscore are normally distributed, however the other features are not, with non-central means. Also when considering the correlation of features with Heroin use, it appears that Country, SS and Impulsive are most correlated, suggesting that they will have greater importance in our models.

Model Creation

Train Test Split

We use Sklearn to split our data into stratified 80% training and 20% testing data.

Naive Bayes

Naive Bayes will provide a good baseline model for comparison due to it's simplicity. We use the Sklearn library to fit the Naive Bayes model to our training data. Naive Bayes uses Bayesian Inference to get information on the posterior distribution. So for every vector of features we can calculate the probability of obtaining either a 1 or a 0 given the data we have observed, and predict the most likely. Note that we assume independence of features, when clearly this is not the case [4].

XGBoost

For a more complex model we turn to XGBoost. This algorithm uses an ensemble approach with many "weak learners" which are weighted according to their error [4]. We use the xgboost python library to fit the model to our training data, and then we use Sklearn GridSearchCV to find the best possible parameters using cross validation.

Random Forest

Finally for our third model, we try Random Forest, which is another decision tree algorithm. The basic idea of Random Forest is to create a "Forest" of smaller decision trees made from random subsets of features, which separate samples based on their "Gini Impurity" and then using ensemble techniques the results are combined [4]. As before we tune our parameters using GridSearchCV in several iterations.

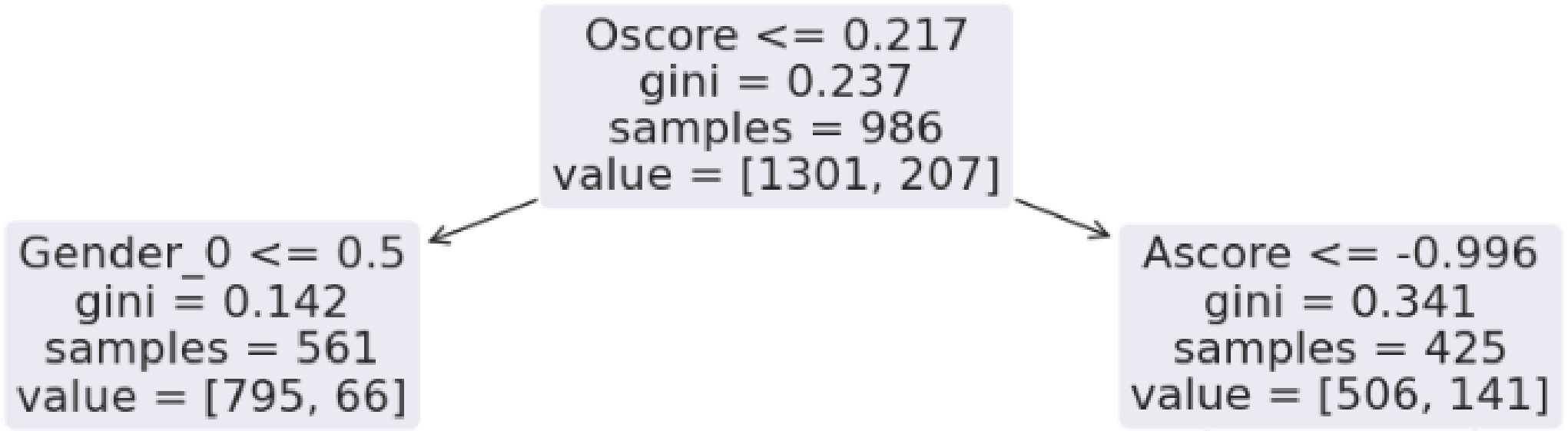
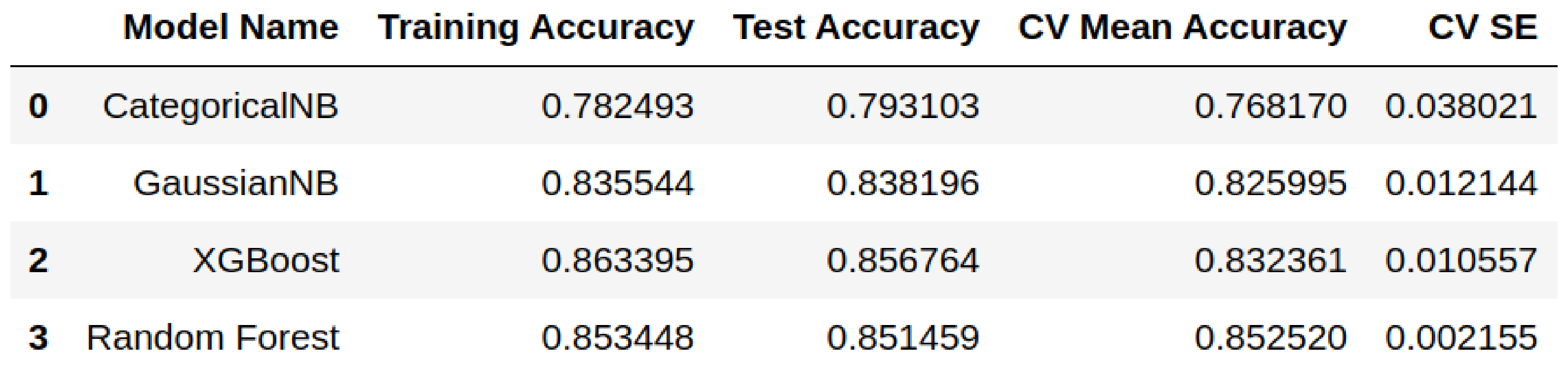


Figure 3:An example of some nodes from a decision tree.

