SC1015 Mini Project

Members:

Luo Maoyuan (U2220249K)

Gupta Raghav (U2222889A)

Gokul Ramesh (U2222182H)

Group member 1:

LUO MAOYUAN

“Computing is my passion!”

GUPTA RAGHAV

"I feel a deep sense of satisfaction from solving complex algorithms"

Group member 3:

GOKUL

“I like doing Data Analysis in my free time!”

Table Of Contents

Project Introduction

Data Cleaning

EDA & Analysis

Machine Learning

Conclusion

01

Project Introduction

Drug Consumption Statistics

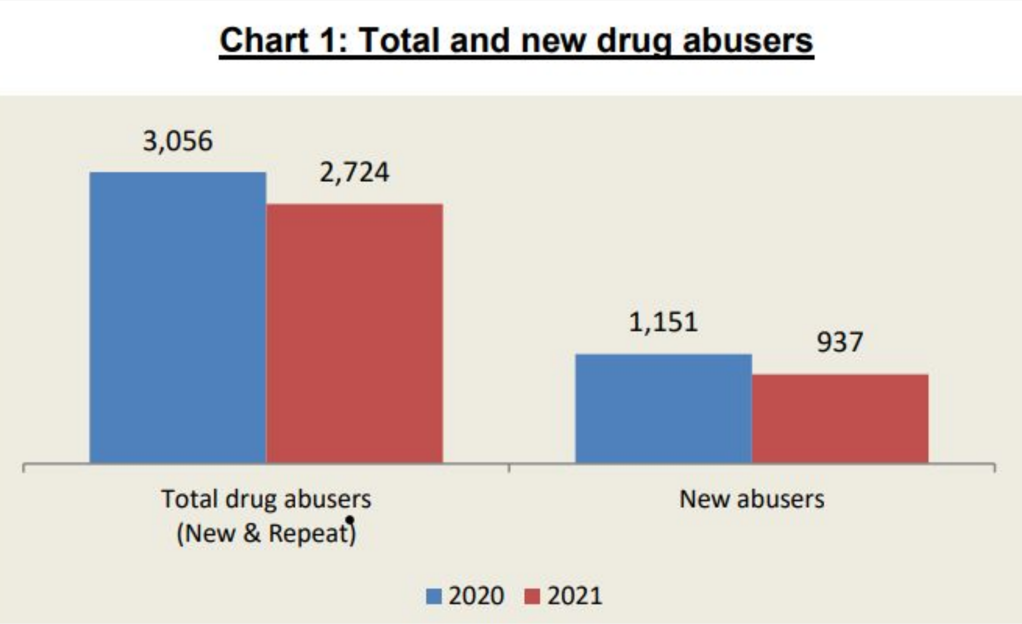


Chart 1: Total and new drug abusers

Breakdown

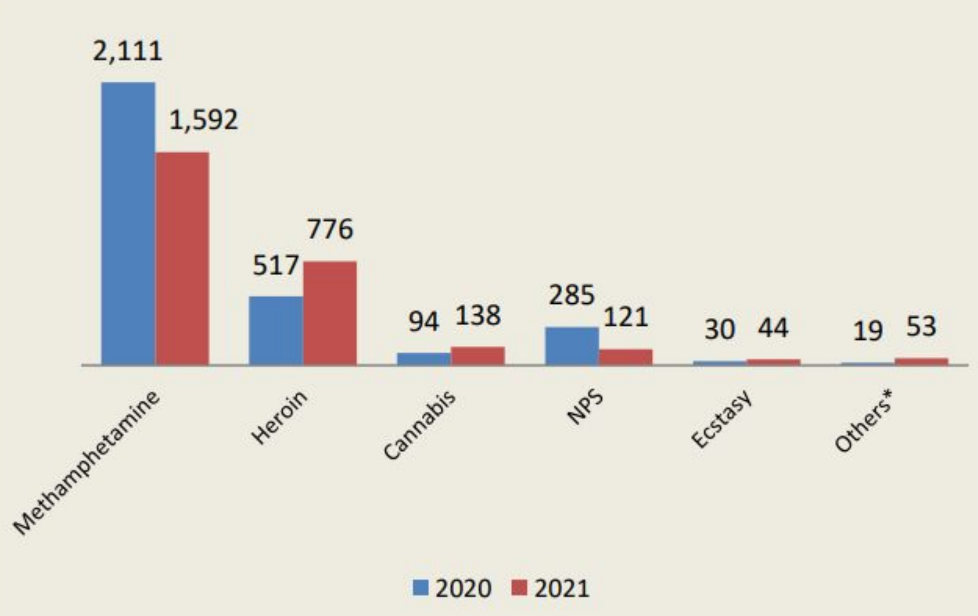
From the graph, we can see that a significant proportion of caught drug users are new offenders

