Drug Consumption Analysis

Manasa Kandimalla
Carlos Antonio McNulty
Shyam Patel

Problem Statement

Which personality traits (e.g., neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsiveness, sensation) and other factors (e.g., age, gender, education) make one susceptible to the usage of various illegal drugs?

Data Sources

E. Fehrman, V. Egan and E. M. Mirkes (2016). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29]. Leicester, UK: University of Leicester, Department of Mathematics.

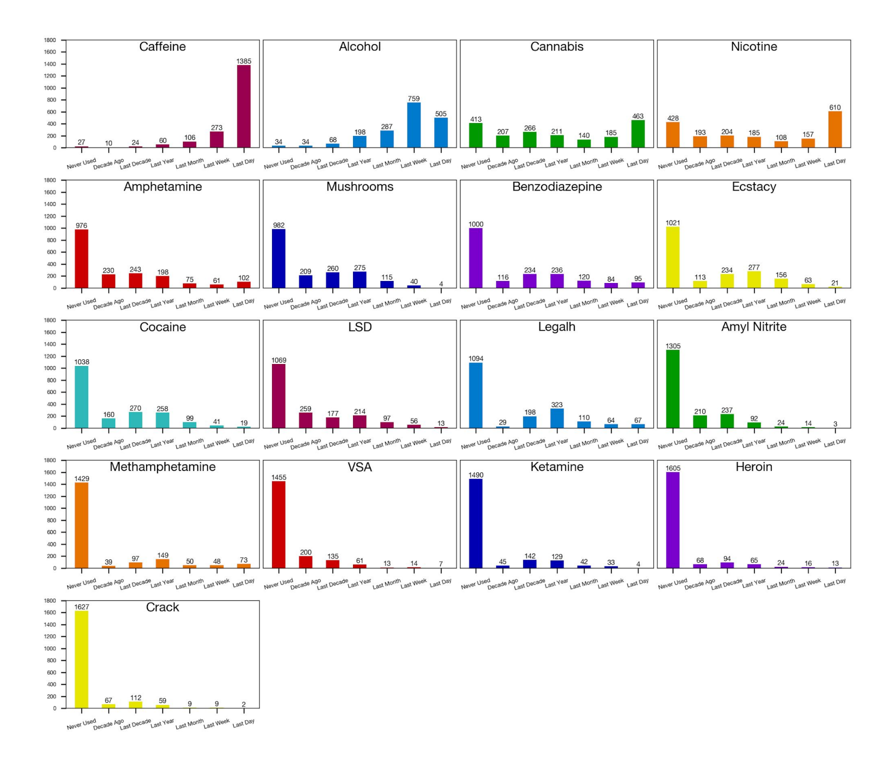
Data Sources

- 1,885 respondents
- NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness)
- BIS-11 (impulsivity)
- ImpSS (sensation seeking)
- Level of education
- Age
- Gender
- Country of residence
- Ethnicity

