ISSN 2348-3156 (Print)
International Journal of Social Science and Humanities Research ISSN 2348-3164 (online)
Vol. 12, Issue 1, pp: (23-35), Month: January - March 2024, Available at: www.researchpublish.com

A Comparative Study of Machine Learning Algorithms for the Prediction of Drug-Abuse - A Classification Approach

[1]Ali Ahmad Siddiqui, [2]Dr. Chao Liu

Contra Costa College

DOI: https://doi.org/10.5281/zenodo.10460926

Published Date: 05-January-2024

Abstract: **This research was conducted as part of ENGIN-298 at Contra Costa College by Ali Ahmad Siddiqui with the assistance of Dr. Chao Liu from Fall 2022 to Fall 2023. Drug-use is a societal crisis that is in-need of desperate mitigation and alleviation. In this paper, we claim that machine learning algorithms are highly effective in assessing the likelihood for an individual to engage in the use of certain drugs. In particular, we show how machine learning algorithms can be trained using real- world data collected from users who consume a variety of different drugs in order to give insight into the drug-use for unknown individuals.**

Keywords: **certain drugs, societal crisis, machine learning algorithms.**

1. INTRODUCTION

Substance abuse in the United States remains an unsolved epidemic. In 2022, there has been a 29% increase in substance abuse (compared to 2019), with over 99,000+ Americans killed due to substance overdose in 2020 alone. Drug overdose deaths in the US since 2000 are nearing one million. And even though the federal budget for drug control nears $35 billion dollars, there are many, unexplored approaches to mitigate the current crisis [4]. Recently, in the realm of programming, machine learning has become a powerful computational tool that serves to identify patterns in large data sets. While research has been conducted on machine learning in the context of drug-abuse, little has been done on using specific and quantifiable personality factors as a means of training those assessing the risk of drug-abuse.

2. GOALS

Machine Learning has the potential to improve drug-abuse diagnosis by providing a non-invasive and efficient method in the process of assessing an individual's risk prediction. Our research aims to identify accurate models and improve the accuracy of these models to identify high-risk individuals early on. In the process, our research contributes to early interventions and reducing the monetary and physical burden of the criminal justice and healthcare systems. However, it is imperative to note that this research is not a substitute for diagnosis and the final conclusions should always be made by industry professionals, including therapists, counselors, and psychiatrists.

As a result, the purpose of this study is to examine the effectiveness of machine learning algorithms to assess the likelihood of specific drug-uses for an individual using personality risk-factors. The importance of this is to identify those at-risk at an earlier state to ensure they seek the proper medical and counseling resources to ensure such interests don't turn into dangerous addictions. The objective of this independent study is to identify machine learning programs in the Python language that can accurately predict the likelihood for abuse of specific drugs and substances for an individual, as a means of alleviating the suffering that society faces due to drug addiction and overdoses.

3. LITERATURE REVIEW

In 2020, over 40 million people in America had at least one substance use disorder. Substance use disorder is used to describe the recurring pattern of using a substance that may cause problems, distress, or even death [3].

Substance abuse not only affects individual drug users, but also their families, friends, and society at large, and are rapidly consuming limited public funds. The economic impact of drug abuse is significant to the economy at large because of lost productivity, burglary, violence, assault, and theft. In 2002, the economic cost of drug abuse to the United States was $180.9 billion. Children of individuals who abuse drugs are often abused or neglected [5]. According to the Federal Bureau of Prisons, around 46% of inmates were incarcerated for drug offenses or drug-related crimes in 2021 [2].

Artificial intelligence systems, with the help of Deep Learning and Machine Learning, can serve as tools to help the current suffering across the nation by finding a way to identify substance abuse in its early phases to ensure it does not exacerbate. Machine learning is a branch of computer science that uses data to overcome challenges that seem impossible through traditional methods.

This is done by allowing computers to learn without directly programming that learning. For example, machine learning has been used to expand drug development prior to critical shortages, especially in the case of infection surges during globalized pandemics [1]. Using Machine Learning to understand risk factors associated with drug-abuse can better enhance counseling and medical intervention services by providing a tool to identify individuals on a risky-road early on.

4. DATASET INFORMATION

For the project, we used a free, publicly available machine learning repository from UC Irvine [7]. The data available in the repository used an online survey methodology to collect data, which includes the Big Five personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information. The purpose of this specific dataset is to diagnostically predict the level of specific drug-use of an individual , based on certain personality and impulsivity factors and measurements available in the dataset for each individual. The database 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, benzodi- azepine, 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.

1. ID is the number of records in the original database. Cannot be related to participants. It can be used for reference only.
2. Age (Real) is age of participant and has one of the values:

Value	Meaning	Cases	Fraction
-0.95197	18-24	643	34.11%
-0.07854	25-34	481	25.52%
0.49788	35-44	356	18.89%
1.09449	45-54	294	15.60%
1.82213	55-64	93	4.93%
2.59171	65+	18	0.95%

Min	Max	Mean	Std.dev.
-0.95197	2.59171	0.03461	0.87813

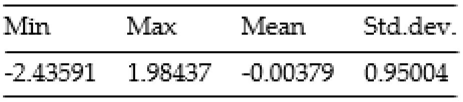
3. Gender (Real) is gender of participant:

Value	Meaning	Cases	Fraction
0.48246	Female	942	49.97%
-0.48246	Male	943	50.03%

Min	Max	Mean	Std.dev.
-0.48246	0.48246	-0.00026	0.48246

4. Education (Real) is level of education of participant and has one of the values:

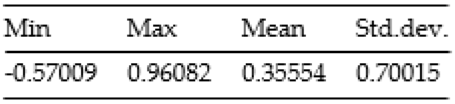
Value	Meaning	Cases	Fraction
-2.43591	Left school before 16 years	28	1.49%
-1.73790	Left school at 16 years	99	5.25%
-1.43719	Left school at 17 years	30	1.59%
-1.22751	Left school at 18 years	100	5.31%
-0.61113	Some college or university, no certificate or degree	506	26.84%
-0.05921	Professional certificate/ diploma	270	14.32%
0.45468	University degree	480	25.46%
1.16365	Masters degree	283	15.01%
1.98437	Doctorate degree	89	4.72%