66%
Repeat Users

34%
New Users

https://www.cnb.gov.sg/docs/default-source/drug-situation-report-documents/cnb-annual-statistics-2021.pdf

Drug Consumption Statistics



Breakdown

Majority of drug users used:

- Methamphetamine
- Heroin
- Cannabis

3703
Meth Users

1293
Heroin Users

242
Cannabis Users

https://www.cnb.gov.sg/docs/default-source/drug-situation-report-documents/cnb-annual-statistics-2021.pdf

IS THERE A WAY TO PREDICT DRUG USAGE?

- Identify relationship between personal factors and drug usages

- Create a model that is able to identify individuals at higher risk of consuming drugs

- An accurate model would allow individuals who are more at risk to be alert to their own susceptibility

Drug Consumption Statistics
(UCI Machine Learning Repository)

Identifying Individuals At Risk Of Drug Usage Based On Personality Type

1885
Respondents

7
Personality Factors

14
Classes Of Drugs

Categorical Drug
Consumption Values

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

7
Personality Factors

- Personality factors values are stated based on a normal distribution

Categorical Drug Consumption values

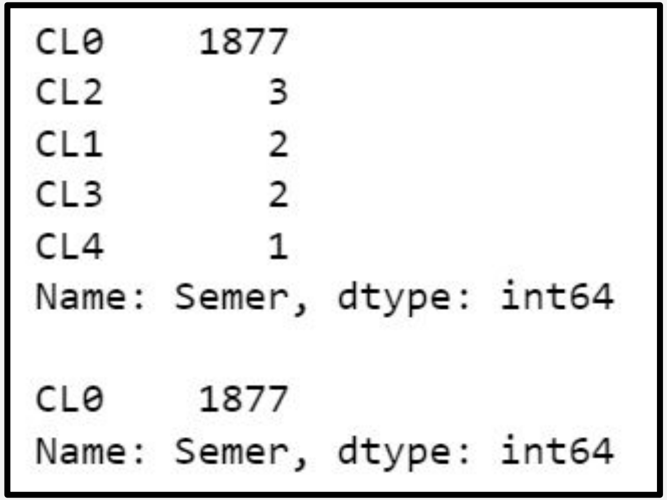
- Drug consumption categories are listed as follows:
  - CL0 Never Used
  - CL1 Used over a Decade Ago
  - CL2 Used in Last Decade
  - CL3 Used in Last Year
  - CL4 Used in Last Month
  - CL5 Used in Last Week
  - CL6 Used in Last Day

02

Data Cleaning

Removal of inaccurate data

Eg : Semer is a fake drug so responses from people that indicated they used Semer will be erroneous



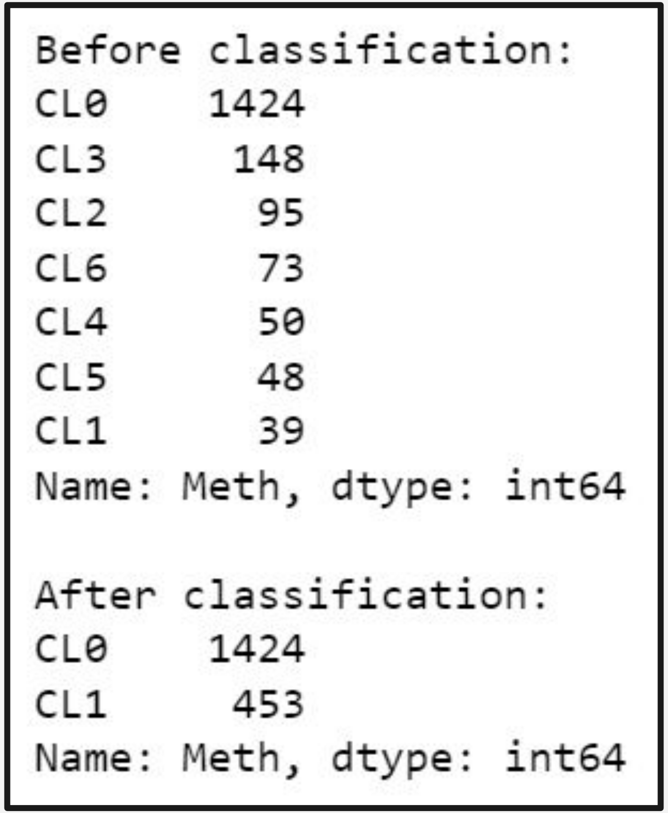
```
CL0     1877
CL2        3
CL1        2
CL3        2
CL4        1
Name: Semer, dtype: int64

CL0     1877
Name: Semer, dtype: int64
```

