



**Project Tittle:**  Drug Consumption Predictive Classification Machine Learning Model

**Date of Submission:** 02/05/23

**Course Title:** Programming in Python

**Section:** [A] **Semester:** Spring 2022-2023

**Course Teacher:** DR. Akinul Islam Jony

**Submitted By:**

**NAME:** Tanjim, Samiul Arif

**ID:** 20-42694-1

**Project Overview:** Drug Consumption Prediction Classification on Cocaine, Heroin, Methamphetamines and nicotine.

Perform feature classification to create target labels for drugs of interest. We will create three separate datasets to asses predict whether an individual uses cocaine, methamphetamines, heroin or nicotine.

**Dataset Overview:** This data was taken from UCI repository.

Link:<https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29>

Dataset contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. 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 one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.

Since all of the features have been quantified into real values, please refer to the link to the original dataset to get more clarity on categorical variables.

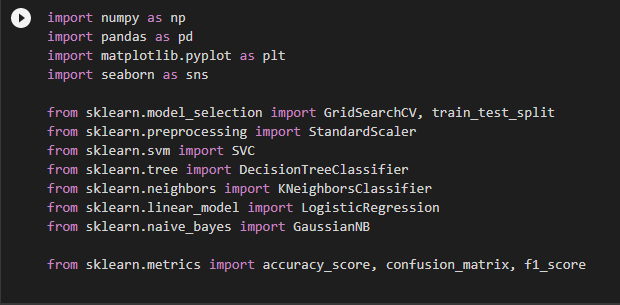
**Feature Attributes for Quantified Data:**

1. ID: is a number of records in an original dataset
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

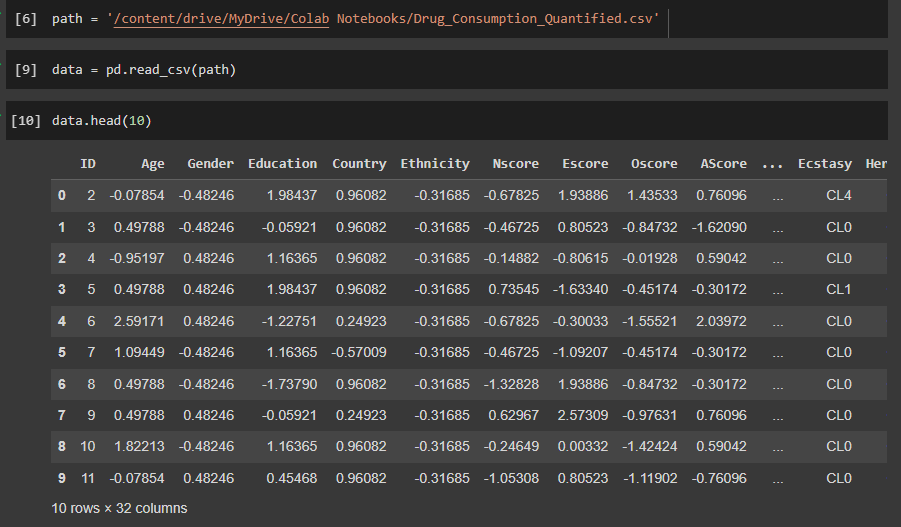
Ratings for Drug Use:

* 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

**Packages and Libraries**

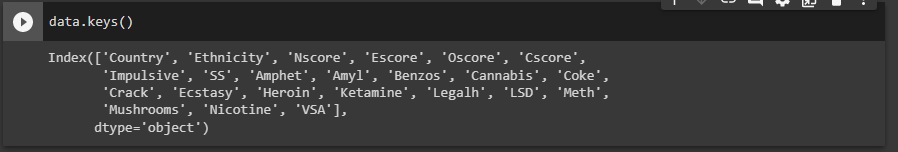


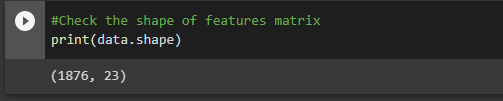
**Load Dataset**



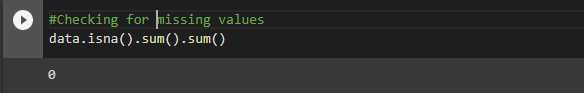
**Explore the Dataset**

Print the keys of ‘Drug Consumption’ dictionary object

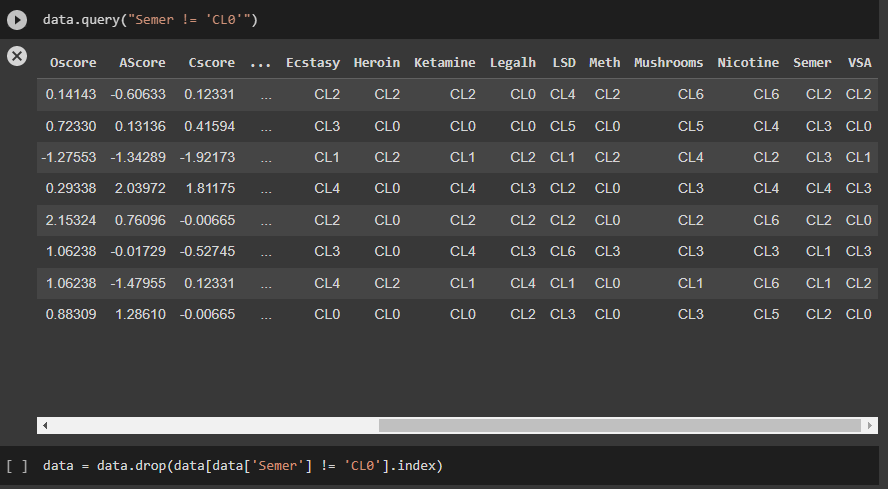




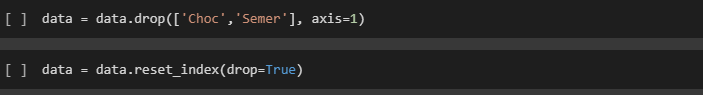
**Data Cleaning**



In the description of data, we mentioned that Semer is not a drug rather it used as a control. Therefore, we will remove these individuals from the data frame.

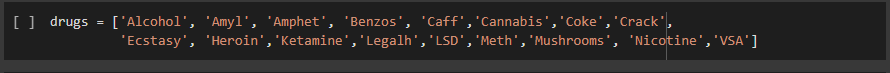


We will also drop unnecessary columns



**Feather Encoding**

As the drugs was labelled in categorical value. We have to encode them with numerical value before applying classification algorithms.