Min	Max	Mean	Std.dev.
-2.43591	1.98437	-0.00379	0.95004

5. Country (Real) is country of current residence of participant and has one of the values:

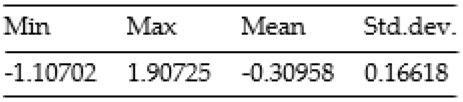
Value	Meaning	Cases	Fraction
-0.09765	Australia	54	2.86%
0.24923	Canada	87	4.62%
-0.46841	New Zealand	5	0.27%
-0.28519	Other	118	6.26%
0.21128	Republic of Ireland	20	1.06%
0.96082	UK	1044	55.38%
-0.57009	USA	557	29.55%



Min	Max	Mean	Std.dev.
-0.57009	0.96082	0.35554	0.70015

6. Ethnicity (Real) is ethnicity of participant and has one of the values:

Value	Meaning	Cases	Fraction
-0.50212	Asian	26	1.38%
-1.10702	Black	33	1.75%
1.90725	Mixed-Black/Asian	3	0.16%
0.12600	Mixed-White/Asian	20	1.06%
-0.22166	Mixed-White/Black	20	1.06%
0.11440	Other	63	3.34%
-0.31685	White	1720	91.25%



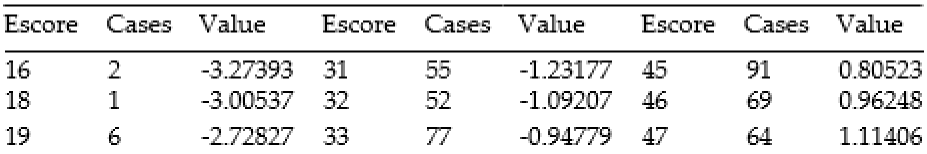
Min	Max	Mean	Std.dev.
-1.10702	1.90725	-0.30958	0.16618

7. Nscore (Real) is NEO-FFI-R Neuroticism. Possible values are presented in table below:

Nscore	Cases	Value	Nscore	Cases	Value	Nscore	Cases	Value
12	1	-3.46436	29	60	-0.67825	46	67	1.02119
13	1	-3.15735	30	61	-0.58016	47	27	1.13281
14	7	-2.75696	31	87	-0.46725	48	49	1.23461
15	4	-2.52197	32	78	-0.34799	49	40	1.37297
16	3	-2.42317	33	68	-0.24649	50	24	1.49158
17	4	-2.34360	34	76	-0.14882	51	27	1.60383
18	10	-2.21844	35	69	-0.05188	52	17	1.72012
19	16	-2.05048	36	73	0.04257	53	20	1.83990
20	24	-1.86962	37	67	0.13606	54	15	1.98437
21	31	-1.69163	38	63	0.22393	55	11	2.12700
22	26	-1.55078	39	66	0.31287	56	10	2.28554
23	29	-1.43907	40	80	0.41667	57	6	2.46262
24	35	-1.32828	41	61	0.52135	58	3	2.61139
25	56	-1.19430	42	77	0.62967	59	5	2.82196
26	57	-1.05308	43	49	0.73545	60	2	3.27393
27	65	-0.92104	44	51	0.82562			
28	70	-0.79151	45	37	0.91093			

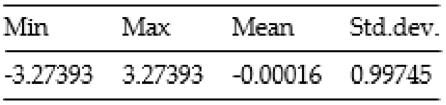
Min	Max	Mean	Std.dev.
-3.46436	3.27393	0.00004	0.99808

8. Escore (Real) is NEO-FFI-R Extraversion. Possible values are presented in table below:



Escore	Cases	Value	Escore	Cases	Value	Escore	Cases	Value
16	2	-3.27393	31	55	-1.23177	45	91	0.80523
18	1	-3.00537	32	52	-1.09207	46	69	0.96248
19	6	-2.72827	33	77	-0.94779	47	64	1.11406

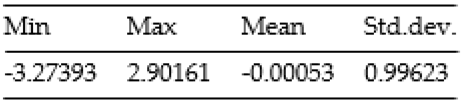
Escore	Cases	Value	Escore	Cases	Value	Escore	Cases	Value
20	3	-2.53830	34	68	-0.80615	48	62	1.28610
21	3	-2.44904	35	58	-0.69509	49	37	1.45421
22	8	-2.32338	36	89	-0.57545	50	25	1.58487
23	5	-2.21069	37	90	-0.43999	51	34	1.74091
24	9	-2.11437	38	106	-0.30033	52	21	1.93886
25	4	-2.03972	39	107	-0.15487	53	15	2.12700
26	21	-1.92173	40	130	0.00332	54	10	2.32338
27	23	-1.76250	41	116	0.16767	55	9	2.57309
28	23	-1.63340	42	109	0.32197	56	2	2.85950
29	32	-1.50796	43	105	0.47617	58	1	3.00537
30	38	-1.37639	44	103	0.63779	59	2	3.27393



