Proactive Risk Assessment of Substance Abuse Using Multi-Class Machine Learning

Andrew Chen 018034679

Meghana Kalle 018045066

CS131: Processing Big Data

May 21 2025

Introduction

Substance abuse is a problem that has plagued society since ancient times and dangerous drugs are more accessible now more than ever in the United States. In 2022 alone, more than 100,000 people in the U.S. died from drug overdoses (NIDA). The 2023 National Survey on Drug Use Health reported that 48.5 million Americans aged 12 or older had a substance use disorder within the last year, including 28.9 million people with an alcohol use disorder and 27.2 million people with a drug use disorder (American Addiction Centers). Substance abuse is also expensive, with the rise of opioid cases during the coronavirus pandemic costing the United States about $1.5 trillion in 2020 (United States Joint Economic Committee). Substance abuse is a crisis that affects millions of people and contributes to increased healthcare costs, loss of productivity, and increased crime rates.

Current methods for identifying individuals at risk of substance abuse include screening, a mainly reactive and self-reported assessment used to detect people who are at risk for developing a substance abuse disorder. However, screening isn't a completely accurate process, as they can fall victim to human bias and underreporting, meaning that at-risk individuals may remain undetected until it is too late to help them.

By using machine learning, risk assessment tools can provide proactive insights rather than reactive responses, allowing healthcare professionals to respond to substance abuse cases more quickly and effectively. We set out to apply Multi-Layer Perceptrons, Random Forest, and Logistic Regression models to create accurate models in order to identify the risk of future substance consumption based on an individual's demographics and psychological data.

Literature Review

"Analysis of substance use and its outcomes by machine learning I. Childhood Evaluation of Liability to Substance Use Disorder" hypothesizes that machine learning can identify the health, psychological, psychiatric, and contextual features to predict the risk of developing a substance use disorder in participants between the ages of 10-22. (Jing) They found that the Random Forest algorithm used in their studies was able to identify 30 psychological, health, environmental and social behavior features that predict a substance abuse disorder, based on questionnaires like the Antisocial Personality Disorder Interview, Constructive Thinking Inventory, Emotional Susceptibility Scale, Sensation Seeking Scale, as well as others.

Methodology

Dataset Overview

The dataset used to train the models come from UC Irvine's Machine Learning Repository (Fehrman) and includes information about the 1884 responses from an online survey. There are 32 total features in the dataset including: ID; age range; gender; education level; country of origin; ethnicity; personality scores from the NEO-FFI-R to measure neuroticism (Nscore), extraversion (Escore), openness to experience (Oscore), agreeableness (A), conscientiousness (C); Barratt Impulsiveness Scale scores to measure impulsiveness (Impulsive); Impulsiveness Sensation-Seeking scores to measure sensation-seeking traits in respondents (ImpSS); and usage data of 19 total drugs. The drugs included are Alcohol, Amphetamines (Amphet), Amyl Nitrite (Amyl), Benzodiazepine (Benzos), Caffeine (Caff), Cannabis, Chocolate (Choc), Cocaine (Coke), Crack, Ecstasy, Heroin, Ketamine, Legal Highs (Legalh), LSD, Magic Mushrooms (Mushrooms), Nicotine, Semeron (Semer), and Volatile Substance Abuse (VSA). It should be noted that Semeron is a fictitious drug that was included to identify usage over-claimers.

Substance usage was quantified using a scale: "CL0" indicates that the respondent never used the substance before, "CL1" indicates that the respondent had last used the substance over a decade ago, "CL2" indicates that the respondent had last used the substance within the last decade, "CL3" indicates that the respondent had used the substance within the last year, "CL4" indicates that the respondent had last used the substance within the last month, "CL5" indicates that the respondent had last used the substance within the last week, and "CL6" indicates that the respondent had last used the substance within the last day.

ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS
2	25-34	M	Doctorate degree	UK	White	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575
3	35-44	M	Professional certificate/ diploma	UK	White	-0.46725	0.80523	-0.84732	-1.6209	-1.0145	-1.37983	0.40148
4	18-24	F	Masters degree	UK	White	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084
5	35-44	F	Doctorate degree	UK	White	0.73545	-1.6334	-0.45174	-0.30172	1.30612	-0.21712	-0.21575
6	65+	F	Left school at 18 years	Canada	White	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858
7	45-54	M	Masters degree	USA	White	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987
8	35-44	M	Left school at 16 years	UK	White	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593
9	35-44	F	Professional certificate/ diploma	Canada	White	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858
10	55-64	M	Masters degree	UK	White	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637
11	25-34	F	University degree	UK	White	-1.05308	0.80523	-1.11902	-0.76096	1.81175	0.19268	0.07987
12	45-54	M	Some college or university, no certificate or degree	Other	White	-1.32828	0.00332	0.14143	-1.92595	-0.52745	0.52975	1.2247
13	55-64	F	University degree	UK	White	2.28554	0.16767	0.44585	-1.6209	-0.78155	1.29221	0.07987
14	55-64	F	Professional certificate/ diploma	Canada	White	-0.79151	0.80523	-0.01928	0.94156	3.46436	-0.71126	-0.84637
15	55-64	F	Professional certificate/ diploma	UK	White	-0.92104	1.45421	0.44585	-0.60633	1.63088	1.29221	0.7654
16	55-64	M	University degree	UK	White	-2.05048	-1.50796	-1.55521	-1.07533	1.13407	-0.71126	-0.52593
17	35-44	F	Some college or university, no certificate or degree	UK	White	-1.55078	-0.80615	-1.68062	0.28783	0.7583	-0.21712	-2.07848
18	45-54	M	Left school at 16 years	UK	White	0.52135	-1.23177	-0.31776	-0.45321	-1.38502	-1.37983	-0.84637
19	55-64	M	University degree	Australia	White	1.37297	-0.15487	-0.17779	-1.92595	-1.5184	-0.71126	-0.21575
20	35-44	M	Professional certificate/ diploma	UK	White	-0.34799	-1.7625	-2.39883	-1.92595	0.7583	-1.37983	-2.07848

Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroom	Nicotine	Semer	VSA
CL5	CL2	CL2	CL0	CL6	CL4	CL6	CL3	CL0	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0	CL0
CL6	CL0	CL0	CL0	CL6	CL3	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0	CL0
CL4	CL0	CL0	CL3	CL5	CL2	CL4	CL2	CL0	CL0	CL0	CL2	CL0	CL0	CL0	CL0	CL2	CL0	CL0
CL4	CL1	CL1	CL0	CL6	CL3	CL6	CL0	CL0	CL1	CL0	CL0	CL1	CL0	CL0	CL2	CL2	CL0	CL0
CL2	CL0	CL0	CL0	CL6	CL0	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
CL6	CL0	CL0	CL0	CL6	CL1	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
CL5	CL0	CL0	CL0	CL6	CL0	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
CL4	CL0	CL0	CL0	CL6	CL0	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
CL6	CL1	CL0	CL1	CL6	CL1	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
CL5	CL0	CL1	CL0	CL6	CL2	CL5	CL2	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL0	CL1
CL5	CL1	CL0	CL0	CL6	CL4	CL5	CL2	CL0	CL3	CL0	CL0	CL0	CL1	CL0	CL2	CL6	CL0	CL0
CL5	CL1	CL0	CL4	CL6	CL3	CL5	CL1	CL0	CL0	CL0	CL0	CL0	CL1	CL1	CL1	CL6	CL0	CL0
CL1	CL0	CL0	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0
CL6	CL0	CL0	CL0	CL6	CL0	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
CL5	CL2	CL2	CL0	CL6	CL1	CL5	CL2	CL0	CL1	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
CL6	CL0	CL0	CL1	CL6	CL3	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
CL6	CL1	CL1	CL0	CL6	CL6	CL4	CL1	CL0	CL1	CL0	CL2	CL0	CL1	CL0	CL1	CL6	CL0	CL0
CL6	CL2	CL0	CL2	CL6	CL3	CL6	CL2	CL0	CL2	CL0	CL0	CL2	CL1	CL0	CL1	CL0	CL0	CL0
CL4	CL1	CL0	CL0	CL6	CL1	CL6	CL0	CL0	CL1	CL0	CL0	CL0	CL0	CL6	CL0	CL1	CL0	CL0

Data Processing:

For the educational level feature, there are some rows with the values “Some college or university, no certificate or degree”. However, the presence of the comma would mess up with later processing commands, so these values are changed to “Some college”.

```
mekalle@CS131A:~/cs131/project$ sed 's/"Some college or university, no certificate or degree"/
Some college/g' ./data/original_dataset.csv > ./data/temp2.csv
```

Before:

ID	Age	Gender	Education	Country	Ethnicity
2	25-34	M	Doctorate degree	UK	White
3	35-44	M	Professional certificate/ diploma	UK	White
4	18-24	F	Masters degree	UK	White
5	35-44	F	Doctorate degree	UK	White
6	65+	F	Left school at 18 years	Canada	White
7	45-54	M	Masters degree	USA	White
8	35-44	M	Left school at 16 years	UK	White
9	35-44	F	Professional certificate/ diploma	Canada	White
10	55-64	M	Masters degree	UK	White
11	25-34	F	University degree	UK	White
12	45-54	M	Some college or university, no certificate or degree	Other	White
13	55-64	F	University degree	UK	White
14	55-64	F	Professional certificate/ diploma	Canada	White
15	55-64	F	Professional certificate/ diploma	UK	White
16	55-64	M	University degree	UK	White
17	35-44	F	Some college or university, no certificate or degree	UK	White
18	45-54	M	Left school at 16 years	UK	White
19	55-64	M	University degree	Australia	White

After:

ID	Age	Gender	Education	Country	Ethnicity
2	25-34	M	Doctorate degree	UK	White
3	35-44	M	Professional certificate/ diploma	UK	White
4	18-24	F	Masters degree	UK	White
5	35-44	F	Doctorate degree	UK	White
6	65+	F	Left school at 18 years	Canada	White
7	45-54	M	Masters degree	USA	White
8	35-44	M	Left school at 16 years	UK	White
9	35-44	F	Professional certificate/ diploma	Canada	White
10	55-64	M	Masters degree	UK	White
11	25-34	F	University degree	UK	White
12	45-54	M	Some college	Other	White
13	55-64	F	University degree	UK	White
14	55-64	F	Professional certificate/ diploma	Canada	White
15	55-64	F	Professional certificate/ diploma	UK	White
16	55-64	M	University degree	UK	White
17	35-44	F	Some college	UK	White
18	45-54	M	Left school at 16 years	UK	White
19	55-64	M	University degree	Australia	White