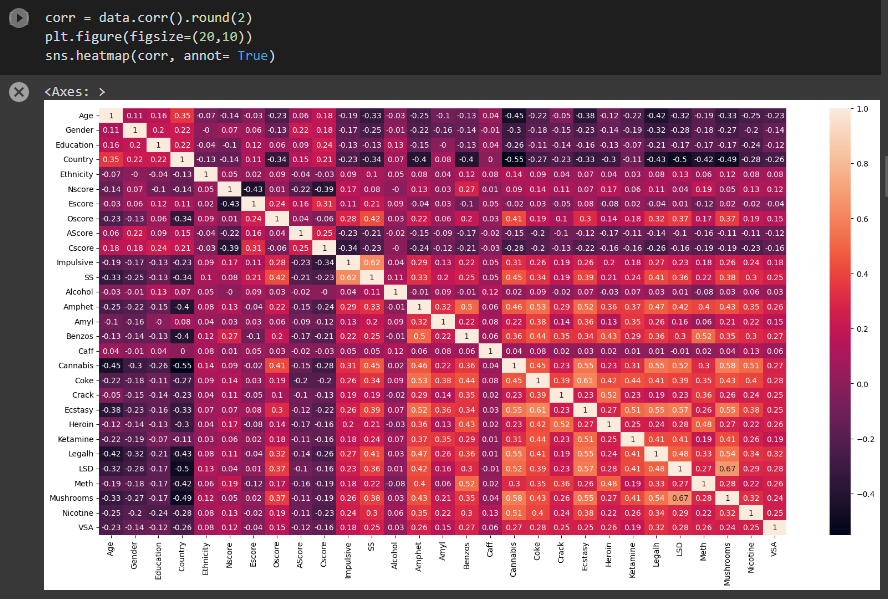


We have replaced ‘CL0 Never Used’ with ‘0’, ‘CL1 Used over a Decade ago’ with ‘1’, ‘CL2 Used in Last Decade’ with ‘2’, ‘CL3 Used in Last Year’ with ‘3’, CL4 Used in Last Month’ with ‘4’, ‘CL5 Used in Last Week’ with ‘5’ and ‘CL6 Used in Last Day’ with ‘6’.

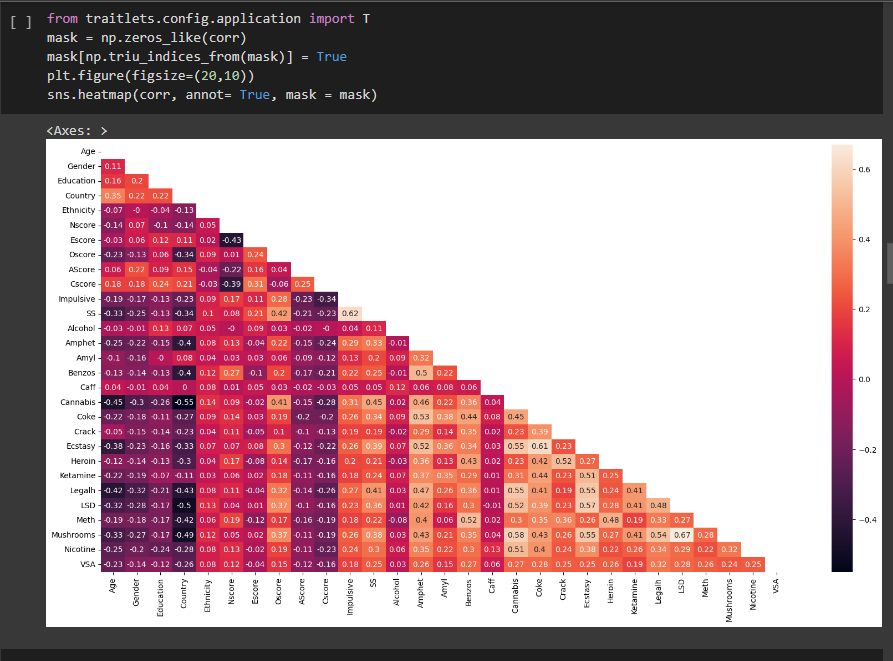


**Exploratory data analysis**

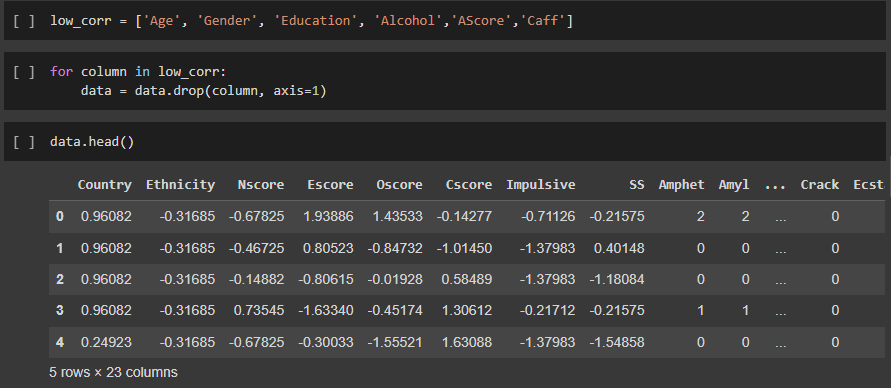
The correlation coefficient ranges from -1 to 1. If the value is close to 1, it means that there is a strong positive correlation between them the two variables.

****

**Removing redundant values**



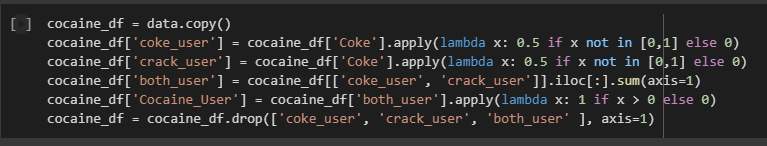
We can see that some of the attributes have low correlation. Those are Age, Gender, Education, Alcohol, AScore, Caff. We will remove these columns for better accuracy of the model.

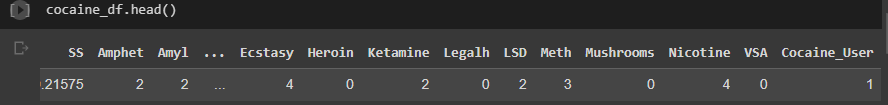


**Feature Selection**

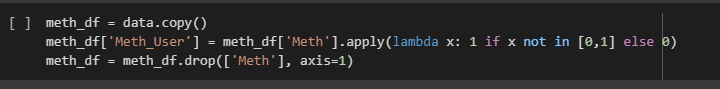
We will create four datasets to asses predict whether an individual uses Cocaine, Methamphetamines, heroin or nicotine.

We will combine **Cocaine and Crack Cocaine** usage into one feature.