Min	Max	Mean	Std.dev.
-3.27393	3.27393	-0.00016	0.99745

9. Oscore (Real) is NEO-FFI-R Openness to experience. Possible values are presented in table below:

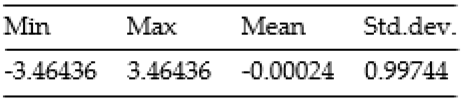
Oscore	Cases	Value	Oscore	Cases	Value	Oscore	Cases	Value
24	2	-3.27393	38	64	-1.11902	50	83	0.58331
26	4	-2.85950	39	60	-0.97631	51	87	0.72330
28	4	-2.63199	40	68	-0.84732	52	87	0.88309
29	11	-2.39883	41	76	-0.71727	53	81	1.06238
30	9	-2.21069	42	87	-0.58331	54	57	1.24033
31	9	-2.09015	43	86	-0.45174	55	63	1.43533
32	13	-1.97495	44	101	-0.31776	56	38	1.65653
33	23	-1.82919	45	103	-0.17779	57	34	1.88511
34	25	-1.68062	46	134	-0.01928	58	19	2.15324
35	26	-1.55521	47	107	0.14143	59	13	2.44904
36	39	-1.42424	48	116	0.29338	60	7	2.90161
37	51	-1.27553	49	98	0.44585			



Min	Max	Mean	Std.dev.
-3.27393	2.90161	-0.00053	0.99623

10. Ascore (Real) is NEO-FFI-R Agreeableness. Possible values are presented in table below:

Ascore	Cases	Value	Ascore	Cases	Value	Ascore	Cases	Value
12	1	-3.46436	34	42	-1.34289	48	104	0.76096
16	1	-3.15735	35	45	-1.21213	49	85	0.94156
18	1	-3.00537	36	62	-1.07533	50	68	1.11406
23	1	-2.90161	37	83	-0.91699	51	58	1.2861
Ascore	Cases	Value	Ascore	Cases	Value	Ascore	Cases	Value
24	2	-2.78793	38	82	-0.76096	52	39	1.45039
25	1	-2.70172	39	102	-0.60633	53	36	1.61108
26	7	-2.5383	40	98	-0.45321	54	36	1.81866
27	7	-2.35413	41	114	-0.30172	55	16	2.03972
28	8	-2.21844	42	101	-0.15487	56	14	2.23427
29	13	-2.07848	43	105	-0.01729	57	8	2.46262
30	18	-1.92595	44	118	0.13136	58	7	2.75696
31	24	-1.772	45	112	0.28783	59	1	3.15735
32	30	-1.6209	46	100	0.43852	60	1	3.46436
33	34	-1.47955	47	100	0.59042			

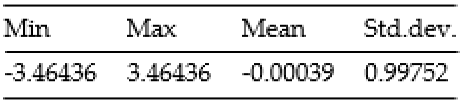


Min	Max	Mean	Std.dev.
-3.46436	3.46436	-0.00024	0.99744

11. Cscore (Real) is NEO-FFI-R Conscientiousness. Possible values are presented in table below:

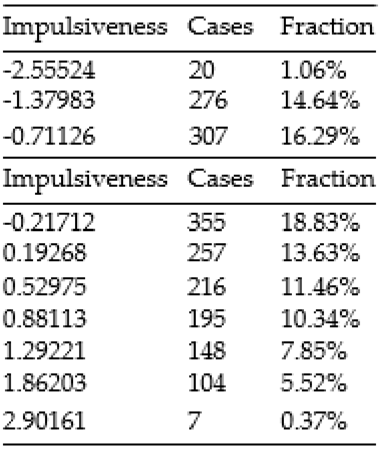
Cscore	Cases	Value	Cscore	Cases	Value	Cscore	Cases	Value
17	1	-3.46436	32	39	-1.25773	46	113	0.58489
19	1	-3.15735	33	49	-1.13788	47	95	0.7583
20	3	-2.90161	34	55	-1.0145	48	95	0.93949
21	2	-2.72827	35	55	-0.89891	49	76	1.13407
22	5	-2.57309	36	69	-0.78155	50	47	1.30612
23	5	-2.42317	37	81	-0.65253	51	43	1.46191
24	6	-2.30408	38	77	-0.52745	52	34	1.63088

25	9	-2.18109	39	87	-0.40581	53	28	1.81175
26	13	-2.04506	40	97	-0.27607	54	27	2.04506
27	13	-1.92173	41	99	-0.14277	55	13	2.33337
28	25	-1.78169	42	105	-0.00665	56	8	2.63199
29	24	-1.64101	43	90	0.12331	57	3	3.00537
30	29	-1.5184	44	111	0.25953	59	1	3.46436
31	41	-1.38502	45	111	0.41594			



Min	Max	Mean	Std.dev.
-3.46436	3.46436	-0.00039	0.99752

12. Impulsive (Real) is impulsiveness measured by BIS-11. Possible values are presented in table below:

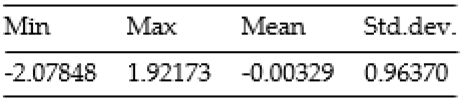


Impulsiveness	Cases	Fraction
-2.55524	20	1.06%
-1.37983	276	14.64%
-0.71126	307	16.29%
Impulsiveness	Cases	Fraction
-0.21712	355	18.83%
0.19268	257	13.63%
0.52975	216	11.46%
0.88113	195	10.34%
1.29221	148	7.85%
1.86203	104	5.52%
2.90161	7	0.37%

Min	Max	Mean	Std.dev.
-2.55524	2.90161	0.00721	0.95446

13. SS (Real) is sensation seeing measured by ImpSS. Possible values are presented in table below:

SS	Cases	Fraction
-2.07848	71	3.77%
-1.54858	87	4.62%
-1.18084	132	7.00%
-0.84637	169	8.97%
-0.52593	211	11.19%
-0.21575	223	11.83%
0.07987	219	11.62%
0.40148	249	13.21%
0.76540	211	11.19%
1.22470	210	11.14%
1.92173	103	5.46%



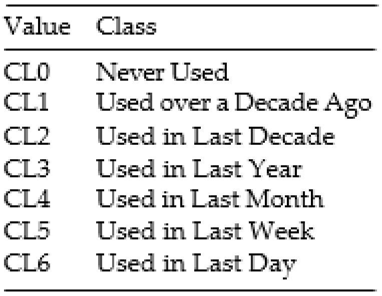
Min	Max	Mean	Std.dev.
-2.07848	1.92173	-0.00329	0.96370