Next, entries that had claimed to use Semeron before were removed and the entire "Semer" column was removed as well. This reduces the total number of columns in the dataset from 32 to 31 and the total number of rows from 1885 to 1877.

```
mekalle@CS131A:~/cs131/project$ (head -n1 ./data/temp2.csv && tail -n +2 ./data/temp2.csv| awk -F',' '{if ($31 == "CL0") print $0}') > ./data/temp3.csv
mekalle@CS131A:~/cs131/project$ paste -d',' <(cut -d',' -f1-30 ./data/temp3.csv) <(cut -d',' -f32 ./data/temp3.csv) > ./data/remove_semer.csv
```

Before:

1820	18-24	M	Masters degree	USA	White	-2.75696	2.57309	1.24033	0.28783	1.46191	-0.21712	1.2247	CL6	CL3	CL2	CL3	CL6	CL5	CL6	CL2	CL0	CL5	CL0	CL0	CL3	CL5	CL2	CL3	CL6	CL0	CL2
1821	18-24	M	Some college	USA	White	1.60383	-2.03972	1.43533	-0.60633	-2.72827	1.86203	-0.21575	CL2	CL0	CL0	CL2	CL6	CL6	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL0	CL0
1822	45-54	F	Some college	USA	White	1.37297	-0.15487	0.44585	0.43852	-0.65253	1.86203	-0.52593	CL3	CL0	CL0	CL0	CL6	CL4	CL6	CL1	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL1	CL3	CL0	CL0
1823	18-24	M	Some college	USA	White	-0.79151	0.32197	1.88511	-1.07533	-0.78155	-0.21712	0.40148	CL3	CL3	CL0	CL3	CL5	CL6	CL5	CL3	CL0	CL3	CL0	CL3	CL5	CL3	CL0	CL3	CL0	CL0	CL0
1824	18-24	F	Some college	USA	White	1.23461	-1.23177	1.06238	-0.30172	-2.30408	0.88113	-0.21575	CL4	CL2	CL2	CL2	CL6	CL3	CL4	CL0	CL0	CL2	CL0	CL0	CL2	CL2	CL2	CL2	CL2	CL0	CL0
1825	25-34	M	University degree	UK	Asian	-0.46725	-1.50796	-0.58331	-0.30172	-0.40581	0.52975	0.40148	CL6	CL0	CL0	CL0	CL0	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
1826	18-24	M	University degree	USA	White	0.04257	-1.09207	0.14143	-1.34289	-0.00665	-0.21712	-0.84637	CL5	CL5	CL0	CL2	CL6	CL6	CL5	CL2	CL0	CL5	CL0	CL4	CL2	CL3	CL3	CL3	CL3	CL0	CL0
1827	18-24	F	University degree	USA	White	0.22393	-0.30033	0.88309	1.2861	-0.00665	0.88113	0.07987	CL4	CL0	CL0	CL2	CL5	CL6	CL5	CL2	CL0	CL0	CL0	CL0	CL2	CL3	CL0	CL3	CL5	CL2	CL0
1828	25-34	F	University degree	USA	White	-1.05308	-1.50796	0.14143	0.28783	0.25953	-0.21712	0.7654	CL4	CL1	CL0	CL2	CL6	CL5	CL5	CL2	CL1	CL0	CL0	CL0	CL2	CL1	CL2	CL2	CL6	CL0	CL1
1829	25-34	F	Masters degree	USA	White	-0.34799	1.11406	0.44585	-2.5383	-0.14277	0.88113	1.2247	CL5	CL3	CL0	CL6	CL6	CL6	CL4	CL4	CL0	CL5	CL0	CL0	CL3	CL0	CL0	CL5	CL2	CL0	CL0
1830	25-34	F	University degree	USA	Other	0.82562	-1.6334	-0.31776	-1.47955	-1.13788	-1.37983	-1.18084	CL2	CL1	CL0	CL6	CL1	CL5	CL3	CL1	CL1	CL0	CL1	CL0	CL0	CL1	CL3	CL0	CL1	CL0	CL0

After:

1820	18-24	M	Masters d	USA	White	-2.75696	2.57309	1.24033	0.28783	1.46191	-0.21712	1.2247	CL6	CL3	CL2	CL3	CL6	CL5	CL6	CL2	CL0	CL5	CL0	CL0	CL3	CL5	CL2	CL3	CL6	CL2
1821	18-24	M	Some coll	USA	White	1.60383	-2.03972	1.43533	-0.60633	-2.72827	1.86203	-0.21575	CL2	CL0	CL0	CL2	CL6	CL6	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL0
1822	45-54	F	Some coll	USA	White	1.37297	-0.15487	0.44585	0.43852	-0.65253	1.86203	-0.52593	CL3	CL0	CL0	CL0	CL6	CL4	CL6	CL1	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL1	CL3	CL0
1823	18-24	M	Some coll	USA	White	-0.79151	0.32197	1.88511	-1.07533	-0.78155	-0.21712	0.40148	CL3	CL3	CL0	CL3	CL5	CL6	CL5	CL3	CL0	CL3	CL0	CL3	CL5	CL3	CL0	CL3	CL0	CL0
1824	18-24	F	Some coll	USA	White	1.23461	-1.23177	1.06238	-0.30172	-2.30408	0.88113	-0.21575	CL4	CL2	CL2	CL2	CL6	CL3	CL4	CL0	CL0	CL2	CL0	CL0	CL2	CL2	CL2	CL2	CL2	CL0
1825	25-34	M	University	UK	Asian	-0.46725	-1.50796	-0.58331	-0.30172	-0.40581	0.52975	0.40148	CL6	CL0	CL0	CL0	CL0	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
1826	18-24	M	University	USA	White	0.04257	-1.09207	0.14143	-1.34289	-0.00665	-0.21712	-0.84637	CL5	CL5	CL0	CL2	CL6	CL6	CL5	CL2	CL0	CL5	CL0	CL4	CL2	CL3	CL3	CL3	CL3	CL0
1828	25-34	F	University	USA	White	-1.05308	-1.50796	0.14143	0.28783	0.25953	-0.21712	0.7654	CL4	CL1	CL0	CL2	CL6	CL5	CL5	CL2	CL1	CL0	CL0	CL0	CL2	CL1	CL2	CL2	CL6	CL1
1829	25-34	F	Masters d	USA	White	-0.34799	1.11406	0.44585	-2.5383	-0.14277	0.88113	1.2247	CL5	CL3	CL0	CL6	CL6	CL6	CL4	CL4	CL0	CL5	CL0	CL0	CL3	CL0	CL0	CL5	CL2	CL0
1830	25-34	F	University	USA	Other	0.82562	-1.6334	-0.31776	-1.47955	-1.13788	-1.37983	-1.18084	CL2	CL1	CL0	CL6	CL1	CL5	CL3	CL1	CL1	CL0	CL1	CL0	CL0	CL1	CL3	CL0	CL1	CL0

Ages were converted from ranges to numbers using ordinal encoding.

```
mekalle@CS131A:~/cs131/project$ awk -F',' '{\
  if ($2 == "18-24") $2 = 1;\
  if ($2 == "25-34") $2 = 2;\
  if ($2 == "35-44") $2 = 3;\
  if ($2 == "45-54") $2 = 4;\
  if ($2 == "55-64") $2 = 5;\
  if ($2 == "65+") $2 = 6;\
  print $0
}' OFS="," ./data/remove_semer.csv > ./data/temp6.csv
```

Before:

ID	Age	Gender	Education	Country	Ethnicity
2	25-34	M	Doctorate degree	UK	White
3	35-44	M	Professional certificate/ diploma	UK	White
4	18-24	F	Masters degree	UK	White
5	35-44	F	Doctorate degree	UK	White
6	65+	F	Left school at 18 years	Canada	White
7	45-54	M	Masters degree	USA	White
8	35-44	M	Left school at 16 years	UK	White
9	35-44	F	Professional certificate/ diploma	Canada	White
10	55-64	M	Masters degree	UK	White
11	25-34	F	University degree	UK	White
12	45-54	M	Some college	Other	White
13	55-64	F	University degree	UK	White
14	55-64	F	Professional certificate/ diploma	Canada	White

After:

ID	Age	Gender	Education	Country	Ethnicity
2	2	M	Doctorate degree	UK	White
3	3	M	Professional certificate/ diploma	UK	White
4	1	F	Masters degree	UK	White
5	3	F	Doctorate degree	UK	White
6	6	F	Left school at 18 years	Canada	White
7	4	M	Masters degree	USA	White
8	3	M	Left school at 16 years	UK	White
9	3	F	Professional certificate/ diploma	Canada	White
10	5	M	Masters degree	UK	White
11	2	F	University degree	UK	White
12	4	M	Some college	Other	White
13	5	F	University degree	UK	White
14	5	F	Professional certificate/ diploma	Canada	White

Education level was converted from text to numbers using ordinal encoding.

```
mekalle@CS131A:~/cs131/project$ awk -F',' '{\
  if ($4 == "Left school at 16 years") $4 = 2;\
  if ($4 == "Left school at 17 years") $4 = 3;\
  if ($4 == "Left school at 18 years") $4 = 4;\
  if ($4 == "Left school before 16 years") $4 = 1;\
  if ($4 == "Professional certificate/ diploma") $4 = 5;\
  if ($4 == "Some college") $4 = 6;\
  if ($4 == "University degree") $4 = 7;\
  if ($4 == "Masters degree") $4 = 8;\
  if ($4 == "Doctorate degree") $4 = 9;\
  print $0\
}' OFS=',' ./data/temp6.csv > ./data/temp7.csv
```

Before:

ID	Age	Gender	Education	Country	Ethnicity
2	2	M	Doctorate degree	UK	White
3	3	M	Professional certificate/ diploma	UK	White
4	1	F	Masters degree	UK	White
5	3	F	Doctorate degree	UK	White
6	6	F	Left school at 18 years	Canada	White
7	4	M	Masters degree	USA	White
8	3	M	Left school at 16 years	UK	White
9	3	F	Professional certificate/ diploma	Canada	White
10	5	M	Masters degree	UK	White
11	2	F	University degree	UK	White
12	4	M	Some college	Other	White
13	5	F	University degree	UK	White
14	5	F	Professional certificate/ diploma	Canada	White

After:

ID	Age	Gender	Education	Country	Ethnicity
2	2	M	9	UK	White
3	3	M	5	UK	White
4	1	F	8	UK	White
5	3	F	9	UK	White
6	6	F	4	Canada	White
7	4	M	8	USA	White
8	3	M	2	UK	White
9	3	F	5	Canada	White
10	5	M	8	UK	White
11	2	F	7	UK	White
12	4	M	6	Other	White
13	5	F	7	UK	White
14	5	F	5	Canada	White