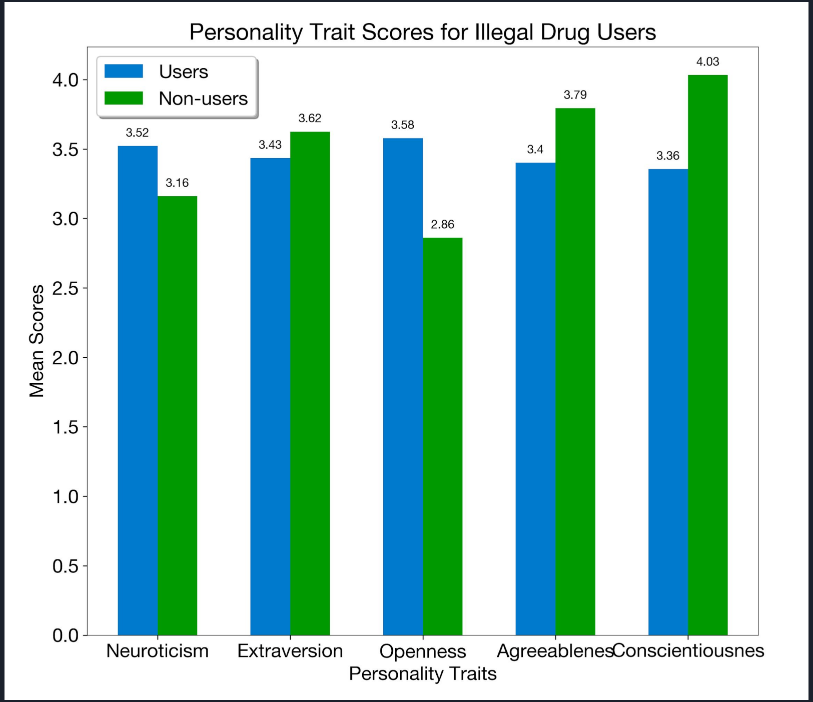
Illegal Drugs

Amphetamines
Amyl nitrite
Benzodiazepine
Cannabis
Cocaine
Crack
Ecstasy
Heroin
Ketamine
LSD
Methadone
Mushrooms