14. Alcohol is a class of alcohol consumption. It is an output attribute with the following distribution of classes.

15. Amphet is a class of amphetamines consumption. It is an output attribute with the following distribution of classes.

16. Amyl is a class of amyl nitrite consumption. It is an output attribute with the following distribution of classes.

17. Benzos is a class of benzodiazepine consumption. It is an output attribute with the following distribution of classes:



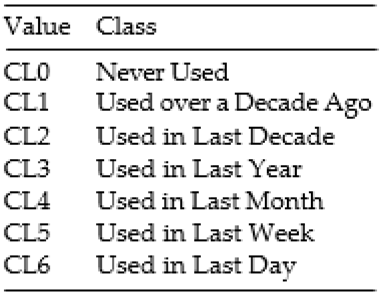
Value	Class
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day

18. Caff is a class of caffeine consumption. It is an output attribute with the following distribution of classes.

19. Cannabis is a class of cannabis consumption. It is an output attribute with the following distribution of classes.

20. Choc is a class of chocolate consumption. It is an output attribute with the following distribution of classes.

21. Coke is a class of cocaine consumption. It is an output attribute with the following distribution of classes:



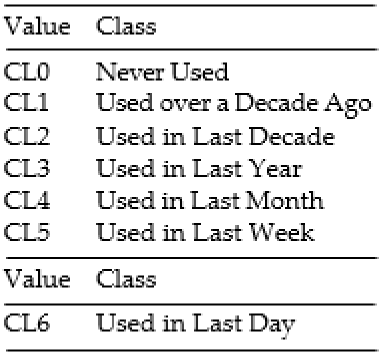
Value	Class
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day

22. Crack is a class of crack consumption. It is an output attribute with the following distribution of classes.

23. Ecstasy is a class of ecstasy consumption. It is an output attribute with the following distribution of classes.

24. Heroin is a class of heroin consumption. It is an output attribute with the following distribution of classes.

25. Ketamine is a class of ketamine consumption. It is an output attribute with the following distribution of classes.



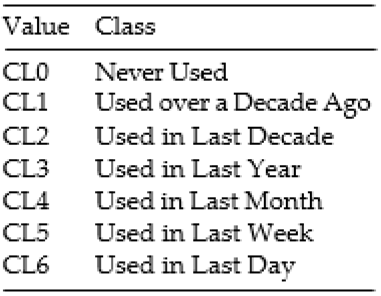
Value	Class
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
Value	Class
CL6	Used in Last Day

26. Legal is a class of legal high consumption. It is an output attribute with the following distribution of classes.

27. LSD is a class of alcohol consumption. It is an output attribute with the following distribution of classes.

28. Meth is a class of methadone consumption. It is an output attribute with the following distribution of classes.

29. Mushrooms is a class of magic mushrooms consumption.It is an output attribute with the following distribution of classes.

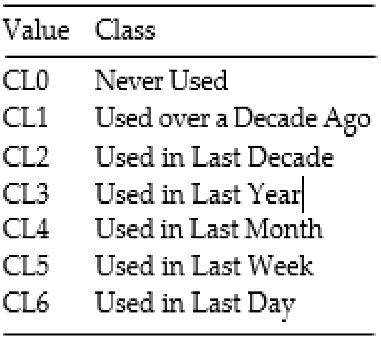


Value	Class
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day

30. Nicotine is a class of nicotine consumption. It is an output attribute with the following distribution of classes.

31. Semer is a class of fictitious drug Semeron consumption. It is an output attribute with the following distribution of classes.

32. VSA is a class of volatile substance abuse consumption. It is an output attribute with the following distribution of classes.



Value	Class
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

5. METHODOLOGY

5.1 Data Collection and Preprocessing

Preprocess the data available from the data repository. Handle missing values and ensure each row has real values to use for the project.

5.2 Correlation Analysis, Feature Selection

Conduct correlation analysis using the Python language among the variables in order to identify re- lationships between features and drug abuse. Select the features that display significant correlation and use that as the X feature data.

5.3 Dataset Preparation

Split the dataset into feature data (X) and target variable (Y), which will be used to create drug-use and abuse predictions in the machine learning models. Ensure the data is formatted properly for training.

5.4 Select Algorithms and Train

Choose a range of machine learning algorithms that can be trained on all 1885 respondents to predict drug-abuse for specific drugs. Use the trained models to generate predictions for drug use for each specific drug based on personality risk factors available from the dataset.

5.5 Model Evaluation, Model Selection

Evaluate the accuracy of each machine learning model's predictions against the actual data from the dataset. To evaluate the performance of the models, use train-test split to divide the dataset into a training and testing set, specifically a 70%-training to 30%-testing ratio. Assess accuracy, precision, recall, and F1 score. Focus on the performance of models where correlations were established between features and drug use for the algorithms.

Evaluate the effectiveness of each model's predictions using cross-validation with 5 folds. Use the results from the cross-validation analysis to evaluate the effectiveness of the model for a specific drug

Select the model that demonstrates the highest combination of AUC score and F1 score in its predictions to strike balance, ensuring it can differentiate and discriminate between non-user and user, and can identify positive instances of drug-use for a specific drug without many false positives.