- The dataset rows with responses that are **NOT CL0** will be removed

Reclassification Of Drug Usage Categorical Values

Values will be reclassified to :

CL0 : Never used before
CL1 : Used before

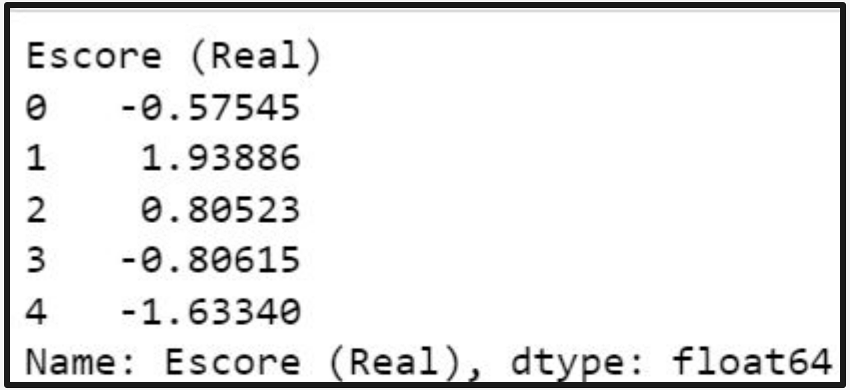


```
Before classification:
CL0    1424
CL3     148
CL2      95
CL6      73
CL4      50
CL5      48
CL1      39
Name: Meth, dtype: int64

After classification:
CL0    1424
CL1     453
Name: Meth, dtype: int64
```

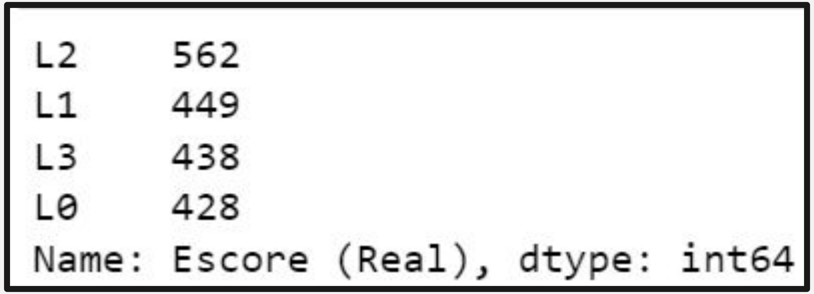
Converting Normalised Personality Factors To Categorical Values

- **L0** - Individual displays the least amount of a certain personality

- **L1** - Individual displays a very high amount of a certain personality



```
Escore (Real)
0   -0.57545
1    1.93886
2    0.80523
3   -0.80615
4   -1.63340
Name: Escore (Real), dtype: float64
```