Drug Usage

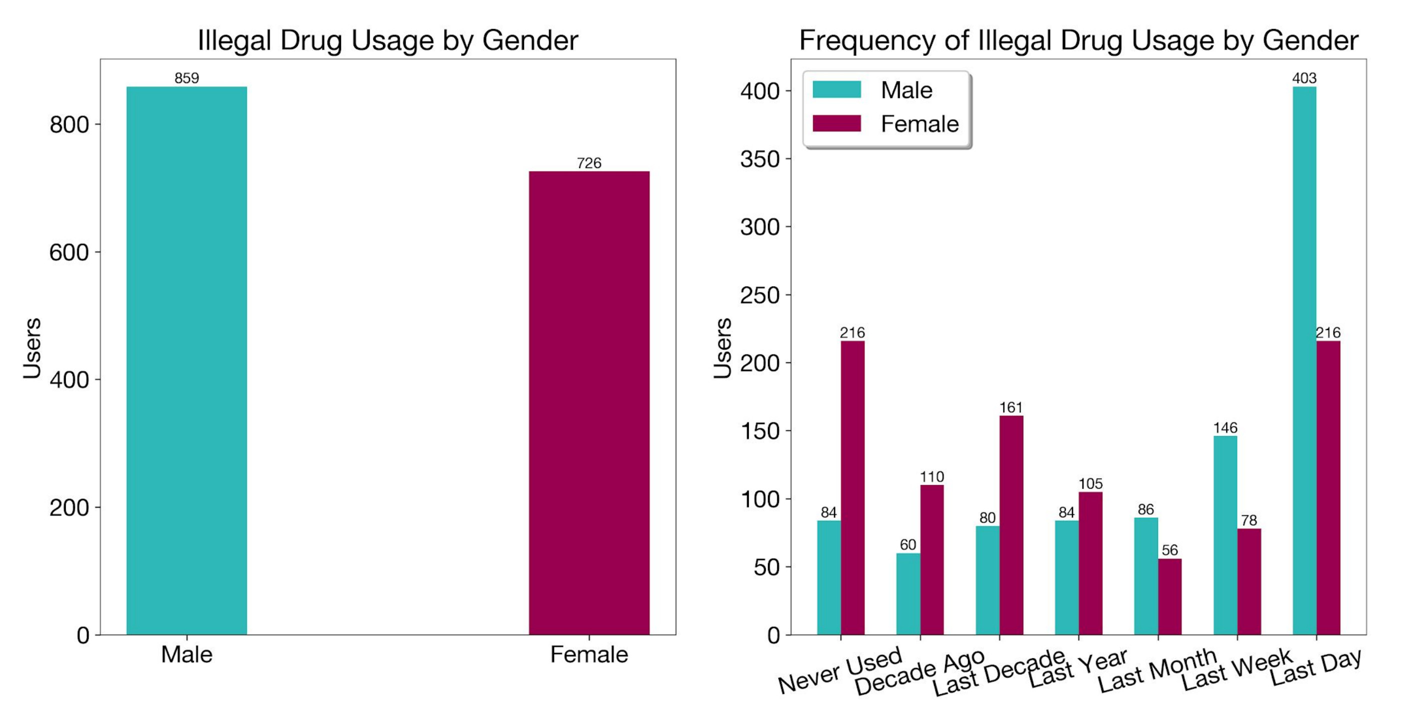

1800
1600
1400
1200
1000
800
600
400
200
0
Never Used
Decade Ago
Last Decade
Last Year
Last Month
Last Week
Last Day
Caffeine
27 10 24 60 106 273 1385
Alcohol
34 34 68 198 287 759 505
Cannabis
413 207 266 211 140 185 463
Nicotine
428 193 204 185 108 157 610
Amphetamine
976 230 243 198 75 61 102
Mushrooms
982 209 260 275 115 40 4
Benzodiazepine
1000 116 234 236 120 84 95
Ecstacy
1021 113 234 277 156 63 21
Cocaine
1038 160 270 258 99 41 19
LSD
1069 259 177 214 97 56 13
Legalh
1094 29 198 323 110 64 67
Amyl Nitrite
1305 210 237 92 24 14 3
Methamphetamine
1429 39 97 149 50 48 73
VSA
1455 200 135 61 13 14 7
Ketamine
1490 45 142 129 42 33 4
Heroin
1605 68 94 65 24 16 13
Crack
1627 67 112 59 9 9 2