One-hot encoding was used to convert country of origin values from text to numbers. Doing this appended the changed values to the end of the dataset, after the target features.

```
awk -F',' '
BEGIN {
  OFS = ",";
  split("Australia,Canada,New_Zealand,Other,Republic_of_Ireland,UK,USA", countries, ",");
}
NR == 1 {
  for (i = 1; i <= NF; i++) {
    gsub(/\r/, "", $i);
    if ($i == "Country") {
      country_col = i;
    } else {
      header = (header ? header OFS : "") $i;
    }
  }
  for (j in countries) {
    header = header OFS "country_" countries[j];
  }
  print header;
}
NR > 1 {
  gsub(/\r/, "", $0);
  original_country = $country_col;
  row = "";
  for (i = 1; i <= NF; i++) {
    if (i != country_col) {
      row = (row ? row OFS : "") $i;
    }
  }
  for (j in countries) {
    row = row OFS ((countries[j] == original_country) ? 1 : 0);
  }
  print row;
}
' ./data/temp7.csv > ./data/one_hot_country.csv
```

```
mekalle@CS131A:~/cs131/project$ chmod +x ./scripts/one_hot_country.txt
mekalle@CS131A:~/cs131/project$ ./scripts/one_hot_country.txt
```

Before:

ID	Age	Gender	Education	Country	Ethnicity
2	2	M	9	UK	White
3	3	M	5	UK	White
4	1	F	8	UK	White
5	3	F	9	UK	White
6	6	F	4	Canada	White
7	4	M	8	USA	White
8	3	M	2	UK	White
9	3	F	5	Canada	White
10	5	M	8	UK	White
11	2	F	7	UK	White
12	4	M	6	Other	White
13	5	F	7	UK	White
14	5	F	5	Canada	White

After:

LSD	Meth	Mushroon	Nicotine	VSA	country_New_Zealand	country_UK	country_Republic_of_Ireland	country_Canada	country_Australia	country_Othe	country_USA
CL2	CL3	CL0	CL4	CL0	0	1	0	0	0	0	0
CL0	CL0	CL1	CL0	CL0	0	1	0	0	0	0	0
CL0	CL0	CL0	CL2	CL0	0	1	0	0	0	0	0
CL0	CL0	CL2	CL2	CL0	0	1	0	0	0	0	0
CL0	CL0	CL0	CL6	CL0	0	0	0	1	0	0	0
CL0	CL0	CL0	CL6	CL0	0	0	0	0	0	0	1
CL0	CL0	CL0	CL0	CL0	0	1	0	0	0	0	0
CL0	CL0	CL0	CL6	CL0	0	0	0	1	0	0	0
CL0	CL0	CL0	CL6	CL0	0	1	0	0	0	0	0
CL0	CL0	CL0	CL2	CL1	0	1	0	0	0	0	0
CL1	CL0	CL2	CL6	CL0	0	0	0	0	0	1	0
CL1	CL1	CL1	CL6	CL0	0	1	0	0	0	0	0
CL0	CL0	CL0	CL1	CL0	0	0	0	1	0	0	0

One-hot encoding was also used to convert gender values to either 0 or 1. Doing this appended the changed values to the end of the dataset, after the country of origin columns.

```
mekalle@CS131A:~/cs131/project$ (paste -d ',' <(head -n1 ./data/one_hot_country.csv) <(echo "gender_F,gender_M") &&
tail -n +2 ./data/one_hot_country.csv| awk -F',' '{if ($3 == "F") print $0 ",1,0"; if ($3 == "M") print $0 ",0,1"}'
OFS=",") > ./data/temp8.csv
mekalle@CS131A:~/cs131/project$ paste -d',' <(cut -d',' -f1-2 ./data/temp8.csv) <(cut -d',' -f4-39 ./data/temp8.csv)
 > ./data/one_hot_gender.csv
```

Before:

ID	Age	Gender	Education	Ethnicity
2	2	M	9	White
3	3	M	5	White
4	1	F	8	White
5	3	F	9	White
6	6	F	4	White
7	4	M	8	White
8	3	M	2	White
9	3	F	5	White
10	5	M	8	White
11	2	F	7	White
12	4	M	6	White
13	5	F	7	White
14	5	F	5	White

After:

country_Australia	country_Other	country_USA	gender_F	gender_M
0	0	0	0	1
0	0	0	0	1
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	1	0	1
0	0	0	0	1
0	0	0	1	0
0	0	0	0	1
0	0	0	1	0
0	1	0	0	1
0	0	0	1	0
0	0	0	1	0

One-hot encoding was also used to convert ethnicity values from text to numbers. Doing this appended the changed values to the end of the dataset, after the gender columns.

```
awk -F',' '
BEGIN {
  OFS = ",";
  split("Asian,Black,Mixed-Black/Asian,Mixed-White/Asian,Mixed-White/Black,Other,White", eths, ",");
}
NR == 1 {
  for (i = 1; i <= NF; i++) {
    gsub(/\r/, "", $i);
    if ($i == "Ethnicity") {
      eth_col = i;
    } else {
      header = (header ? header OFS : "") $i;
    }
  }
  for (j in eths) {
    header = header OFS "eth_" eths[j];
  }
  print header;
}
NR > 1 {
  gsub(/\r/, "", $0);
  original_eth = $eth_col;
  row = "";
  for (i = 1; i <= NF; i++) {
    if (i != eth_col) {
      row = (row ? row OFS : "") $i;
    }
  }
  for (j in eths) {
    row = row OFS ((eths[j] == original_eth) ? 1 : 0);
  }
  print row;
}
' ./data/one_hot_gender.csv > ./data/one_hot_eth.csv
```

```
mekalle@CS131A:~/cs131/project$ chmod +x ./scripts/one_hot_eth.txt
mekalle@CS131A:~/cs131/project$ ./scripts/one_hot_eth.txt
```

Before:

ID	Age	Education	Ethnicity
2	2	9	White
3	3	5	White
4	1	8	White
5	3	9	White
6	6	4	White
7	4	8	White
8	3	2	White
9	3	5	White
10	5	8	White
11	2	7	White
12	4	6	White
13	5	7	White
14	5	5	White

After:

gender_F	gender_M	eth_Mixed-Black/Asian	eth_Other	eth_Mixed-White/Black	eth_Black	eth_Asian	eth_Mixed-White/Asian	eth_White
0	1	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1

The drug usage scale was changed from text to numbers and also categorized: with “CL0” and “CL1” changed to 0 to indicate “not at risk”; “CL2” and “CL3” changed to 1 to indicate “potentially at risk”; and “CL4”, “CL5” and “CL6” changed to 2 to indicate “at risk”.

```
mekalle@CS131A:~/cs131/project$ sed -e 's/CL0/0/g' -e 's/CL1/0/g' -e 's/CL2/1/g' -e 's/CL3/1/g'
 -e 's/CL4/2/g' -e 's/CL5/2/g' -e 's/CL6/2/g' ./data/one_hot_eth.csv > ./data/temp8.csv
```

Before:

Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroon	Nicotine	VSA
CL5	CL2	CL2	CL0	CL6	CL4	CL6	CL3	CL0	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0
CL6	CL0	CL0	CL0	CL6	CL3	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0
CL4	CL0	CL0	CL3	CL5	CL2	CL4	CL2	CL0	CL0	CL0	CL2	CL0	CL0	CL0	CL0	CL2	CL0
CL4	CL1	CL1	CL0	CL6	CL3	CL6	CL0	CL0	CL1	CL0	CL0	CL1	CL0	CL0	CL2	CL2	CL0
CL2	CL0	CL0	CL0	CL6	CL0	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0
CL6	CL0	CL0	CL0	CL6	CL1	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0
CL5	CL0	CL0	CL0	CL6	CL0	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
CL4	CL0	CL0	CL0	CL6	CL0	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0
CL6	CL1	CL0	CL1	CL6	CL1	CL6	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0
CL5	CL0	CL1	CL0	CL6	CL2	CL5	CL2	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL1
CL5	CL1	CL0	CL0	CL6	CL4	CL5	CL2	CL0	CL3	CL0	CL0	CL0	CL1	CL0	CL2	CL6	CL0
CL5	CL1	CL0	CL4	CL6	CL3	CL5	CL1	CL0	CL0	CL0	CL0	CL0	CL1	CL1	CL1	CL6	CL0
CL1	CL0	CL0	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0

After:

Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroon	Nicotine	VSA
2	1	1	0	2	2	2	1	0	2	0	1	0	1	1	0	2	0
2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	2	1	2	1	0	0	0	1	0	0	0	0	1	0
2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	1	1	0
1	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	1	2	1	0	0	0	0	0	0	0	0	1	0
2	0	0	0	2	2	2	1	0	1	0	0	0	0	0	1	2	0
2	0	0	2	2	1	2	0	0	0	0	0	0	0	0	0	2	0
0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0

The “ID” column was removed as it isn’t needed when training the models.

```
mekalle@CS131A:~/cs131/project$ cut -d',' -f2-44 ./data/temp8.csv > ./data/temp9.csv
```

Before:

ID	Age	Education	Nscore	Escore
2	2	9	-0.67825	1.93886
3	3	5	-0.46725	0.80523
4	1	8	-0.14882	-0.80615
5	3	9	0.73545	-1.6334
6	6	4	-0.67825	-0.30033
7	4	8	-0.46725	-1.09207
8	3	2	-1.32828	1.93886
9	3	5	0.62967	2.57309
10	5	8	-0.24649	0.00332
11	2	7	-1.05308	0.80523
12	4	6	-1.32828	0.00332
13	5	7	2.28554	0.16767
14	5	5	-0.79151	0.80523

After:

Age	Education	Nscore	Escore
2	9	-0.67825	1.93886
3	5	-0.46725	0.80523
1	8	-0.14882	-0.80615
3	9	0.73545	-1.6334
6	4	-0.67825	-0.30033
4	8	-0.46725	-1.09207
3	2	-1.32828	1.93886
3	5	0.62967	2.57309
5	8	-0.24649	0.00332
2	7	-1.05308	0.80523
4	6	-1.32828	0.00332
5	7	2.28554	0.16767
5	5	-0.79151	0.80523

The targets (substance columns) were moved to the end.

```
mekalle@CS131A:~/cs131/project$ paste -d ',' <(cut -d ',' -f1-9 ./data/temp9.csv) <(cut -d ',' -f28-43 ./data/temp9.csv) <(cut -d ',' -f10-27 ./data/temp9.csv) > ./data/temp10.csv
```

Before:

Age	Education	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy
2	9	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	2	1	1	0	2	2	2	1	0	2
3	5	-0.46725	0.80523	-0.84732	-1.6209	-1.0145	-1.37983	0.40148	2	0	0	0	2	1	2	0	0	0
1	8	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	2	0	0	1	2	1	2	1	0	0
3	9	0.73545	-1.6334	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	2	0	0	0	2	1	2	0	0	0
6	4	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858	1	0	0	0	2	0	2	0	0	0
4	8	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987	2	0	0	0	2	0	2	0	0	0
3	2	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593	2	0	0	0	2	0	2	0	0	0
3	5	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858	2	0	0	0	2	0	2	0	0	0
5	8	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637	2	0	0	0	2	0	2	0	0	0
2	7	-1.05308	0.80523	-1.11902	-0.76096	1.81175	0.19268	0.07987	2	0	0	0	2	1	2	1	0	0
4	6	-1.32828	0.00332	0.14143	-1.92595	-0.52745	0.52975	1.2247	2	0	0	0	2	2	2	1	0	1
5	7	2.28554	0.16767	0.44585	-1.6209	-0.78155	1.29221	0.07987	2	0	0	2	2	1	2	0	0	0
5	5	-0.79151	0.80523	-0.01928	0.94156	3.46436	-0.71126	-0.84637	0	0	0	0	2	0	0	0	0	0

After:

Age	Education	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	country_New_Zealand	country_UK	country_Republic_of_Ireland	country_Canada
2	9	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	0	1	0	0
3	5	-0.46725	0.80523	-0.84732	-1.6209	-1.0145	-1.37983	0.40148	0	1	0	0
1	8	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	0	1	0	0
3	9	0.73545	-1.6334	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	0	1	0	0
6	4	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858	0	0	0	1
4	8	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987	0	0	0	0
3	2	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593	0	1	0	0
3	5	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858	0	0	0	1
5	8	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637	0	1	0	0
2	7	-1.05308	0.80523	-1.11902	-0.76096	1.81175	0.19268	0.07987	0	1	0	0
4	6	-1.32828	0.00332	0.14143	-1.92595	-0.52745	0.52975	1.2247	0	0	0	0
5	7	2.28554	0.16767	0.44585	-1.6209	-0.78155	1.29221	0.07987	0	1	0	0
5	5	-0.79151	0.80523	-0.01928	0.94156	3.46436	-0.71126	-0.84637	0	0	0	1

eth_Asian	eth_Mixed-White/Asian	eth_White	Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroon	Nicotine	VSA
0	0	1	2	1	1	0	2	2	2	1	0	2	0	1	0	1	1	0	2	0
0	0	1	2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	0	0	0
0	0	1	2	0	0	1	2	1	2	1	0	0	0	1	0	0	0	0	1	0
0	0	1	2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	1	1	0
0	0	1	1	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
0	0	1	2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
0	0	1	2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0
0	0	1	2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
0	0	1	2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
0	0	1	2	0	0	0	2	1	2	1	0	0	0	0	0	0	0	0	1	0
0	0	1	2	0	0	0	2	2	2	1	0	1	0	0	0	0	0	1	2	0
0	0	1	2	0	0	2	2	1	2	0	0	0	0	0	0	0	0	0	2	0
0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0

The NEO-FFI-R personality scores, Impulsive score, and ImpSS score were normalized.

```
#!/bin/bash

input_file="./data/temp10.csv"
output_file="./data/processed.csv"
nscore=3
ss=9

awk -F"," -v first_c="$nscore" -v last_c="$ss" 'BEGIN {OFS = ","}
  NR == 1 {
    header = $0;
    next;
  }

  {
    row[NR] = $0;
    split($0, cols, ",");
    row_count[NR] = length(cols);  # Save NF
    for (col = first_c; col <= last_c; col++) {
      val = cols[col] + 0;
      if (!(col in min) || val < min[col]) min[col] = val;
      if (!(col in max) || val > max[col]) max[col] = val;
    }
  }

  END {
    print header;

    for (i = 2; i <= NR; i++) {
      split(row[i], cols, ",");
      for (col = first_c; col <= last_c; col++) {
        val = cols[col] + 0;
        if (min[col] == max[col]) {
          cols[col] = 0;
        } else {
          cols[col] = (val - min[col]) / (max[col] - min[col]);
        }
      }

      for (j = 1; j <= row_count[i]; j++) {
        printf("%s", cols[j]);
        if (j < row_count[i]) printf(",");
      }
      print "";
    }
  }
' "$input_file" > "$output_file"
```

```
mekalle@CS131A:~/cs131/project$ chmod +x ./scripts/normalization.sh
mekalle@CS131A:~/cs131/project$ ./scripts/normalization.sh
```

Before:

Age	Education	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS
2	9	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575
3	5	-0.46725	0.80523	-0.84732	-1.6209	-1.0145	-1.37983	0.40148
1	8	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084
3	9	0.73545	-1.6334	-0.45174	-0.30172	1.30612	-0.21712	-0.21575
6	4	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858
4	8	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987
3	2	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593
3	5	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858
5	8	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637
2	7	-1.05308	0.80523	-1.11902	-0.76096	1.81175	0.19268	0.07987
4	6	-1.32828	0.00332	0.14143	-1.92595	-0.52745	0.52975	1.2247
5	7	2.28554	0.16767	0.44585	-1.6209	-0.78155	1.29221	0.07987
5	5	-0.79151	0.80523	-0.01928	0.94156	3.46436	-0.71126	-0.84637

After:

Age	Education	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS
2	9	0.413474	0.796106	0.762567	0.609827	0.479394	0.33792	0.465658
3	5	0.444788	0.622976	0.392939	0.266061	0.35358	0.215401	0.619957
1	8	0.492045	0.376883	0.527023	0.585213	0.584415	0.215401	0.224398
3	9	0.623275	0.250544	0.456995	0.456454	0.688508	0.428474	0.465658
6	4	0.413474	0.454133	0.278311	0.794386	0.73538	0.215401	0.132468
4	8	0.444788	0.333217	0.456995	0.456454	0.635594	0.428474	0.539559
3	2	0.317006	0.796106	0.392939	0.456454	0.73538	0.503573	0.388117
3	5	0.607577	0.892967	0.372052	0.609827	0.663677	0.215401	0.132468
5	8	0.47755	0.500507	0.299519	0.585213	0.517797	0.215401	0.308011
2	7	0.357847	0.622976	0.348943	0.390173	0.761484	0.503573	0.539559
4	6	0.317006	0.500507	0.553046	0.222034	0.423875	0.565343	0.825752
5	7	0.853317	0.525607	0.602341	0.266061	0.387201	0.705068	0.539559
5	5	0.396666	0.622976	0.527023	0.635892	1	0.33792	0.308011

Processed Dataset:

Age	Education	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	country_N	country_U	country_R	country_C	country_A	country_O	country_U	gender_F	gender_M	eth_Mixed	eth_Other	eth_Mixed	eth_Black	eth_Asian	eth_Mixed	eth_White
2	9	0.413474	0.796106	0.762567	0.609827	0.479394	0.33792	0.465658	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
3	5	0.444788	0.622976	0.392939	0.266061	0.35358	0.215401	0.619957	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
1	8	0.492045	0.376883	0.527023	0.585213	0.584415	0.215401	0.224398	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
3	9	0.623275	0.250544	0.456995	0.456454	0.688508	0.428474	0.465658	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
6	4	0.413474	0.454133	0.278311	0.794386	0.73538	0.215401	0.132468	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
4	8	0.444788	0.333217	0.456995	0.456454	0.635594	0.428474	0.539559	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1
3	2	0.317006	0.796106	0.392939	0.456454	0.73538	0.503573	0.388117	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
3	5	0.607577	0.892967	0.372052	0.609827	0.663677	0.215401	0.132468	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
5	8	0.47755	0.500507	0.299519	0.585213	0.517797	0.215401	0.308011	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
2	7	0.357847	0.622976	0.348943	0.390173	0.761484	0.503573	0.539559	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
4	6	0.317006	0.500507	0.553046	0.222034	0.423875	0.565343	0.825752	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
5	7	0.853317	0.525607	0.602341	0.266061	0.387201	0.705068	0.539559	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
5	5	0.396666	0.622976	0.527023	0.635892	1	0.33792	0.308011	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
5	5	0.377443	0.722089	0.602341	0.41249	0.73538	0.705068	0.710933	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
5	7	0.209828	0.269702	0.278311	0.344801	0.663677	0.33792	0.388117	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
3	6	0.283986	0.376883	0.258003	0.541542	0.609443	0.428474	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
4	2	0.591502	0.311882	0.47869	0.43459	0.300104	0.215401	0.308011	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
5	7	0.717887	0.476348	0.501355	0.222034	0.280854	0.33792	0.465658	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
3	5	0.462487	0.230828	0.141704	0.222034	0.609443	0.215401	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1

Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroon	Nicotine	VSA
2	1	1	0	2	2	2	1	0	2	0	1	0	1	1	0	2	0
2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	2	1	2	1	0	0	0	1	0	0	0	0	1	0
2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	1	1	0
1	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	1	2	1	0	0	0	0	0	0	0	0	1	0
2	0	0	0	2	2	2	1	0	1	0	0	0	0	0	1	2	0
2	0	0	2	2	1	2	0	0	0	0	0	0	0	0	0	2	0
0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0
2	1	1	0	2	0	2	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	1	2	0	0	0	0	0	0	0	0	0	2	0
2	0	0	0	2	2	2	0	0	0	0	1	0	0	0	0	2	0
2	1	0	1	2	1	2	1	0	1	0	0	1	0	0	0	0	0
2	0	0	0	2	0	2	0	0	0	0	0	0	0	2	0	0	0

Sampling Techniques

To address class imbalance in the dataset, which consists primarily of categorical features, we techniques that were compatible with such data:

1. SMOTENC (SMOTE for Nominal and Continuous features)
   SMOTENC generates synthetic samples for the minority class by creating points between existing data points. Unlike Random Over-sampling and Under-sampling, this method adds new examples to the dataset. SMOTENC was useful in most cases as it created new data without significantly distorting the original distribution.
2. Random Over-Sampling
   This technique randomly duplicates samples of the minority class until there are an even number of each class. This method did not introduce new synthetic data and only duplicated existing points, which was not extremely helpful in training the model
3. Random Under-sampling
   This technique randomly deletes samples of the majority class until there were an even number of samples of each class. Unfortunately, it did not prove to be very useful as our minority class was extremely small in some cases, causing RUS to remove large portions of the dataset.

To visualize the impact of the techniques, these visuals are provided to compare the data before and after resampling. Added points are marked with a '+' while removed points are marked with an 'X'.