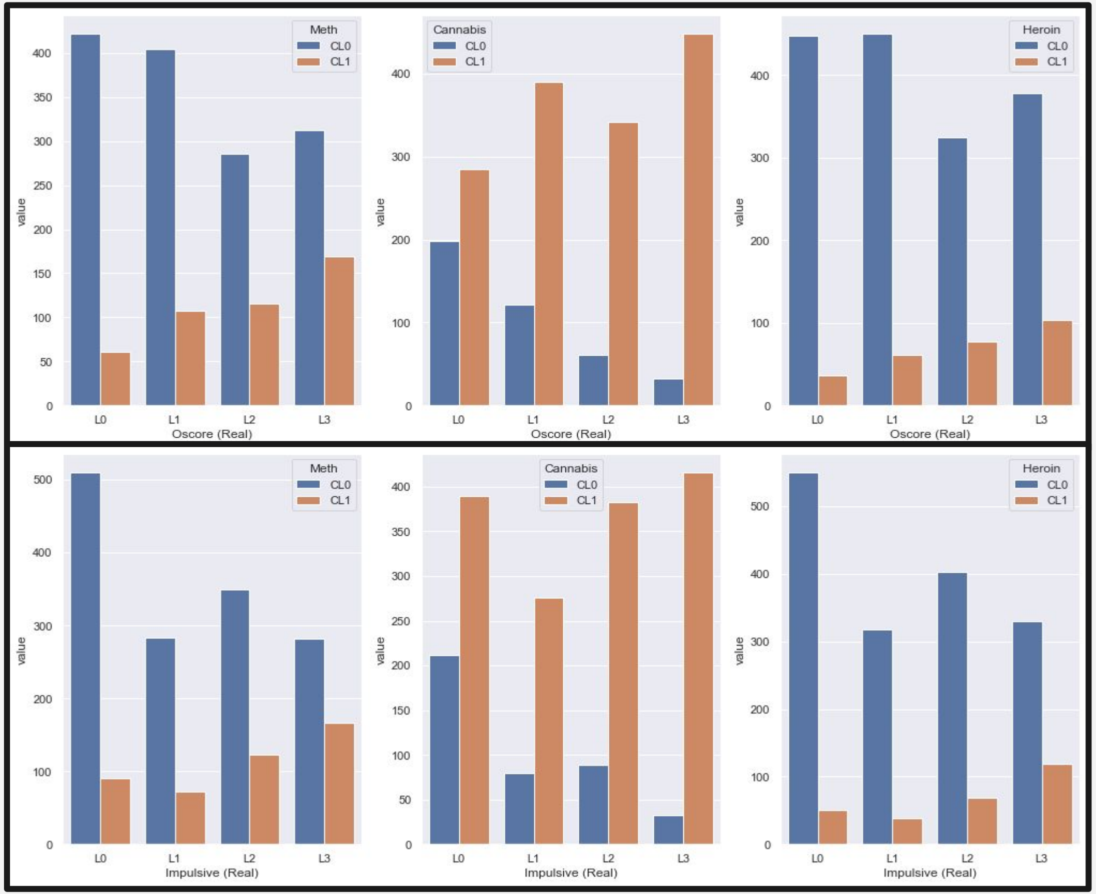


```
L2    562
L1    449
L3    438
L0    428
Name: Escore (Real), dtype: int64
```

03

EDA & Insights

Multiple Barplots



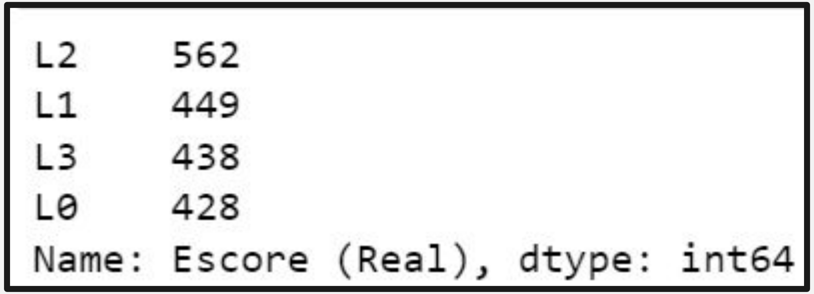
- Compare the relationship between usage of drugs and the spectrum of the personality factor that the individual lies on

- This is done for every single personality factor against every high consumption drugs

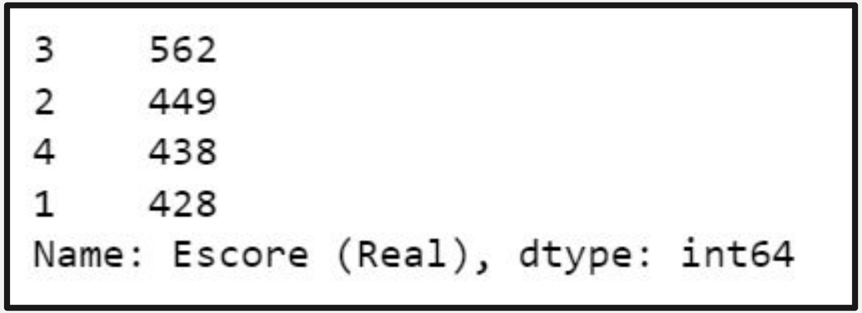
- Eg: For people with a higher Oscore, they generally see an increase in drug usage across all drug types