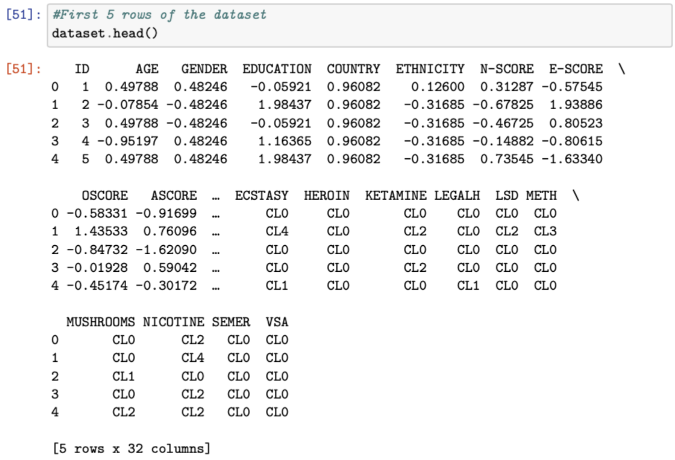
5.6 Assessment Metrics

Assess the final select model's performance using additional evaluation metrics. This includes precision, recall, and F1-score to ensure its reliability. Assess the AUC Score of the model for the specific drug using cross-validation.

5.7 Interpretation and Conclusion

Interpret the outcomes, and draw conclusions regarding effective models and their predictions for certain drugs.

6. DATASET PREPROCESSING



```
[51]: #First 5 rows of the dataset
      dataset.head()
```

```
[51]:    ID      AGE    GENDER  EDUCATION  COUNTRY  ETHNICITY  N-SCORE  E-SCORE  \
      0   1  0.49788  0.48246   -0.05921  0.96082    0.12600  0.31287 -0.57545
      1   2 -0.07854 -0.48246    1.98437  0.96082   -0.31685 -0.67825  1.93886
      2   3  0.49788 -0.48246   -0.05921  0.96082   -0.31685 -0.46725  0.80523
      3   4 -0.95197  0.48246    1.16365  0.96082   -0.31685 -0.14882 -0.80615
      4   5  0.49788  0.48246    1.98437  0.96082   -0.31685  0.73545 -1.63340

          OSCORE   ASCORE  ...  ECSTASY  HEROIN  KETAMINE LEGALH  LSD METH  \
      0 -0.58331 -0.91699  ...      CL0     CL0       CL0    CL0  CL0  CL0
      1  1.43533  0.76096  ...      CL4     CL0       CL2    CL0  CL2  CL3
      2 -0.84732 -1.62090  ...      CL0     CL0       CL0    CL0  CL0  CL0
      3 -0.01928  0.59042  ...      CL0     CL0       CL2    CL0  CL0  CL0
      4 -0.45174 -0.30172  ...      CL1     CL0       CL0    CL1  CL0  CL0

        MUSHROOMS NICOTINE SEMER  VSA
      0       CL0      CL2   CL0  CL0
      1       CL0      CL4   CL0  CL0
      2       CL1      CL0   CL0  CL0
      3       CL0      CL2   CL0  CL0
      4       CL2      CL2   CL0  CL0

      [5 rows x 32 columns]
```

7. FINDINGS

The code systematically compares each pair of columns in the correlation matrix and identifies those that surpass the specified correlation threshold. Based on the correlation analysis, using a correlation threshold of 0.35, we will use the following data attributes for each respondent from the dataset in order to train machine learning algorithms. - AGE - COUNTRY - N-SCORE - E-SCORE - OSCORE - IMPULSIVE - SS - AMPHET - AMYL - BENZOS - CANNABIS - COKE - CRACK - ECSTASY - HEROIN - KETAMINE - LEGALH - LSD - METH - MUSHROOMS - NICOTINE

The main takeaway is that the following data attributes cannot be used in this project due to their weaker relationships or lack of dependencies with other columns in the dataset. To prevent negatively affecting the performance of the algorithm and models, the following data attributes will not be used and have been removed from the data. - GENDER - ETHNICITY - EDUCATION - ASCORE - ALCOHOL - CAFF - CHOC - SEMER - VSA

8. FEATURE SELECTION

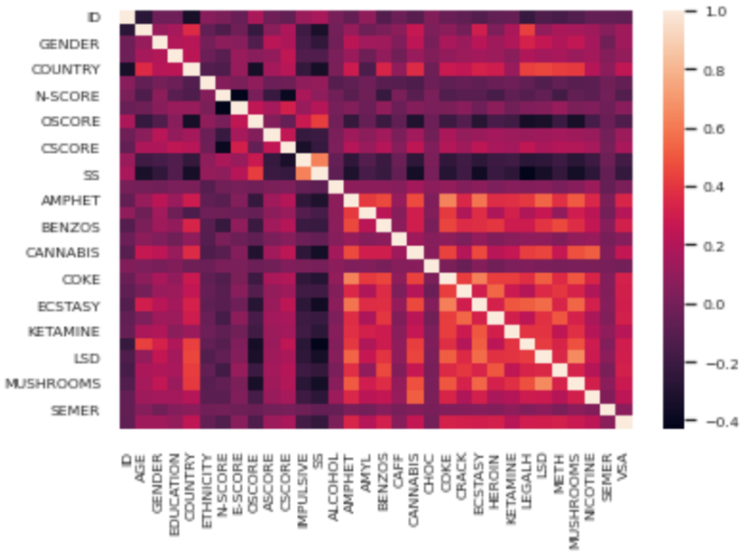
For the project, we are selecting the following as Features (X) to train the programs, as they are a comprehensive set of features that can be easily identified and determined in individuals. The following are relevant and impactful features that come from the results of the correlation analysis