Model Results

In the table below we provide the results on both the training and testing data for each algorithm. Also we use 5-fold cross validation on the **whole** data set for another measure of accuracy. Finally we have the standard error for the 5-fold cross validation results for each model.



	Model Name	Training Accuracy	Test Accuracy	CV Mean Accuracy	CV SE
0	CategoricalNB	0.782493	0.793103	0.768170	0.038021
1	GaussianNB	0.835544	0.838196	0.825995	0.012144
2	XGBoost	0.863395	0.856764	0.832361	0.010557
3	Random Forest	0.853448	0.851459	0.852520	0.002155

Figure 4:Model Results

Interpretation of Results

Firstly we note that the accuracy for XGBoost and Random Forest is decently higher than Naive Bayes, which is to be expected. Also note that we are getting some overfitting on our XGBoost model, since the training data accuracy is higher than the test data accuracy. Also whilst the accuracy for XGBoost on the test data is higher than Random Forest, when we look at the 5-fold cross validation scores Random Forest is the clear winner, with it also having a much lower standard error than any of the other models. It is also interesting to consider the confusion matrix heatmaps for each model, because they show that the decision tree based algorithms actually performed worse when identifying users, tending to largely favour non-users.

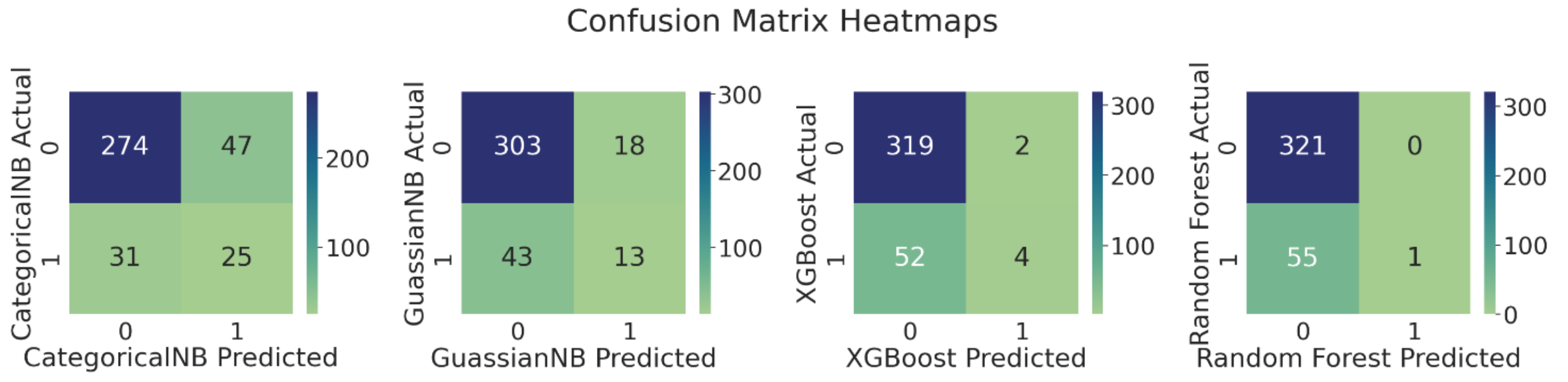


Figure 5:Comparing predicted values with actual for each model.

Conclusions

We can now answer the questions we first set out: Yes! (Accepting an error of approx 15%), it is possible to predict whether or not someone has used Heroin at some point in their life, suggesting that (most) Heroin users have distinguishable characteristics.

GitHub & Oral Presentation

- GitHub Repository for source Jupyter Notebook code:
`https://github.com/isaacjeffersonlee/Project_6_Y1_ICL`
- Oral Presentation Link:
`https://www.youtube.com/watch?v=dQw4w9WgXcQ`

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[2] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk, 2017.

[3] Michael C. Ashton. Chapter 2 - Personality Traits and the Inventories that Measure Them. Academic Press, San Diego, second edition edition, 2013.

[4] Tibshirani R. Friedman J. H. Hastie, T. The elements of statistical learning: Data mining, inference, and prediction : with 200 full-color illustrations. New York : Springer, New York, second edition edition, 2001.