Converting Categorical Values to Numerical Categorical Format

- **Prepare the dataset for Machine Learning & Chi-Squared EDA by converting all categorical values to numerical values**



```
L2    562
L1    449
L3    438
L0    428
Name: Escore (Real), dtype: int64
```



```
3    562
2    449
4    438
1    428
Name: Escore (Real), dtype: int64
```

Chi-Squared Statistics

p_Value ~ 0	Nscore	Escore	Oscore	Ascore	Cscore	Impulsiveness
Meth	67.3386	22.6983	74.4518	40.7313	73.2611	72.2018
Cannabis	25.7094	10.7557	177.5247	36.8214	106.9018	120.5951
Heroin	61.2076	10.6299	46.8324	28.6642	51.3036	72.6781

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- All the personality factors were relevant in predicting drug usages for Meth, Cannabis & Heroin

Insights Gained

Bar-Graph

- ❖ An **increase** in NScore, Oscore & Impulsiveness correlated to a **increase** in consumption of drugs
- ❖ An **decrease** in EScore & Ascore correlated to a **decrease** in consumption of drugs

Chi-Squared

- ❖ Values for Chi-Squared obtained showcases that all the personality factors are **relevant in predicting drug usage** due to good association between the categorical values

04

ML Techniques

ML TECHNIQUES

SUPPORT VECTOR

Works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes

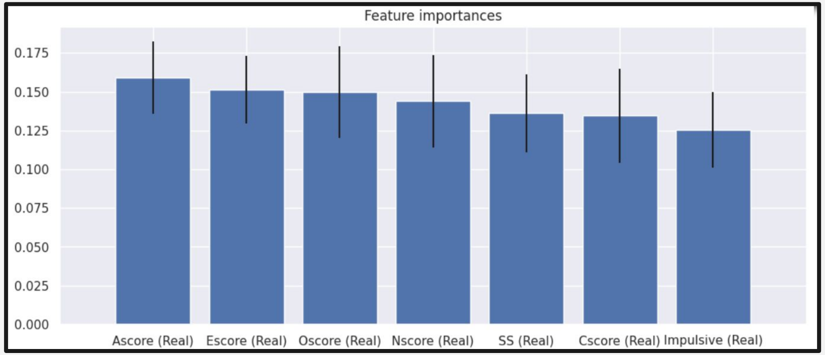
RANDOM FOREST

Combines the output of multiple decision trees to reach a single result. It is able to handle both classification and regression problems

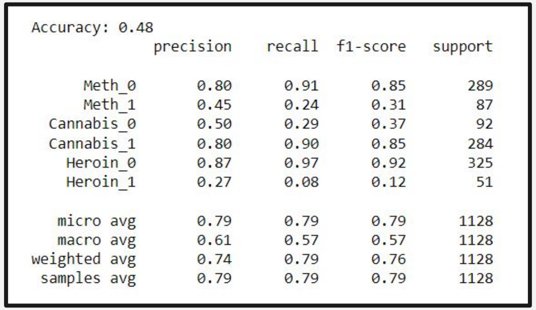
NEURAL NETWORK

A series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates

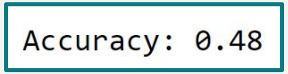
RANDOM FOREST



Graph shows the importance of each of the personality factors in creating the random forest classifier model



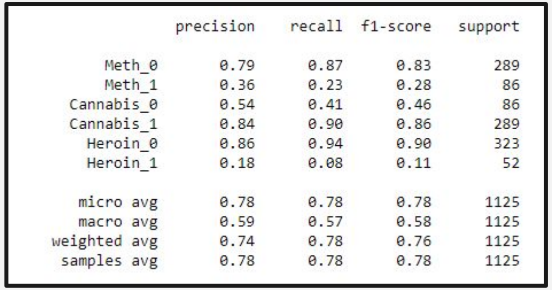
```
Accuracy: 0.48
              precision    recall  f1-score   support

      Meth_0       0.80      0.91      0.85       289
      Meth_1       0.45      0.24      0.31        87
  Cannabis_0       0.50      0.29      0.37        92
  Cannabis_1       0.80      0.90      0.85       284
    Heroin_0       0.87      0.97      0.92       325
    Heroin_1       0.27      0.08      0.12        51

   micro avg       0.79      0.79      0.79      1128
   macro avg       0.61      0.57      0.57      1128
weighted avg       0.74      0.79      0.76      1128
 samples avg       0.79      0.79      0.79      1128
```