Illegal Drug Usage by Personality Trait

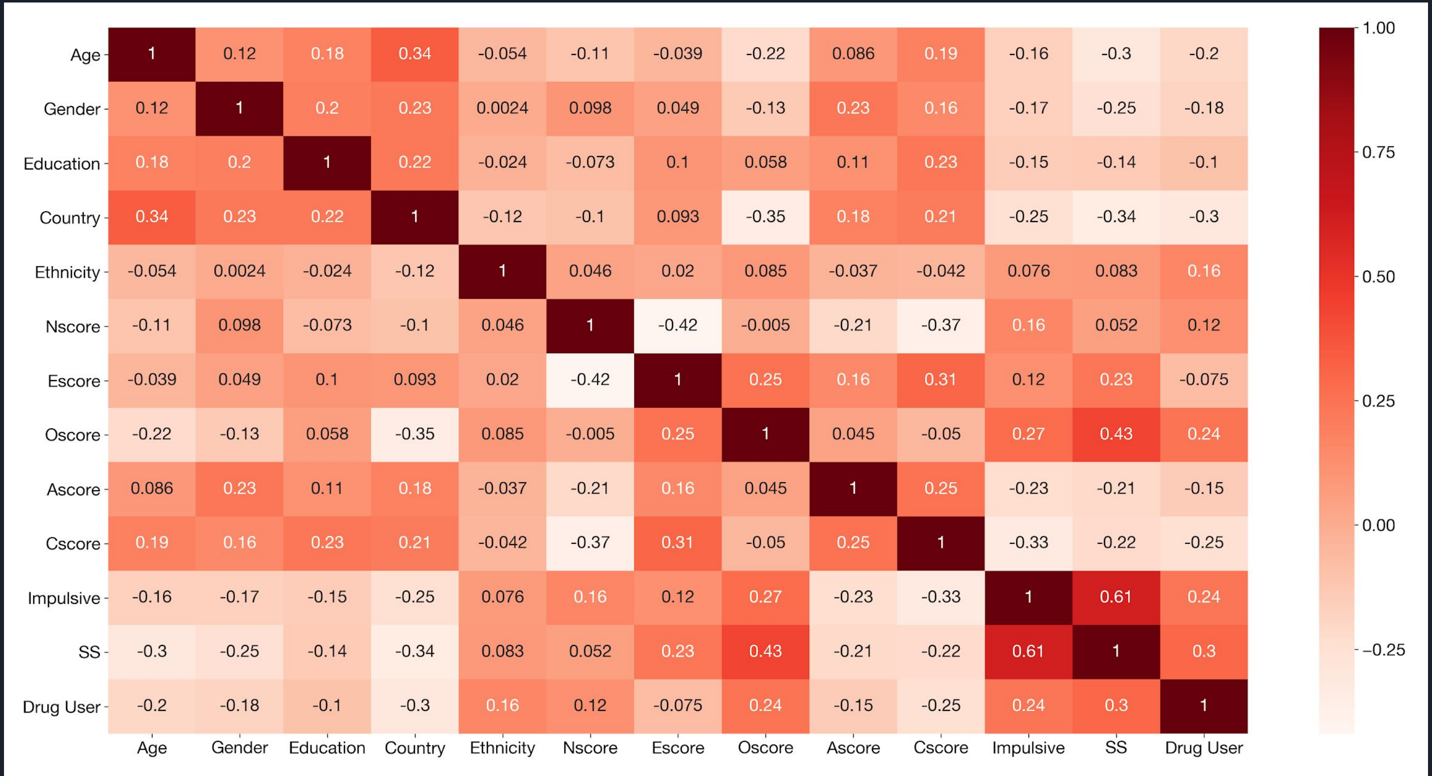


	t-statistic	p-value
Nscore	6.3434	2.6924e-10
Escore	-3.4179	0.0003
Oscore	13.2261	1.8035e-34
Ascore	-6.3017	1.8286e-10
Cscore	-12.7530	1.2412e-32

Illegal Drug Usage and Frequency by Gender



Heatmap of Correlation Matrix



	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Drug User
Age	1	0.12	0.18	0.34	-0.054	-0.11	-0.039	-0.22	0.086	0.19	-0.16	-0.3	-0.2
Gender	0.12	1	0.2	0.23	0.0024	0.098	0.049	-0.13	0.23	0.16	-0.17	-0.25	-0.18
Education	0.18	0.2	1	0.22	-0.024	-0.073	0.1	0.058	0.11	0.23	-0.15	-0.14	-0.1
Country	0.34	0.23	0.22	1	-0.12	-0.1	0.093	-0.35	0.18	0.21	-0.25	-0.34	-0.3
Ethnicity	-0.054	0.0024	-0.024	-0.12	1	0.046	0.02	0.085	-0.037	-0.042	0.076	0.083	0.16
Nscore	-0.11	0.098	-0.073	-0.1	0.046	1	-0.42	-0.005	-0.21	-0.37	0.16	0.052	0.12
Escore	-0.039	0.049	0.1	0.093	0.02	-0.42	1	0.25	0.16	0.31	0.12	0.23	-0.075
Oscore	-0.22	-0.13	0.058	-0.35	0.085	-0.005	0.25	1	0.045	-0.05	0.27	0.43	0.24
Ascore	0.086	0.23	0.11	0.18	-0.037	-0.21	0.16	0.045	1	0.25	-0.23	-0.21	-0.15
Cscore	0.19	0.16	0.23	0.21	-0.042	-0.37	0.31	-0.05	0.25	1	-0.33	-0.22	-0.25
Impulsive	-0.16	-0.17	-0.15	-0.25	0.076	0.16	0.12	0.27	-0.23	-0.33	1	0.61	0.24
SS	-0.3	-0.25	-0.14	-0.34	0.083	0.052	0.23	0.43	-0.21	-0.22	0.61	1	0.3
Drug User	-0.2	-0.18	-0.1	-0.3	0.16	0.12	-0.075	0.24	-0.15	-0.25	0.24	0.3	1

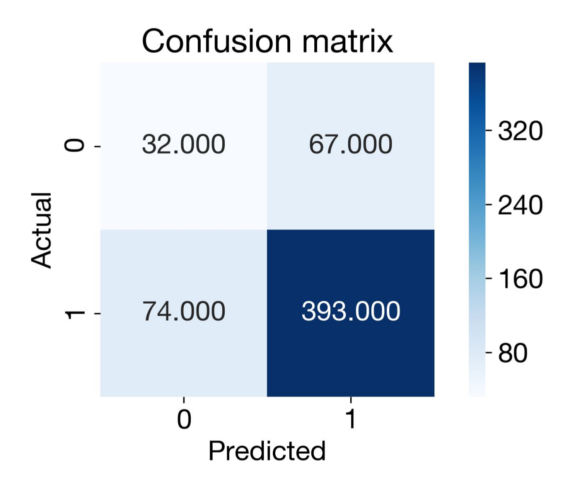
Data Modeling:
Decision Tree Classifier

- Used Pearson correlation heatmap to select variables
  - *Age*, *Country*, *Oscore*, *Cscore*, *Impulsive* and *SS* were found to be highly correlated with the output variable *Drug User*
- 475 nodes, may be evidence of overfitting
- 10–fold cross validation → 79.0% mean **accuracy**
  87.4% mean **F1 score**