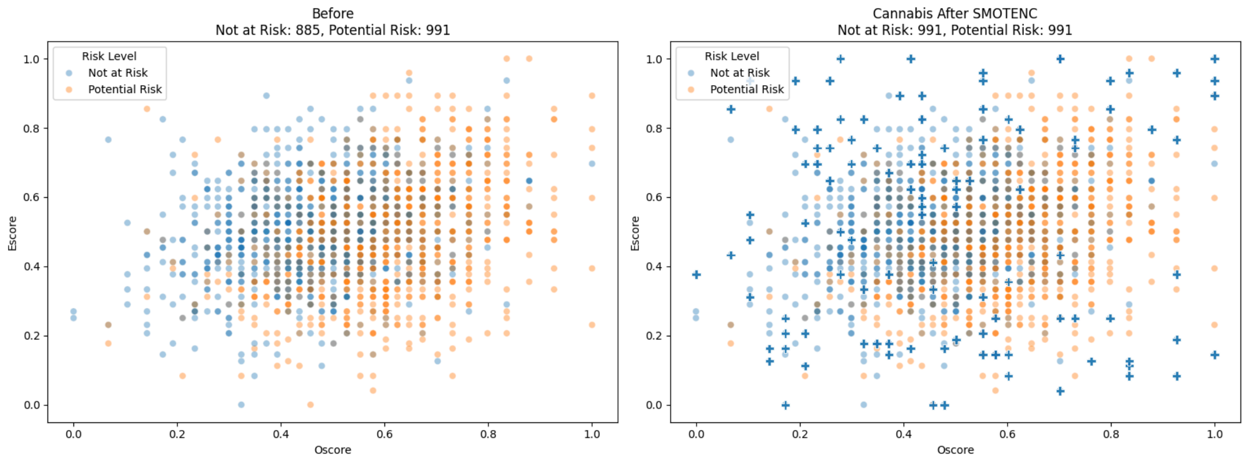


Figure 3.1 Cannabis Data Before and After SMOTENC

After resampling, the cannabis data was generally very similar to the original as the data for this drug was relatively balanced.

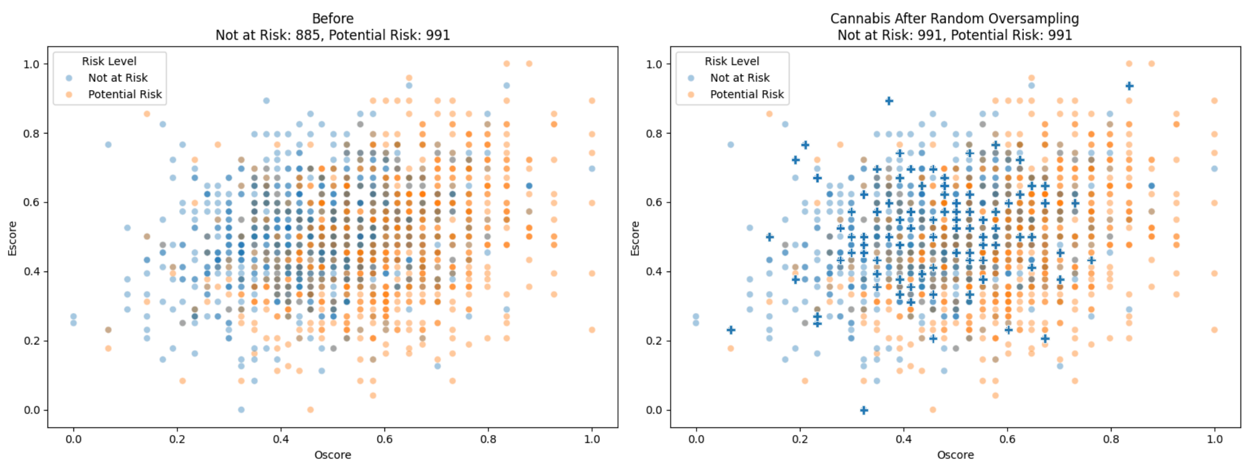


Figure 3.2 Cannabis Data Before and After Random Oversampling

Like before, the data is relatively similar to the original data, however, it is obvious that the Random Oversampling simply duplicated points, represented by the '+' markers on the already existing data whereas SMOTENC created new data, represented by the '+' markers in new areas.

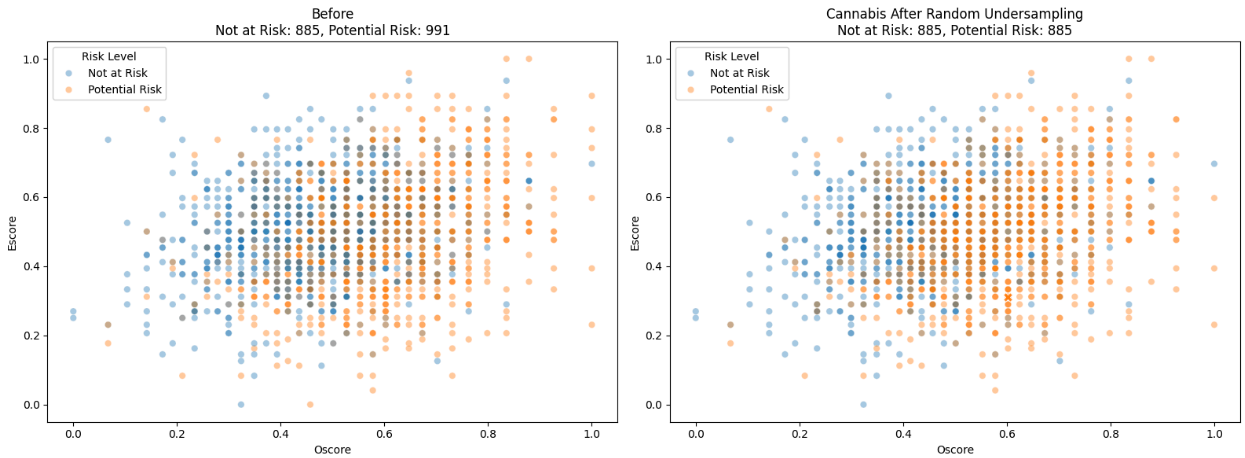


Figure 3.3 Cannabis Data Before and After Random Undersampling

Random Undersampling barely affects the cannabis data yet again as it is already relatively balanced.

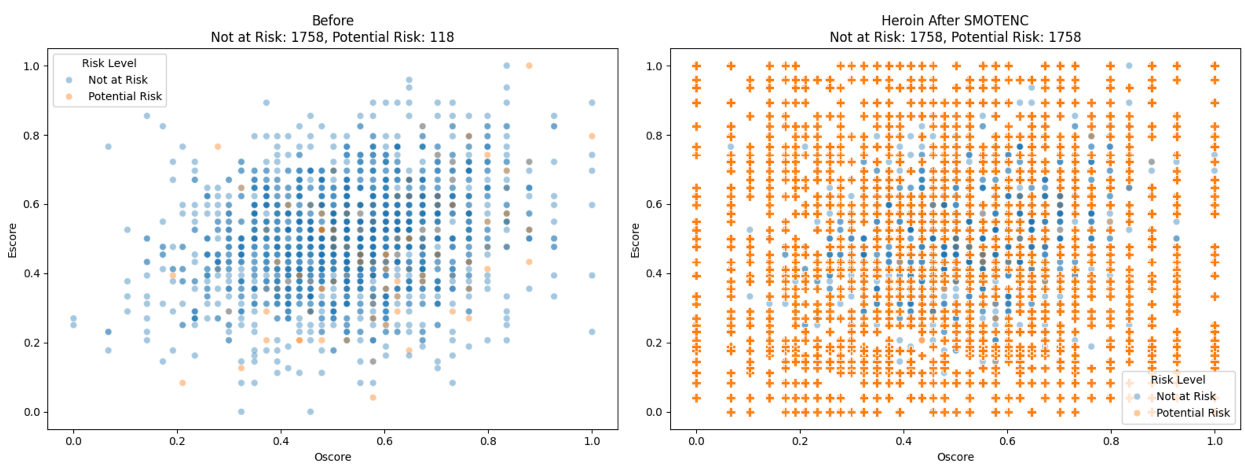


Figure 3.4 Heroin Data Before and After SMOTENC

After resampling, there is clearly many more ‘Potential Risk’ samples, however SMOTENC struggles to see a pattern in the minority class as they are sparse and scattered, resulting in a seemingly random graph after resampling.

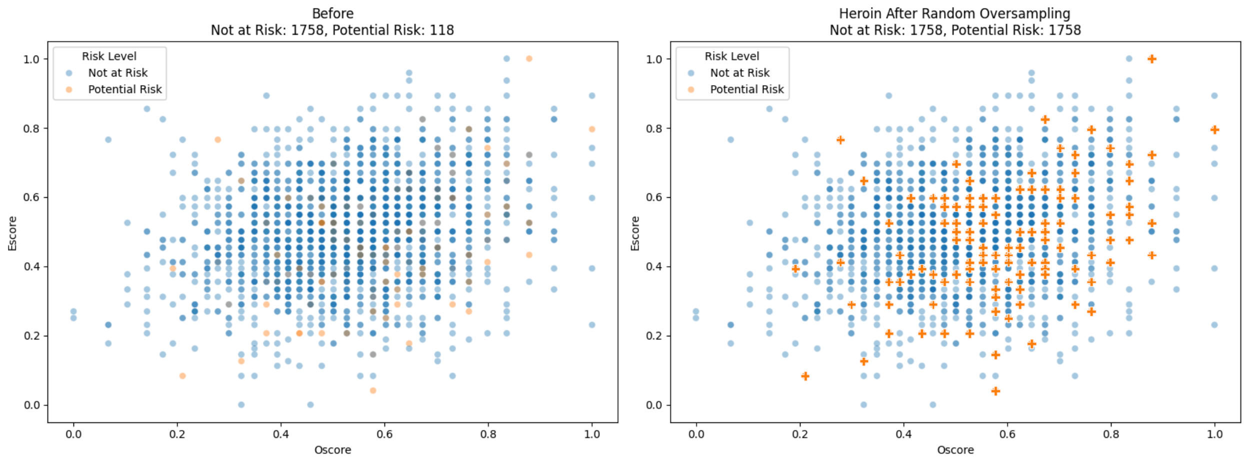


Figure 3.5 Heroin Data Before and After Random Oversampling

Like the cannabis data, Random Oversampling simply duplicates the existing minority samples. Since there is such a large imbalance, all of the minority points are duplicated multiple times as depicted in the graph.