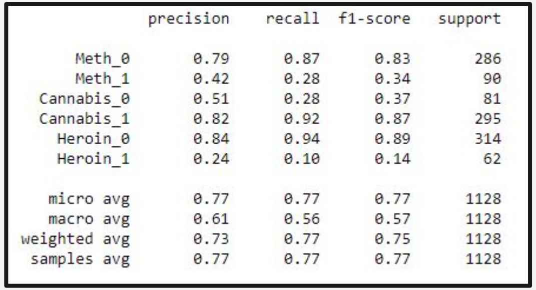
```
Accuracy: 0.48
```

Results of random forest classifier using random train test split

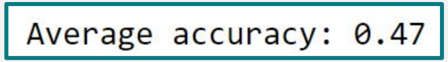
KFOLD and SHUFFLESPLIT



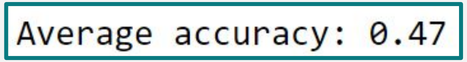
```
              precision    recall  f1-score   support

      Meth_0       0.79      0.87      0.83       289
      Meth_1       0.36      0.23      0.28        86
  Cannabis_0       0.54      0.41      0.46        86
  Cannabis_1       0.84      0.90      0.86       289
    Heroin_0       0.86      0.94      0.90       323
    Heroin_1       0.18      0.08      0.11        52

   micro avg       0.78      0.78      0.78      1125
   macro avg       0.59      0.57      0.58      1125
weighted avg       0.74      0.78      0.76      1125
 samples avg       0.78      0.78      0.78      1125
```



```
              precision    recall  f1-score   support

      Meth_0       0.79      0.87      0.83       286
      Meth_1       0.42      0.28      0.34        90
  Cannabis_0       0.51      0.28      0.37        81
  Cannabis_1       0.82      0.92      0.87       295
    Heroin_0       0.84      0.94      0.89       314
    Heroin_1       0.24      0.10      0.14        62

   micro avg       0.77      0.77      0.77      1128
   macro avg       0.61      0.56      0.57      1128
weighted avg       0.73      0.77      0.75      1128
 samples avg       0.77      0.77      0.77      1128
```



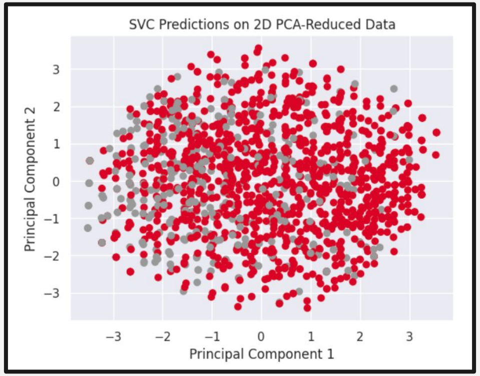
```
Average accuracy: 0.47
```

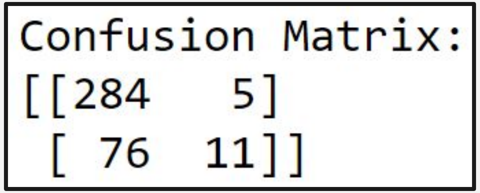


```
Average accuracy: 0.47
```

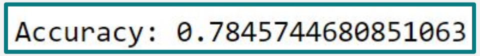
SUPPORT VECTOR

Meth



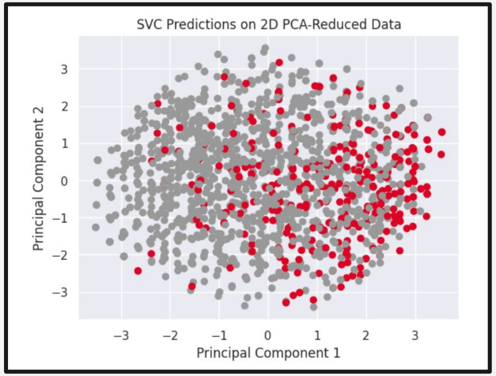


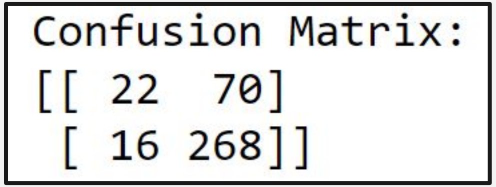
```
Confusion Matrix:
[[284   5]
 [ 76  11]]
```