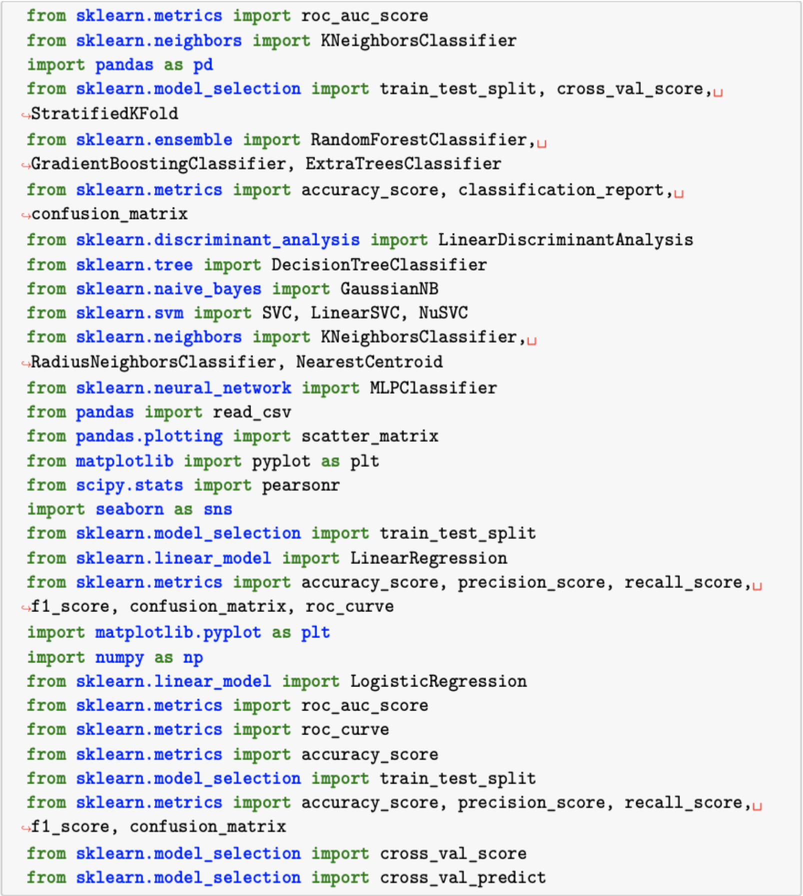
(X). - AGE - COUNTRY - N-SCORE - E-SCORE - OSCORE - CSCORE - IMPULSIVE - SS

The following are the target variables (Y) for the project. - AMPHET - AMYL - BENZOS - CANNABIS - COKE - CRACK - ECSTASY - HEROIN - KETAMINE - LEGALH - LSD - METH, MUSHROOMS - NICOTINE

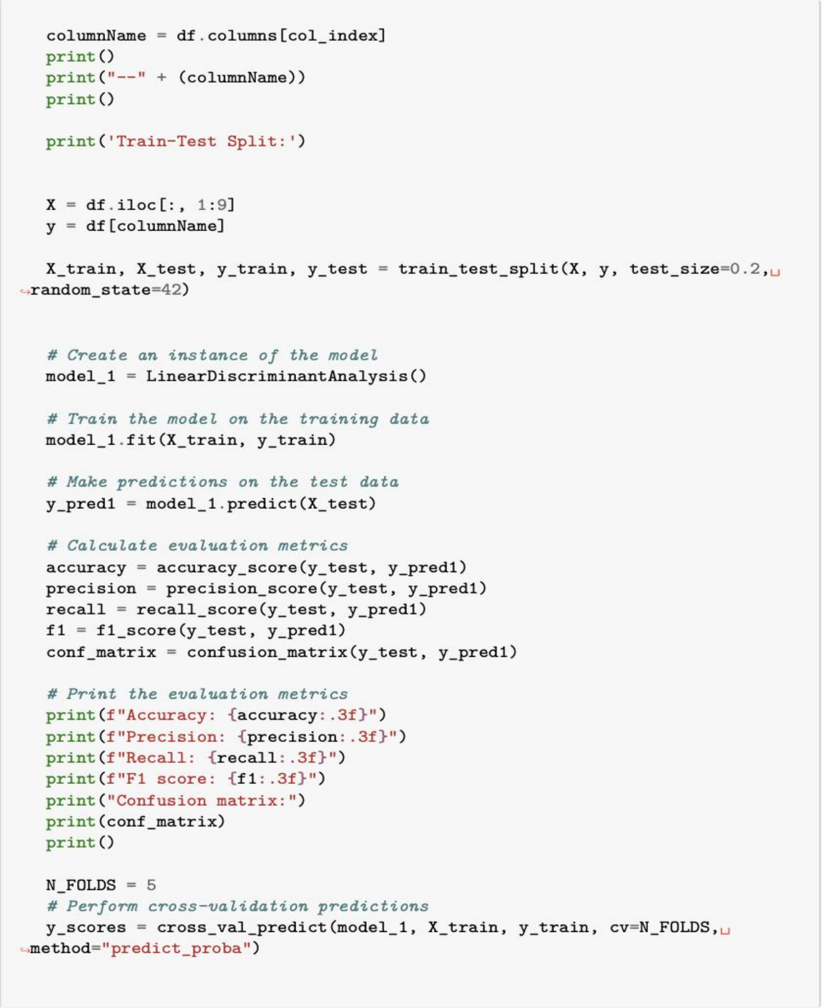
9. MODEL TRAINING AND LEARNING

We used a Supervised Learning setting, where we trained the models using features (X) to predict the target variable (Y) to maximize model performance. This ensured the model learned the relationship between input features and output variable based on label data. The models learn patterns from the training data and then predict on the unseen testing data to evaluate how accurate it is to new, unseen data. We used a 70% training to 30% testing ratio for the project while using a train-test split. After, we evaluated the effectiveness of the model's predictions using cross-validation with 5 folds. We used the results from the cross-validation analysis to evaluate the effectiveness of the model for a specific drug because it reduces variability in model performance estimation as train-test splits can produce variability due to the randomness in the data-split. Cross-validation ensured the results are more comprehensive by using different subsets of the data. We determined the AUC score and the ROC graph using cross-validation. We used the following machine learning models in this study, and evaluated the effectiveness and accuracy of each for all 14 drugs. - Linear Discriminant Analysis (LDA) - K-Nearest Neighbors (KNN) - Decision Tree (CART) - Gaussian Naive Bayes (NB) - Random Forest (RF) - Gradient Boosting Machine (GBM), Support Vector Machine (SVM) - Logistic Regression (LR)