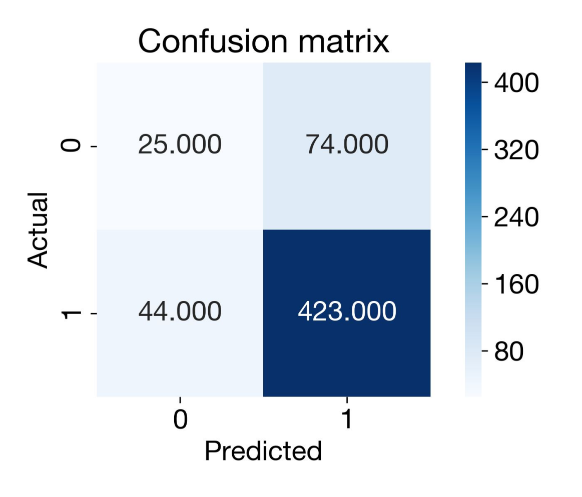
Data Modeling:
k-Nearest Neighbors Classifier

- Used same 6 independent variables
  (*Age*, *Country*, *Oscore*, *Cscore*, *Impulsive* and *SS*)
- 10–fold cross validation → 80.9% mean **accuracy**
  88.8% mean **F1 score**

Data Modeling:
Decision Tree vs k-Nearest Neighbors



↑ **Precision** of 85.4%
↓ **Recall** of 84.2%



↓ **Precision** of 85.1%
↑ **Recall** of 90.6%

Data Modeling:
Random Forest Classifier

- Used Grid Search
  - 30,240 base estimators
  - evaluated using 10-fold cross validation
  - 100 iterations
- 10–fold cross validation
  - 91.3% mean **accuracy**
  - 84.4% mean **F1 score**
- 82.7% **accuracy** on the test set
- 90.2% **F1 score** on the set

Data Modeling

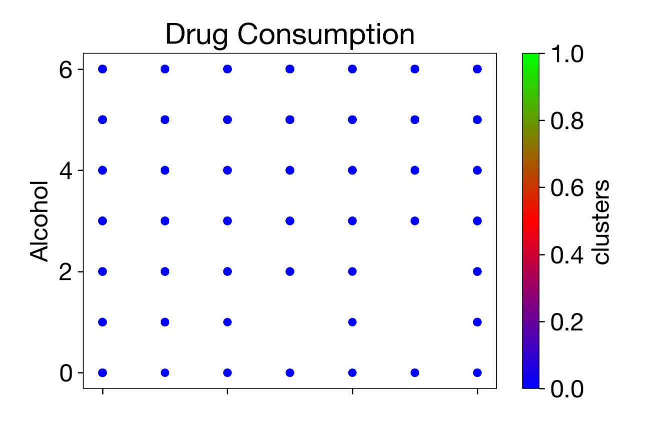
Clustering

- Silhouette Coefficient as the Evaluation Metric
- Clusters with 8 subsets of features
- Hierarchical Clustering with Single Linkage, Hierarchical Clustering with Complete Linkage, K-Means Clustering, DBSCAN applied to each cluster.
- Hierarchical Clustering performed the best and DBSCAN performed the worst.

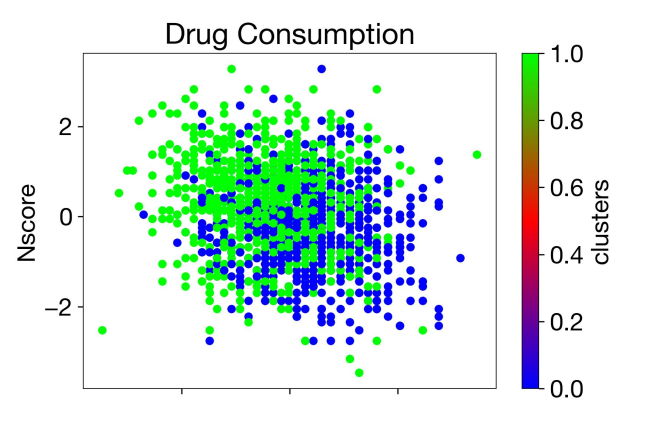
Data Modeling

Clustering

Cluster with only Non-Illegal Drugs(alcohol VS Nicotine)



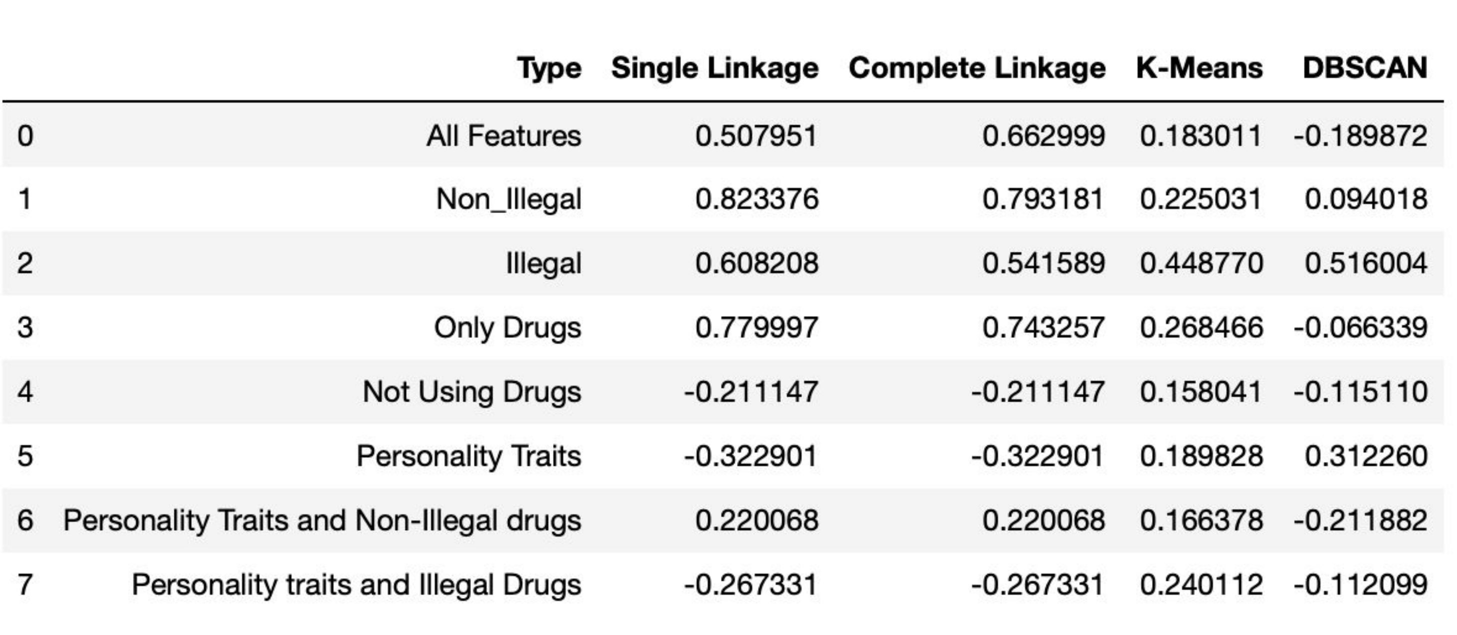
Cluster with only Personality Traits(Ascore VS Nscore)



Data Modeling

Clustering

Silhouette Coefficients



	Type	Single Linkage	Complete Linkage	K-Means	DBSCAN
0	All Features	0.507951	0.662999	0.183011	-0.189872
1	Non_Illegal	0.823376	0.793181	0.225031	0.094018
2	Illegal	0.608208	0.541589	0.448770	0.516004
3	Only Drugs	0.779997	0.743257	0.268466	-0.066339
4	Not Using Drugs	-0.211147	-0.211147	0.158041	-0.115110
5	Personality Traits	-0.322901	-0.322901	0.189828	0.312260
6	Personality Traits and Non-Illegal drugs	0.220068	0.220068	0.166378	-0.211882
7	Personality traits and Illegal Drugs	-0.267331	-0.267331	0.240112	-0.112099

Conclusions

- We observed that the clusters that are made up of Only Illegal, Only Non-Illegal and Only Drugs have the highest Silhouette Coefficient which means they are good clusters compared to the others.
- The best performing binary classifier was the random forest classifier with a mean accuracy of 91.2% when tested with 10-fold cross validation. On the test set, it had an accuracy of 82.7% and an F1 score of 90.2%.
- We are satisfied with our model's ability to predict drug usage by using personality traits and other characteristics.