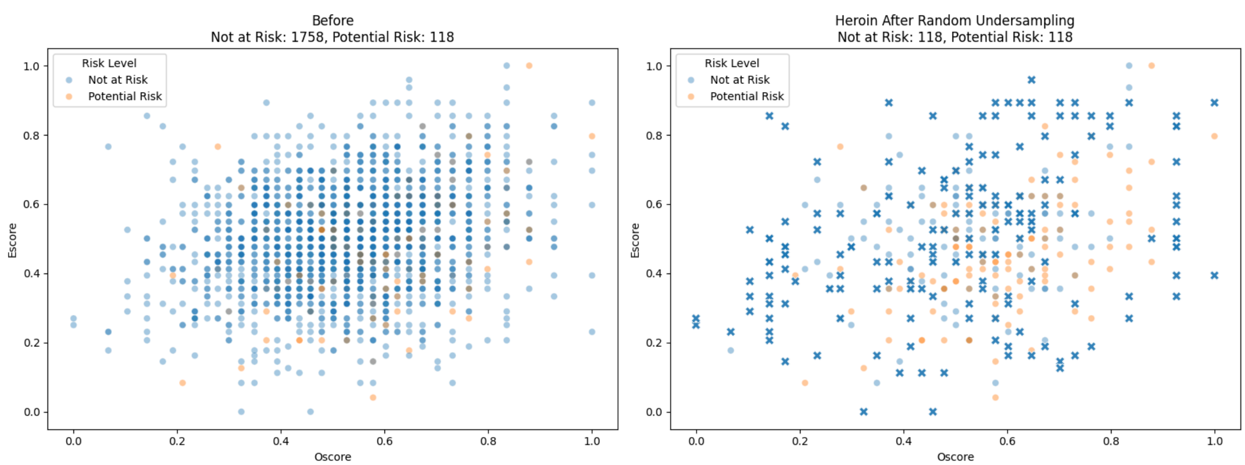


Figure 3.6 Heroin Data Before and After Random Undersampling

In Figure 3.6, the weaknesses of random undersampling are highlighted as a large imbalance results in removing a significant portion of the 'Not at Risk' samples. This data was not useful for training the models as it discards nearly all of it, limiting the models' ability to generalize.

Implementation:

To predict substance abuse risk from the data, we implemented three supervised learning models: Multilayer Perceptron, Random Forest, and Logistic Regression. Each model was evaluated on Cannabis (balanced) and Heroin (imbalanced) usage. Because the two drugs vary in class imbalance, hyperparameters may be different.

GridSearchCV from scikit-learn was used to find the best hyperparameters with 5-fold StratifiedKFold cross-validation to ensure that the minority classes were represented in all folds. Models were optimized using macro recall score since we felt that it was extremely important to identify minority samples (usually positive samples for dangerous drugs).

Model 1: Multilayer Perceptron

```
param_grid = {
    'activation': ['tanh', 'relu'],
    'alpha': [0.0001, 0.1],
    'solver': ['adam'],
    'learning_rate_init': [0.001, 0.01],
    'hidden_layer_sizes': [
        (32,), (16,),
    ]
}
```

Figure 4.1 MLP Parameter Grid for Grid Search Cross Validation

The MLP model was used to capture non-linear patterns in the data. Through GridSearchCV, the hyperparameters for Cannabis and Heroin were found:

Cannabis: {'activation': 'relu', 'alpha': 0.1, 'hidden_layer_sizes': (16,), 'learning_rate_init': 0.01, 'solver': 'adam'}

Heroin: {'activation': 'relu', 'alpha': 0.1, 'hidden_layer_sizes': (32,), 'learning_rate_init': 0.01, 'solver': 'adam'}

We used RandomOverSampler for resampling because SMOTENC often created synthetic samples that led to overfitting, especially with the Heroin data since it was so imbalanced. ROS provided more stable data by simply duplicating real samples without distortion.

Model 2: Random Forests

```python
param_grid = {
    'n_estimators': [50, 100],
    'max_depth': [2],
    'min_samples_split': [10, 20, 40],
    'min_samples_leaf': [5, 10]
}
```

Figure 4.2 Random Forests Parameter Grid for Grid Search Cross Validation

Next, we used the Random Forest model because it is good at handling imbalanced data. Like MLP, hyperparameters were tuned using GridSearchCV and stratified 5-fold cross-validation. The hyperparameters used were:

Cannabis: {'max_depth': 2, 'min_samples_leaf': 5, 'min_samples_split': 20, 'n_estimators': 100}

Heroin: {'max_depth': 2, 'min_samples_leaf': 10, 'min_samples_split': 10, 'n_estimators': 100}

For this model, instead of Random Oversampling, we used SMOTENC to create synthetic data as we learned duplicate data may cause it to overfit.

Model 3: Logistic Regression

```python
param_grid = {
    'penalty': ['l2'],
    'C': [0.01, 0.1, 1, 10],
    'solver': ['lbfgs'],
}
```

Figure 4.3 Logistic Regression Parameter Grid for Grid Search Cross Validation

Lastly, we used logistic regression as a baseline due to its simplicity. Like the other models, we used GridSearchCV and stratified 5-fold cross-validation. Final hyperparameters were selected to be:

Cannabis: {'C': 1, 'penalty': 'l2', 'solver': 'lbfgs'}

Heroin: {'C': 10, 'penalty': 'l2', 'solver': 'lbfgs'}

For resampling, we used RandomOverSampler again as SMOTENC seemed to generate unrealistic samples for the minority class. ROS provided a more simple dataset which minimized overfitting.

Experimental Setup

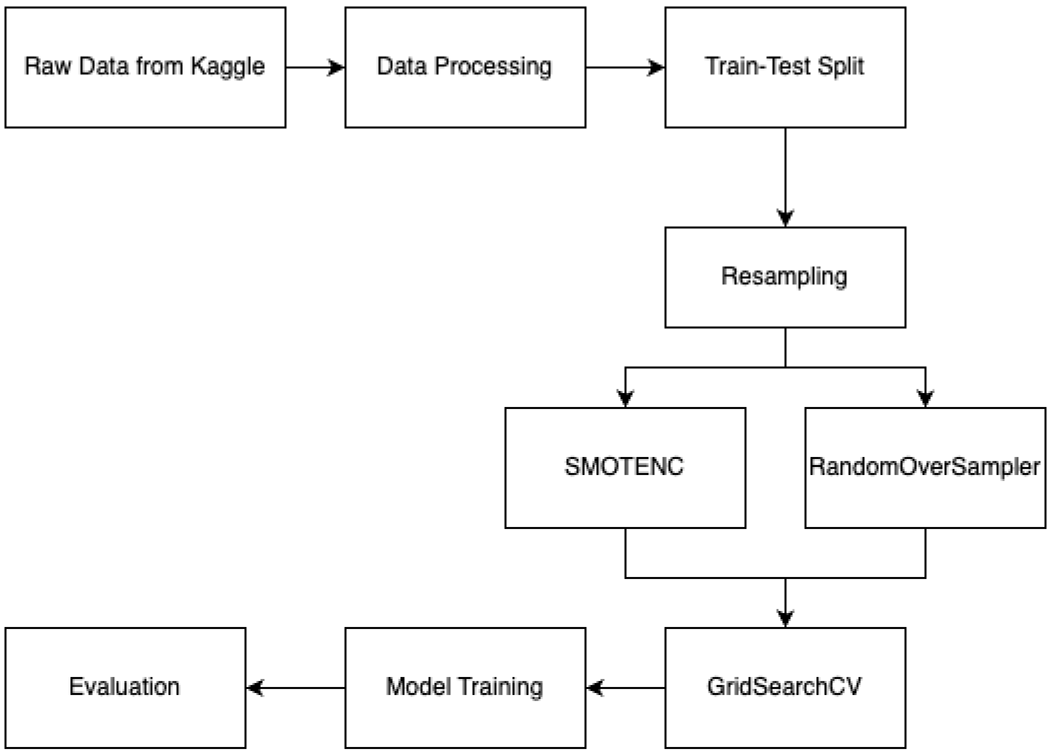


Figure 5 Architectural Design

Our project followed a typical machine learning workflow consisting of data preprocessing, resampling to address class imbalance, hyperparameter tuning, model training, and finally, evaluation. Three models (MLP, Random Forest, and Logistic Regression) were applied to predict the risk of substance abuse. For this project, we decided to focus on two target variables: Cannabis (balanced) and Heroin (imbalanced).

The dataset was split into 80% training and 20% testing using scikit-learn's train_test_split. The training set was then resampled using either SMOTENC or ROS to compare the difference in performance between unbalanced and balanced data. The test data was held until the final evaluation to examine the models' performance on unseen data.

Results

To evaluate the effectiveness of our models and resampling strategies, we compared performance of the models trained on imbalanced data to models trained on balanced data. Each model was assessed using precision, recall, and F1-score, with an emphasis on macro recall as we believed it was extremely important to identify minority cases.

Model 1: Multilayer Perceptron

Unbalanced results for Cannabis

	precision	recall	f1-score	support
0	0.78	0.78	0.78	167
1	0.83	0.83	0.83	209
accuracy			0.81	376
macro avg	0.81	0.81	0.81	376
weighted avg	0.81	0.81	0.81	376

Unbalanced Train results for Cannabis

	precision	recall	f1-score	support
0	0.80	0.81	0.80	718
1	0.82	0.81	0.82	782
accuracy			0.81	1500
macro avg	0.81	0.81	0.81	1500
weighted avg	0.81	0.81	0.81	1500

Figure 6.1 Unbalanced MLP Results for Cannabis

For the unbalanced MLP, training and test metrics were extremely similar, indicating less overfitting or underfitting.

Balanced results for Cannabis

	precision	recall	f1-score	support
0	0.79	0.78	0.78	167
1	0.82	0.83	0.83	209
accuracy			0.81	376
macro avg	0.81	0.81	0.81	376
weighted avg	0.81	0.81	0.81	376

Balanced Train results for Cannabis

	precision	recall	f1-score	support
0	0.81	0.80	0.81	782
1	0.80	0.82	0.81	782
accuracy			0.81	1564
macro avg	0.81	0.81	0.81	1564
weighted avg	0.81	0.81	0.81	1564

Figure 6.2 Balanced MLP Results for Cannabis

For the balanced results, again the train and test metrics are relatively similar indicating that there is not much overfitting or underfitting. The results from Figure 6.1 and 6.2 are also similar to each other because balancing the data did not have much effect on the already fairly balanced Cannabis data. Precision of ~0.8 for both models indicates that the model was correct about 80% of the time when predicting ‘Not at risk’. Recall scores of ~0.8 indicate that the model correctly caught 80% of cases for both ‘Not at risk’ and ‘Potential risk’

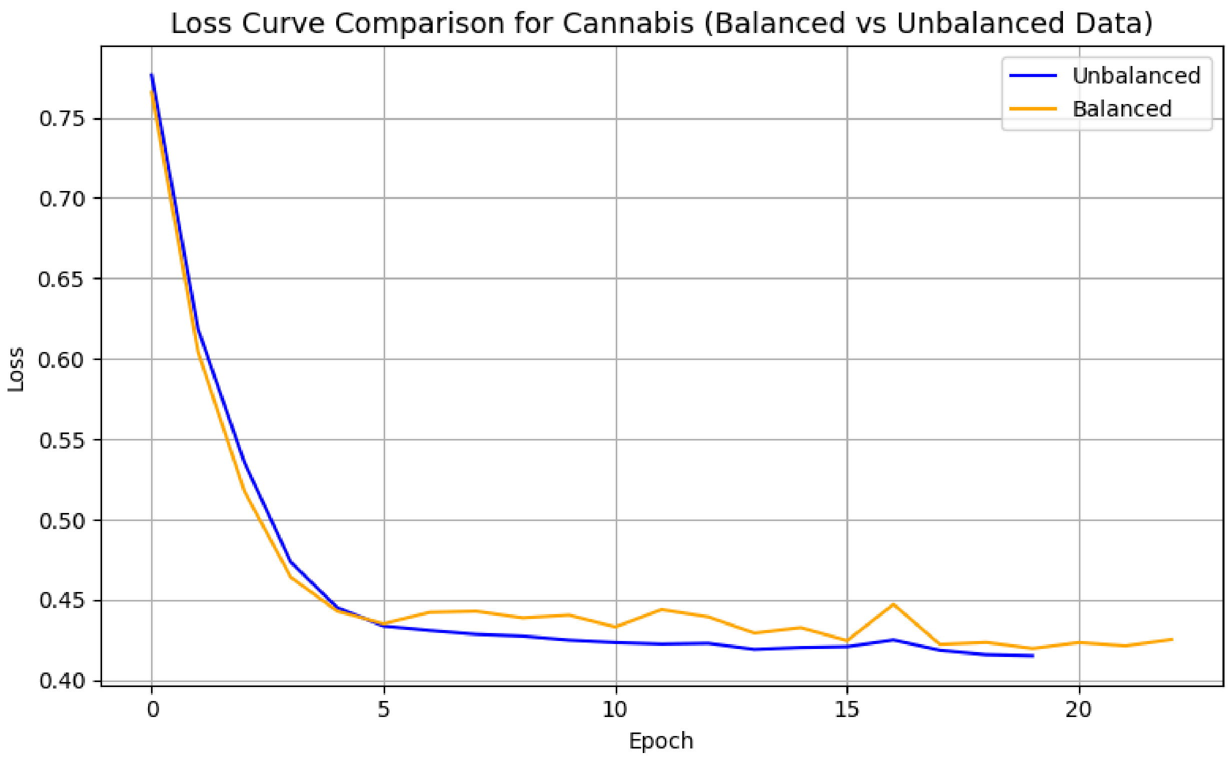


Figure 6.3 MLP Loss Curve for Cannabis

The loss curves for the two models are extremely similar since they are working on nearly the same data.

Unbalanced results for Heroin

	precision	recall	f1-score	support
0	0.92	1.00	0.96	345
1	0.00	0.00	0.00	31
accuracy			0.92	376
macro avg	0.46	0.50	0.48	376
weighted avg	0.84	0.92	0.88	376

Unbalanced Train results for Heroin

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1413
1	0.00	0.00	0.00	87
accuracy			0.94	1500
macro avg	0.47	0.50	0.49	1500
weighted avg	0.89	0.94	0.91	1500

Figure 6.4 Unbalanced MLP Results for Heroin

The training results and test results are again relatively similar indicating less overfitting. Because of the strongly imbalanced heroin data, the model focused only on the majority class and never predicted ‘Potential Risk’ (0), giving the model zeroes in precision, recall and f1-score meaning it did extremely poorly with the imbalanced data.

```
 Balanced results for Heroin
              precision    recall  f1-score   support

           0       0.97      0.77      0.86       345
           1       0.22      0.74      0.34        31

    accuracy                           0.76       376
   macro avg       0.60      0.75      0.60       376
weighted avg       0.91      0.76      0.81       376


 Balanced Train results for Heroin
              precision    recall  f1-score   support

           0       0.94      0.75      0.84      1413
           1       0.79      0.95      0.87      1413

    accuracy                           0.85      2826
   macro avg       0.87      0.85      0.85      2826
weighted avg       0.87      0.85      0.85      2826
```

Figure 6.5 Balanced MLP Results for Heroin

Compared to the results depicted in Figure 6.4, it is clear that the MLP model trained on resampled data performed extremely well. While it performed very well in training, the test results are significantly worse, indicating that there was overfitting. In the test results, the model had an extremely low precision of 0.22 for the minority class, indicating that it was frequently incorrectly predicting ‘potential risk.’ However, it was able to correctly identify 74% of ‘potential risk’ individuals, which is more important in this problem as false-negatives are much more dangerous than false-positives.

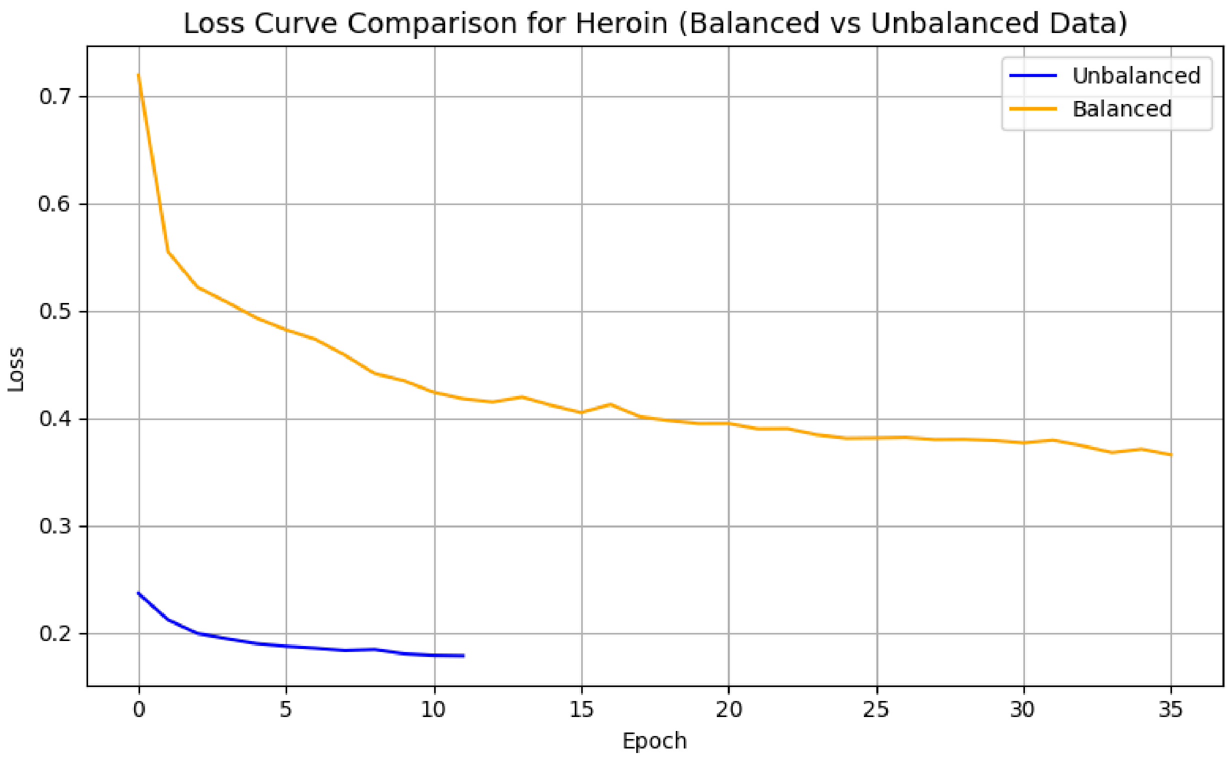
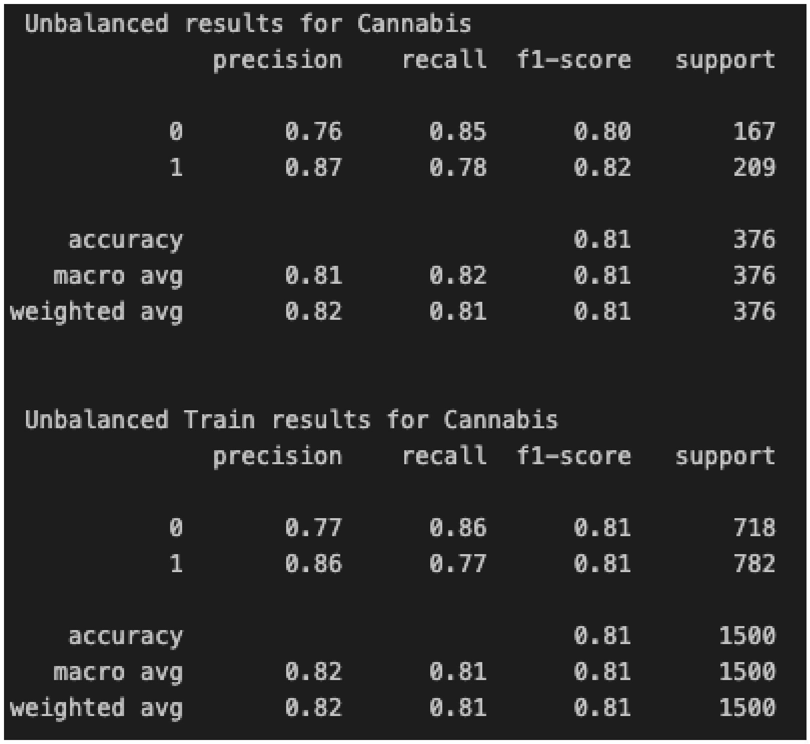


Figure 6.6 MLP Loss Curve for Heroin

In Figure 6.6, we see that the balanced data had a much higher loss but this is because the model actually predicted the minority class instead of only focusing on the majority. We also see that the model takes much longer to finish training as it is more difficult to train when the model is not only predicting one class.

Model 2: Random Forest



Unbalanced results for Cannabis

	precision	recall	f1-score	support
0	0.76	0.85	0.80	167
1	0.87	0.78	0.82	209
accuracy			0.81	376
macro avg	0.81	0.82	0.81	376
weighted avg	0.82	0.81	0.81	376

Unbalanced Train results for Cannabis

	precision	recall	f1-score	support
0	0.77	0.86	0.81	718
1	0.86	0.77	0.81	782
accuracy			0.81	1500
macro avg	0.82	0.81	0.81	1500
weighted avg	0.82	0.81	0.81	1500

Figure 7.1 Unbalanced RF Results for Cannabis

There was not much overfitting with the unbalanced random forest for Cannabis. This is likely because the Cannabis data was already extremely balanced.

Balanced results for Cannabis

	precision	recall	f1-score	support
0	0.76	0.84	0.80	167
1	0.86	0.78	0.82	209
accuracy			0.81	376
macro avg	0.81	0.81	0.81	376
weighted avg	0.82	0.81	0.81	376

Balanced Train results for Cannabis

	precision	recall	f1-score	support
0	0.79	0.85	0.82	782
1	0.84	0.78	0.81	782
accuracy			0.82	1564
macro avg	0.82	0.82	0.81	1564
weighted avg	0.82	0.82	0.81	1564

Figure 7.2 Balanced RF Results for Cannabis

Compared to Figure 7.1, the results are not very different, again because the data is nearly unchanged after resampling.

Unbalanced results for Heroin

	precision	recall	f1-score	support
0	0.98	0.71	0.82	345
1	0.20	0.81	0.32	31
accuracy			0.72	376
macro avg	0.59	0.76	0.57	376
weighted avg	0.91	0.72	0.78	376

Unbalanced Train results for Heroin

	precision	recall	f1-score	support
0	0.99	0.73	0.84	1413
1	0.16	0.84	0.27	87
accuracy			0.74	1500
macro avg	0.57	0.79	0.56	1500
weighted avg	0.94	0.74	0.81	1500

Figure 7.3 Unbalanced RF Results for Heroin

Between the training and testing results, the models had very similar results, indicating little overfitting.

```
 Balanced results for Heroin
              precision    recall  f1-score   support

           0       0.92      0.99      0.95       345
           1       0.00      0.00      0.00        31

    accuracy                           0.91       376
   macro avg       0.46      0.50      0.48       376
weighted avg       0.84      0.91      0.88       376


 Balanced Train results for Heroin
              precision    recall  f1-score   support

           0       0.90      1.00      0.95      1413
           1       1.00      0.90      0.94      1413

    accuracy                           0.95      2826
   macro avg       0.95      0.95      0.95      2826
weighted avg       0.95      0.95      0.95      2826
```

Figure 7.4 Balanced RF Results for Heroin

For the balanced Random Forest results, we see that the training data performed very well, however, the testing results were extremely poor. This model displays extreme overfitting and does not really provide any useful information. This model was extremely difficult to tune.

Model 3: Logistic Regression

Unbalanced results for Cannabis

	precision	recall	f1-score	support
0	0.80	0.83	0.82	167
1	0.86	0.83	0.85	209
accuracy			0.83	376
macro avg	0.83	0.83	0.83	376
weighted avg	0.83	0.83	0.83	376

Unbalanced Train results for Cannabis

	precision	recall	f1-score	support
0	0.79	0.84	0.81	718
1	0.84	0.79	0.82	782
accuracy			0.81	1500
macro avg	0.82	0.82	0.81	1500
weighted avg	0.82	0.81	0.81	1500

Figure 8.1 Unbalanced LR Results for Cannabis

Again, there were pretty similar results indicating little overfitting.

Balanced results for Cannabis

	precision	recall	f1-score	support
0	0.79	0.84	0.81	167
1	0.86	0.82	0.84	209
accuracy			0.83	376
macro avg	0.83	0.83	0.83	376
weighted avg	0.83	0.83	0.83	376

Balanced Train results for Cannabis

	precision	recall	f1-score	support
0	0.80	0.85	0.82	782
1	0.84	0.78	0.81	782
accuracy			0.82	1564
macro avg	0.82	0.82	0.82	1564
weighted avg	0.82	0.82	0.82	1564

Figure 8.2 Balanced LR Results for Cannabis

The unbalanced and balanced logistic regression models were extremely similar as seen in the MLP and RF models because the data is relatively unchanged.

Unbalanced results for Heroin

	precision	recall	f1-score	support
0	0.92	1.00	0.96	345
1	0.00	0.00	0.00	31
accuracy			0.92	376
macro avg	0.46	0.50	0.48	376
weighted avg	0.84	0.92	0.88	376

Unbalanced Train results for Heroin

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1413
1	0.00	0.00	0.00	87
accuracy			0.94	1500
macro avg	0.47	0.50	0.49	1500
weighted avg	0.89	0.94	0.91	1500

Figure 8.3 Unbalanced LR Results for Heroin

Although the model performed very poorly in both training and testing for the unbalanced logistic regression model, the results are very similar, indicating little overfitting.

```
Balanced results for Heroin
              precision    recall  f1-score   support

           0       0.98      0.72      0.83       345
           1       0.21      0.84      0.34        31

    accuracy                           0.73       376
   macro avg       0.60      0.78      0.58       376
weighted avg       0.92      0.73      0.79       376


Balanced Train results for Heroin
              precision    recall  f1-score   support

           0       0.83      0.73      0.78      1413
           1       0.76      0.85      0.80      1413

    accuracy                           0.79      2826
   macro avg       0.79      0.79      0.79      2826
weighted avg       0.79      0.79      0.79      2826
```

Figure 8.4 Balanced LR Results for Heroin

After resampling, the logistic regression model was able to perform much better than before. Although it had a low score of 0.21 for precision for the minority class, the recall score was extremely high, which is again what we focused on since false-negatives are much more dangerous than false-positives in this scenario.

Model Comparison:

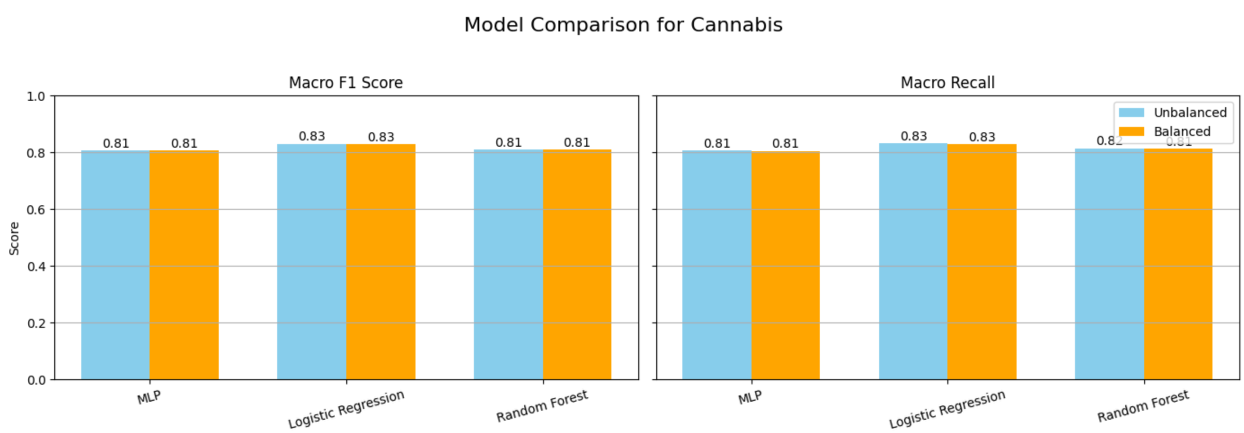


Figure 9.1 Model Comparison for Cannabis

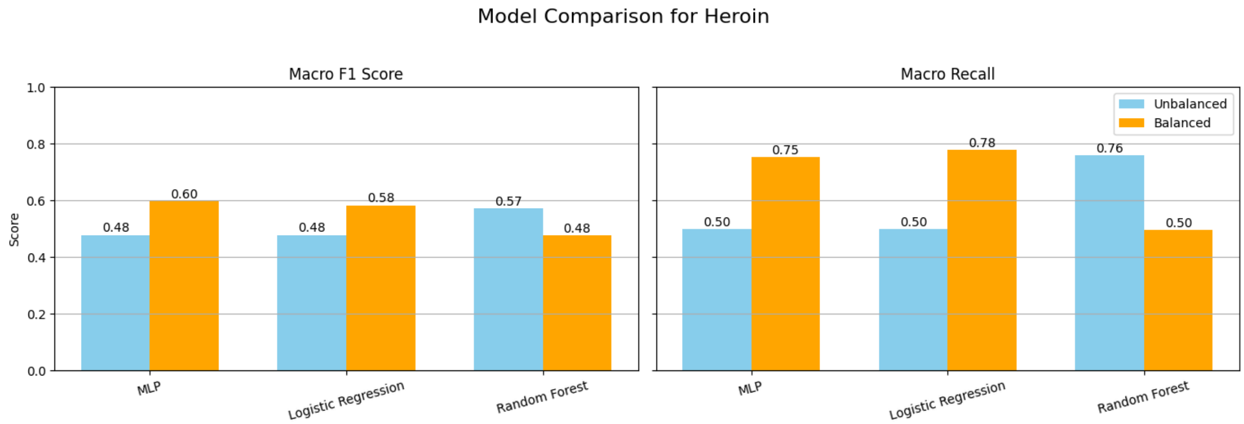


Figure 9.2 Model Comparison for Heroin

Overall, for cannabis, the models performed almost identically across the board for all three models and both imbalanced and balanced data. For heroin, most of the models performed better after resampling, likely because the data was so imbalanced and hard to train on. We see that the random forest model is an outlier, however, this is probably due to human error as we struggled significantly to tune it.

Conclusion:

Overall, the models mostly performed well and could be a great alternative to existing diagnostic techniques because of its speed and low cost.

Our model also works with other drug data, but for the sake of brevity, they are not all included in the report. In the future, the model could be trained on other data beyond personality information to be more accurate and precise.

References

"Family Addiction: How Does Addiction Affect Families?" *American Addiction Centers*, 15 Jan. 2025, americanaddictioncenters.org/rehab-guide/family-members/addiction-effects-on-family.

Fehrman, Elaine, Vincent Egan, and Evgeny Mirkes. "Drug Consumption (Quantified)." UCI Machine Learning Repository, 2015, https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified.

Jing, Yankang et al. "Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder." *Drug and alcohol dependence* vol. 206 (2020): 107605. doi:10.1016/j.drugalcdep.2019.107605

NIDA. "Drug Overdose Deaths: Facts and Figures ." *National Institute on Drug Abuse*, 21 Aug. 2024, https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates.

United States Joint Economic Committee. "The Economic Toll of the Opioid Crisis Reached Nearly $1.5 Trillion in 2020." *United States Joint Economic Committee*, 28 Sept. 2022, www.jec.senate.gov/public/index.cfm/democrats/2022/9/the-economic-toll-of-the-opioid-crisis-reached-nearly-1-5-trillion-in-2020.