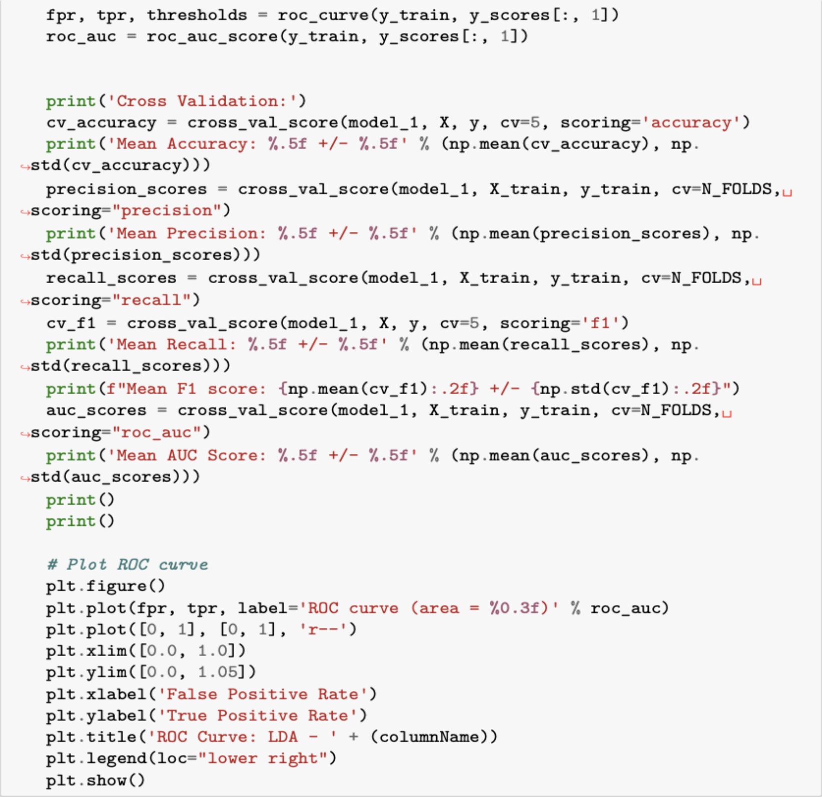




```
from sklearn.metrics import roc_auc_score
from sklearn.neighbors import KNeighborsClassifier
import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_score,␣
↪StratifiedKFold
from sklearn.ensemble import RandomForestClassifier,␣
↪GradientBoostingClassifier, ExtraTreesClassifier
from sklearn.metrics import accuracy_score, classification_report,␣
↪confusion_matrix
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC, LinearSVC, NuSVC
from sklearn.neighbors import KNeighborsClassifier,␣
↪RadiusNeighborsClassifier, NearestCentroid
from sklearn.neural_network import MLPClassifier
from pandas import read_csv
from pandas.plotting import scatter_matrix
from matplotlib import pyplot as plt
from scipy.stats import pearsonr
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score,␣
↪f1_score, confusion_matrix, roc_curve
import matplotlib.pyplot as plt
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score,␣
↪f1_score, confusion_matrix
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_val_predict
```



```
    columnName = df.columns[col_index]
    print()
    print("--" + (columnName))
    print()

    print('Train-Test Split:')


    X = df.iloc[:, 1:9]
    y = df[columnName]

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
  ↪random_state=42)


    # Create an instance of the model
    model_1 = LinearDiscriminantAnalysis()

    # Train the model on the training data
    model_1.fit(X_train, y_train)

    # Make predictions on the test data
    y_pred1 = model_1.predict(X_test)

    # Calculate evaluation metrics
    accuracy = accuracy_score(y_test, y_pred1)
    precision = precision_score(y_test, y_pred1)
    recall = recall_score(y_test, y_pred1)
    f1 = f1_score(y_test, y_pred1)
    conf_matrix = confusion_matrix(y_test, y_pred1)

    # Print the evaluation metrics
    print(f"Accuracy: {accuracy:.3f}")
    print(f"Precision: {precision:.3f}")
    print(f"Recall: {recall:.3f}")
    print(f"F1 score: {f1:.3f}")
    print("Confusion matrix:")
    print(conf_matrix)
    print()

    N_FOLDS = 5
    # Perform cross-validation predictions
    y_scores = cross_val_predict(model_1, X_train, y_train, cv=N_FOLDS,␣
  ↪method="predict_proba")
```



```
    fpr, tpr, thresholds = roc_curve(y_train, y_scores[:, 1])
    roc_auc = roc_auc_score(y_train, y_scores[:, 1])


    print('Cross Validation:')
    cv_accuracy = cross_val_score(model_1, X, y, cv=5, scoring='accuracy')
    print('Mean Accuracy: %.5f +/- %.5f' % (np.mean(cv_accuracy), np.
  ↪std(cv_accuracy)))
    precision_scores = cross_val_score(model_1, X_train, y_train, cv=N_FOLDS,␣
  ↪scoring="precision")
    print('Mean Precision: %.5f +/- %.5f' % (np.mean(precision_scores), np.
  ↪std(precision_scores)))
    recall_scores = cross_val_score(model_1, X_train, y_train, cv=N_FOLDS,␣
  ↪scoring="recall")
    cv_f1 = cross_val_score(model_1, X, y, cv=5, scoring='f1')
    print('Mean Recall: %.5f +/- %.5f' % (np.mean(recall_scores), np.
  ↪std(recall_scores)))
    print(f"Mean F1 score: {np.mean(cv_f1):.2f} +/- {np.std(cv_f1):.2f}")
    auc_scores = cross_val_score(model_1, X_train, y_train, cv=N_FOLDS,␣
  ↪scoring="roc_auc")
    print('Mean AUC Score: %.5f +/- %.5f' % (np.mean(auc_scores), np.
  ↪std(auc_scores)))
    print()
    print()

    # Plot ROC curve
    plt.figure()
    plt.plot(fpr, tpr, label='ROC curve (area = %0.3f)' % roc_auc)
    plt.plot([0, 1], [0, 1], 'r--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('ROC Curve: LDA - ' + (columnName))
    plt.legend(loc="lower right")
    plt.show()
```