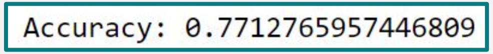
```
Accuracy: 0.7845744680851063
```

Cannabis



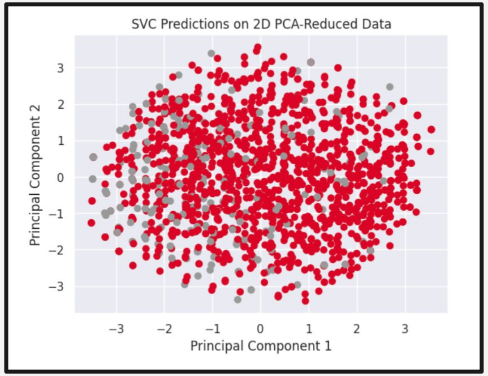


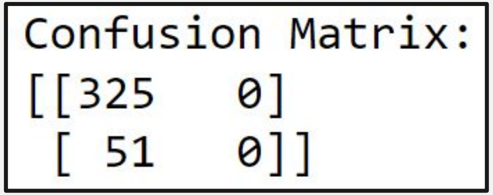
```
Confusion Matrix:
[[ 22  70]
 [ 16 268]]
```



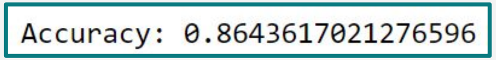
```
Accuracy: 0.7712765957446809
```

Heroin





```
Confusion Matrix:
[[325   0]
 [ 51   0]]
```

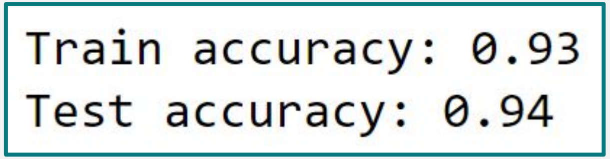


```
Accuracy: 0.8643617021276596
```

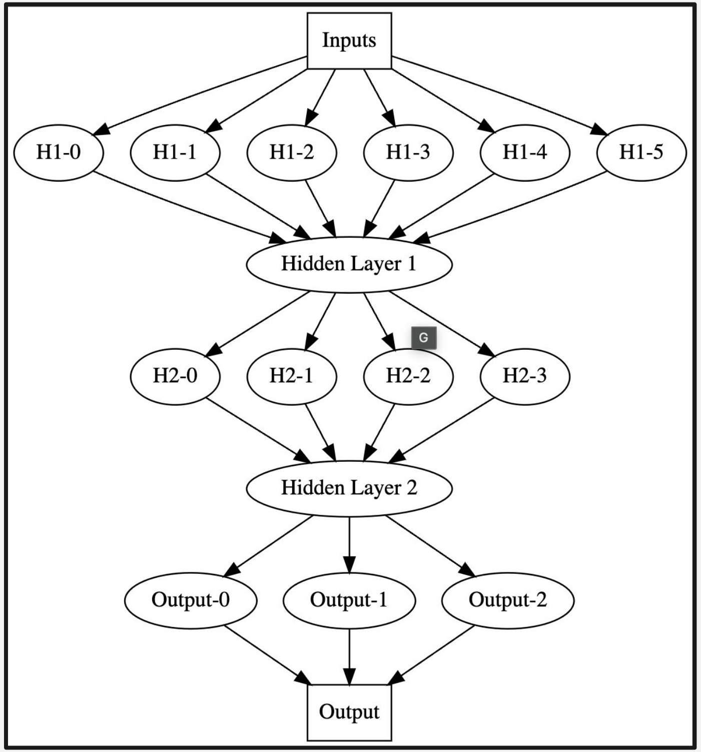
NEURAL NETWORK

- A series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way human brain operates

- 2 Hidden layers which contain 6 nodes and 4 nodes respectively and uses the keras library to train



```
Train accuracy: 0.93
Test accuracy: 0.94
```



05

Conclusion

CONCLUSION

CHI-SQUARED

All personality factors are useful in predicting drug usage

IMPORTANCE

Ascore (real) is the most important personality factor in predicting drug usage, while least important are Cscore (real) and Impulsive (real)

BAR-GRAPH

Increase in Nscore, Oscore, Impulsiveness led to increase in drug usage, while decrease in Escore & Ascore correlated to decrease in usage of drugs

CLASS IMBALANCE

High class imbalance could lead to worse performance of ML models since minority groups are underrepresented

NEURAL NETWORKS

Neural network allows us to most accurately predict drug usage based on personality factors

THANK YOU!