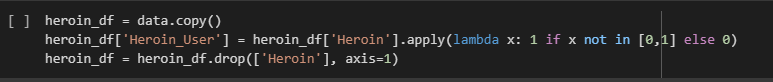


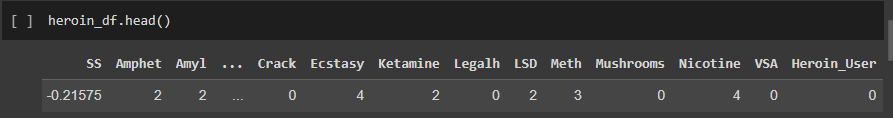
**Methamphetamine**



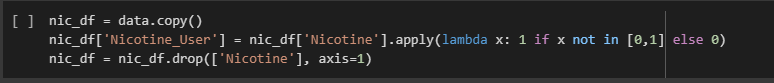


**Heroin**



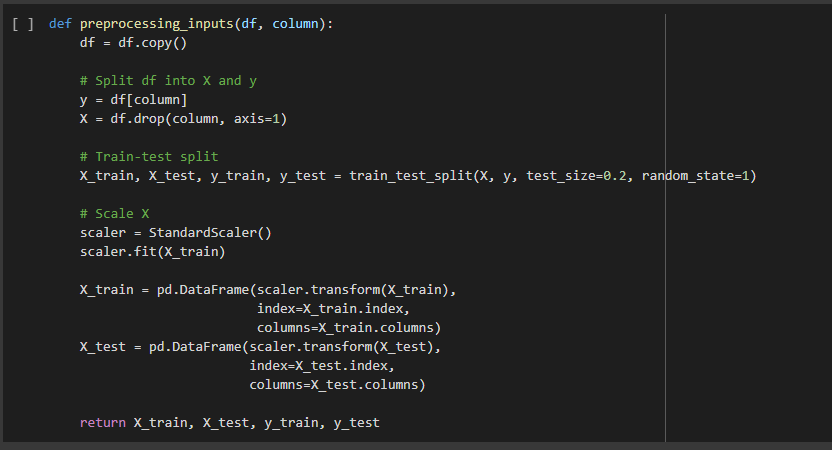


**Nicotine**

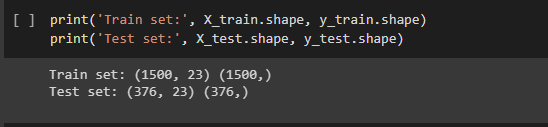




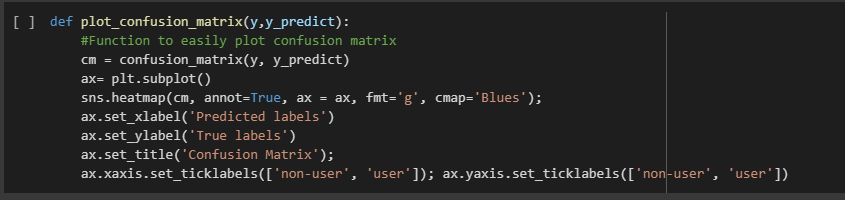
**Spit the dataset**



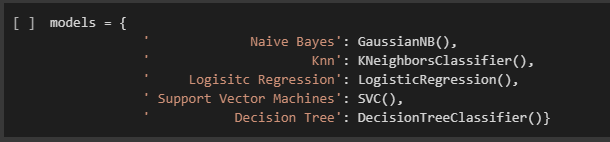
After splitting dataset into training set and test set



**Function for Confusion Matrix**

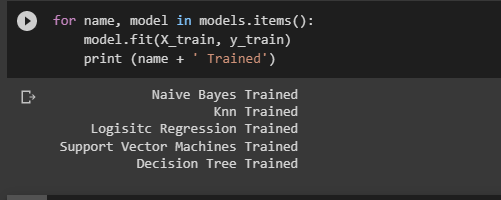


**Model**

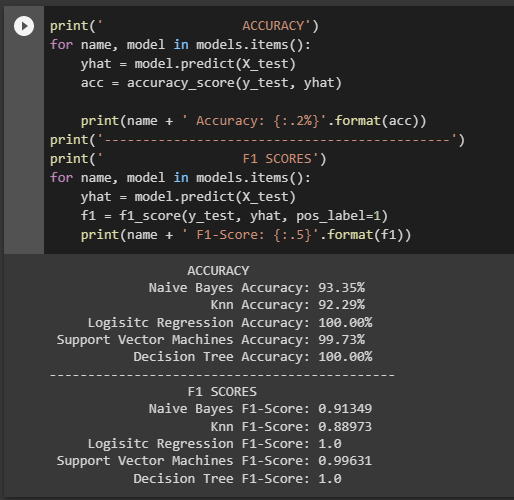


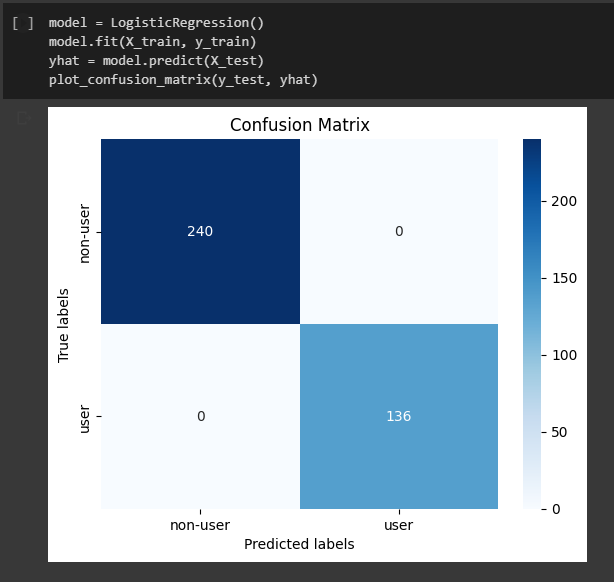
**Cocaine Model Training**



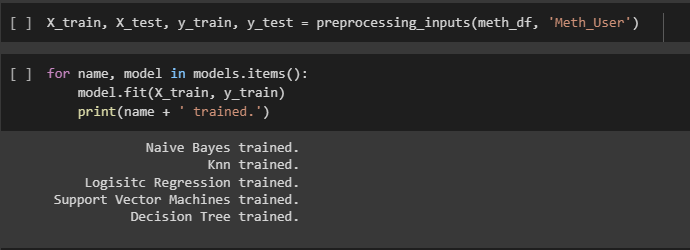


**Cocaine Model Result**

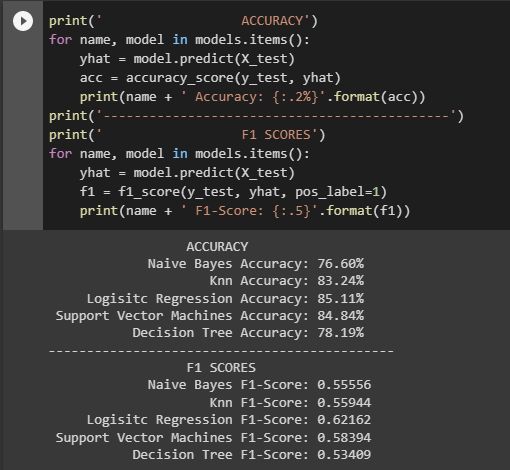




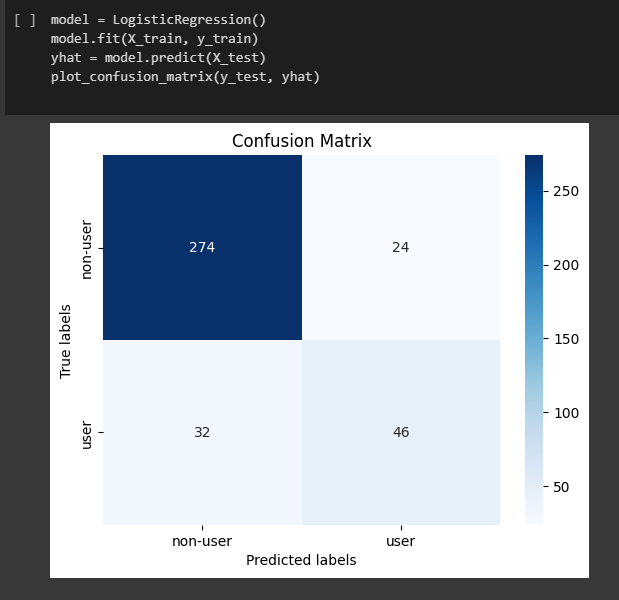
**Methamphetamine Model Training**



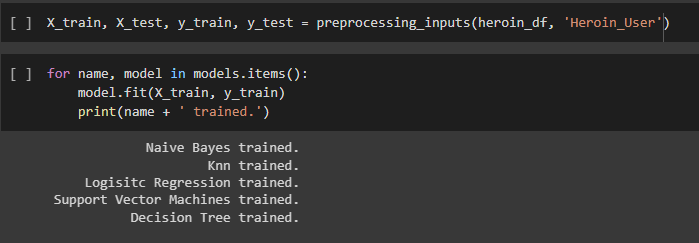
**Methamphetamine Model Result**



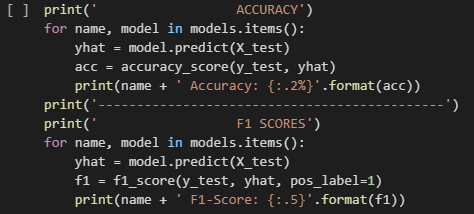
**Methamphetamine Logistic Regression model confusion matrix**

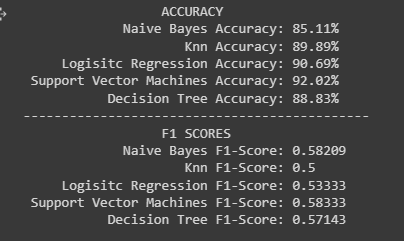


**Heroin Model Training**

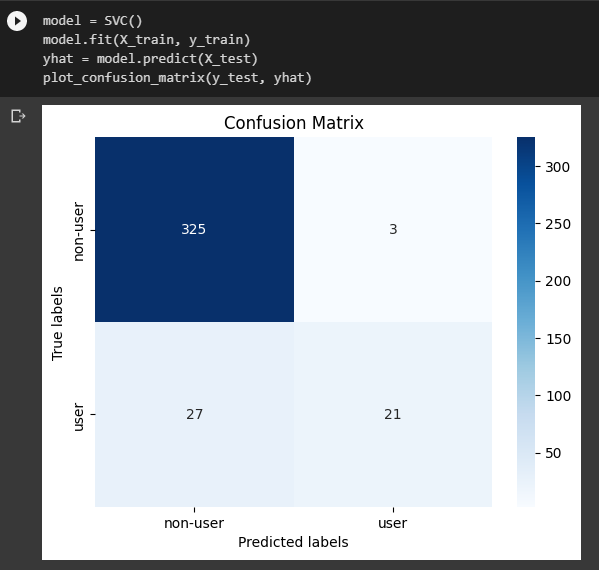


**Heroin Model result**

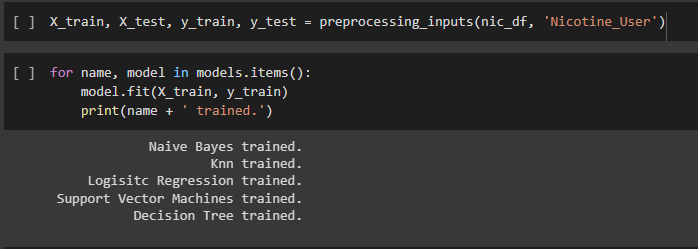




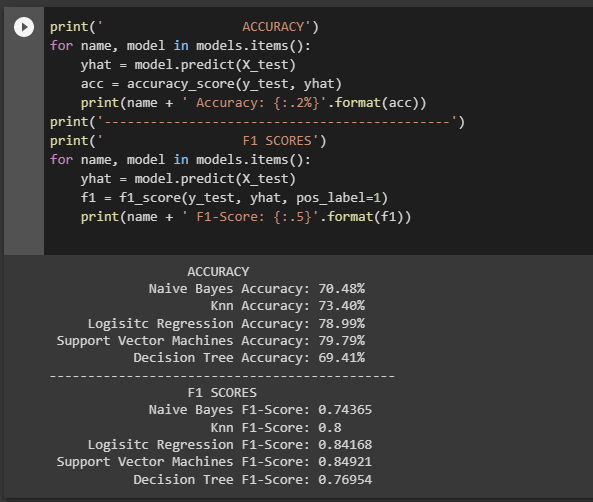
**Heroin Support Vector Machine (SVM) Model Confusion Matrix**



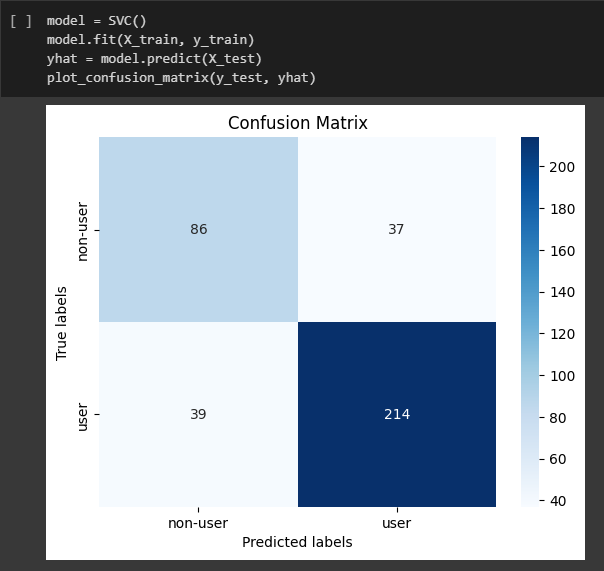
**Nicotine Model Training**



**Nicotine Model Result**



**Nicotine Support vector Machine Model Confusion Matrix**



**Discussion And Conclusion**

* Both logistic regression and decision tree exhibited 100% accuracy in classifying Cocaine users.
* For Methamphetamine user classification, logistic regression provides the maximum accuracy of 85%.
* The support vector machine model has a 92% accuracy rate for heroin user classification.
* For nicotine user classification, the support vector outperforms all other classifiers by about 80%.

<https://colab.research.google.com/drive/16KCHl2C82hP3MLZJl6E1lgdkN_wH9Nmc?usp=sharing>