10. RESULTS & DISCUSSION

After training and testing the different classification models, there are models that have high percentages for their accuracy, precision, recall, and F1 Score for specific drugs. However, in the context of this research, it is important to note that these percentages on their own do not necessarily give us the entire picture of the effectiveness of these models. There were many models that had recall percentages well above 90%, however, for this paper, we are also taking into account the AUC score and the F1 score because it gives us a greater understanding of the model's ability to discriminate between classes and its true positive and false positive rates, rather than predicting the majority of outputs.

For example, there were many drugs under the SVM model that scored relatively high for a specific drug's accuracy, precision, recall, and F1 Score. However, the AUC score was a little over 0.5, which is essentially random chance. Since the AUC score was not near 1 for some of these results even with the high accuracy, we cannot use these results because it demonstrates that the dataset was imbalanced for that specific drug (such as a very hard-core, uncommon drug with very few users) where one class significantly outnumbered the other. For such hard drugs, the model can accurately predict that most people were non-users, however, its AUC score tells us that it lacks the ability to discriminate between users and non-users. The classifier may have learned to predict the majority class more frequently, which explains the high accuracy but the poor performance in distinguishing the minority class. As a result, for this paper, we finalized our results with models that had both relatively high F1 and AUC scores for specific drugs with a model.

For this research, we purposefully also calculated the AUC and F1 scores of each drug with each model. Together, analyzing the AUC score alongside the model's F1 Score for each drug allowed us to conclude the most effective model that accurately predicts true drug-use and has the ability to distinguish across various thresholds.

The following are the results from the study. The table showcases the models with relatively high F1 and AUC scores for specific drugs across all models.

Model	Drug Name	F1 Score	AUC Score
Linear Discriminant Analysis	Cannabis	0.87	0.835
Linear Discriminant Analysis	LegalH	0.70	0.846
K-Nearest Neighbors	Cannabis	0.86	0.771
Gaussian Naive Bayes	Cannabis	0.82	0.834
Gaussian Naive Bayes	LegalH	0.7	0.8377
Random Forest	Cannabis	0.87	0.815
Random Forest	Coke	0.61	0.837
Random Forest	LSD	0.67	0.795
Random Forest	Mushrooms	0.70	0.792
Gradient Boosting Machine	Cannabis	0.87	0.818
Gradient Boosting Machine	Ecstasy	0.67	0.785
Gradient Boosting Machine	LegalH	0.68	0.809
Support Vector Machine	Cannabis	0.88	0.781
Support Vector Machine	Ecstasy	0.68	0.782
Support Vector Machine	LegalH	0.70	0.835
Support Vector Machine	Mushrooms	0.70	0.796
Logistic Regression	Cannabis	0.88	0.836
Logistic Regression	Ecstasy	0.62	0.781
Logistic Regression	LegalH	0.69	0.846
Logistic Regression	LSD	0.67	0.817
Logistic Regression	Mushrooms	0.67	0.799
Logistic Regression	Nicotine	0.87	0.696

The following is a condensed version of the above list, showing models that have great potential for future applications of our research results to increase the effectiveness of these models, and using other models for these specific drugs.

Model	Drug Name	F1 Score	AUC Score
Logistic Regression	Cannabis	0.88	0.836
Linear Discriminant Analysis	LegalH	0.70	0.846
Logistic Regression	Nicotine	0.87	0.696

11. CONCLUSION & FUTURE PLANS

In summary, our study demonstrates that machine learning algorithms are capable of accurately predicting the use of drugs for an individual using their data related to demographics and personality factors. Specifically, the ability for these models to predict the use of Cannabis, LegalH, and Nicotine was effective. Based on these results, the classification model we identified to be the most effective for predicting the use of the above drugs, given high F1 and AUC scores, was Logistic Regression for the drug Cannabis. Over the next few years, we plan to conduct additional testing using different machine learning training and testing techniques to increase the F1 and AUC scores for these drugs and other untested drugs using different drug-use data repositories. Following these additional tests, the next step is to create a website and an IOS software app to ensure the widespread accessibility of the results of this study for individuals, families, and counselors. The results of this study can be used as an additional tool in assessing the risk of drug use for specific drugs in adolescents to counteract the drug abuse epidemic in America.

REFERENCES

[1] Black, J. (2022). An Introduction to Machine Learning for Classification and Prediction. Family Practice, XX, 1–5.

[2] Federal Bureau of Prisons.(2023, November 11).BOP statistics: In- mate offenses. Federal Bureau of Prisons; Federal Bureau of Prisons. https://www.bop.gov/about/statistics/statistics_inmate_offenses.jsp

[3] Johns Hopkins Medicine. (2019). Substance Abuse / Chemical Dependency. John Hop- kins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/substance- abuse-chemical-dependency

[4] National Center for Drug Abuse Statistics. (2020). NCDAS: Substance Abuse and Addiction Statistics [2020]. National Center for Drug Abuse Statistics. https://drugabusestatistics.org/

[5] National Drug Intelligence Center.(2006, January).The Impact of Drugs on Society - National Drug Threat Assessment 2006. Justice.gov. https://www.justice.gov/archive/ndic/pubs11/18862/impact.htm

[6] Substance Use Has Risen During COVID-19 Pandemic. (2022, March 15). Www.rpc.senate.gov. https://www.rpc.senate.gov/policy-papers/substance-use-has-risen- during-covid-19-pandemic

[7] UCI Machine Learning Repository. (2016). Archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified