Exploring the Correlation Between Drug Use and Mental Illness

Introduction:

In our project, we will try to find a correlation between drug use and mental illness. Our dataset contains records of people who were hospitalized in rehabilitation centers in the USA. It includes information about their social status, drug use history, and demographic information.

Finding mental illness among people is a challenging problem that involves multiple tests and experts analyzing a person's behavior over a period of time, which can even take up to one year. Furthermore, individuals with mental illness may not always be aware of it due to the difficulty of the diagnostic process, which can be critical in rehabilitation problem.

By working the data, we built a regression model that predict mental illness among drug users based on their information. This regression model can aid in identifying mental illness among drug users. Finding these illnesses/ being aware of them can help the users and or the people treating them, by giving them special care and providing them with proper treatment, which could help the treatment and even save lives, because of the acknowledgement of the mental illness. Our model can help to find about mental illness by putting in information that he already has and help finding if a person is sick with high probability.

There are many studies that show correlation between drug use and mental illness, and that drug use can cause it. In our project we try to find among the drug users, who is more likely to have a specific illness.

Data Overview:

For our project, we utilized the Treatment Episode Data Set: Admissions (TEDS-A).TEDS is a system that collects treatment data from states to monitor their substance use treatment systems. TEDS is composed of two major components: TEDS-A, which focuses on admissions data, and TEDS-D, which captures discharge records.

- TEDS-A provides demographic, clinical, and substance use characteristics of admissions to alcohol or drug treatment in facilities that report to state administrative data systems. TEDS-A has two parts: a minimum data set (collected by all states) and a supplemental data set (collected by some states).The minimum data set consists of 19 items that include demographic information, primary, secondary, and tertiary substances used by the subject, their route of administration, frequency of use, and age at first use ,source of referral to treatment, number of prior treatment episodes, and service type. The supplemental data set includes 15 psychiatric, social, and economic items.

- TEDS-D is the discharge records component of the TEDS system. It includes the same variables as the admissions (TEDS-A) component, with the addition of type of service at discharge, length of stay, and reason for discharge or discontinuation of service. In our project, we utilized features encompassing demographics (age, gender, race), clinical factors (primary substance use, frequency of use, DSM diagnosis), treatment-related details (service type, medication-assisted therapy), and additional variables including education, veteran status, employment, and attendance at self-help groups.

In our project, we utilized features encompassing demographics (age, gender, race), clinical factors (primary substance use, frequency of use, DSM diagnosis), treatment-related details (service type, medication-assisted therapy), and additional variables including education, veteran status, employment, and attendance at self-help groups.

Methods and Results:

In order to investigate our research question, we opted to perform logistic regression models for each of the four mental diseases. Initially, we assessed the relationship between variables in our dataset. We utilized the Cramer's V coefficient to calculate the correlation between each pair of variables in the dataset, and the results were recorded in a correlation matrix. By examining this matrix, we can detect any significant correlations between variables.

	CASEID	STFIPS	VET	EDUC	SERVICES	NOPRIOR	PSOURCE	EMPLOY	METHUSE	PSYPROB	GENDER	DSMCRIT	AGE	RACE	SUB1	FREQ1	FREQ_ATND_SELF_HELP
CASEID	1	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
STFIPS	1	1.00000	0.05862	0.13550	0.33960	0.19140	0.23640	0.14440	0.39120	0.61900	0.14800	0.16740	0.11520	0.35070	0.16220	0.27970	0.20240
VET	1	0.05862	1.00000	0.05402	0.08432	0.03028	0.03669	0.03636	0.01720	0.02838	0.09145	0.03048	0.11200	0.04325	0.07203	0.05027	0.02959
EDUC	1	0.13550	0.05402	1.00000	0.05970	0.03331	0.06327	0.09224	0.04162	0.11270	0.07333	0.06035	0.19780	0.06110	0.10940	0.05838	0.03982
SERVICES	1	0.33960	0.08432	0.05970	1.00000	0.10570	0.08498	0.10530	0.09561	0.13300	0.08939	0.07639	0.08121	0.08053	0.13600	0.23020	0.09129
NOPRIOR	1	0.19140	0.03028	0.03331	0.10570	1.00000	0.05948	0.05817	0.11720	0.06001	0.03281	0.10180	0.09556	0.04995	0.12220	0.07442	0.08466
PSOURCE	1	0.23640	0.03669	0.06327	0.08498	0.05948	1.00000	0.05015	0.14660	0.11930	0.10440	0.08527	0.08157	0.06940	0.10630	0.10750	0.07831
EMPLOY	1	0.14440	0.03636	0.09224	0.10530	0.05817	0.05015	1.00000	0.01446	0.05494	0.06998	0.09824	0.20290	0.08144	0.10130	0.03347	0.06073
METHUSE	1	0.39120	0.01720	0.04162	0.09561	0.11720	0.14660	0.01446	1.00000	0.07449	0.03739	0.04192	0.07916	0.19890	0.21960	0.02871	0.06044
PSYPROB	1	0.61900	0.02838	0.11270	0.13300	0.06001	0.11930	0.05494	0.07449	1.00000	0.02944	0.06454	0.09656	0.09667	0.12470	0.09279	0.08733
GENDER	1	0.14800	0.09145	0.07333	0.08939	0.03281	0.10440	0.06998	0.03739	0.02944	1.00000	0.17580	0.09115	0.07850	0.14720	0.03808	0.02986
DSMCRIT	1	0.16740	0.03048	0.06035	0.07639	0.10180	0.08527	0.09824	0.04192	0.06454	0.17580	1.00000	0.09030	0.10470	0.10290	0.03434	0.04563
AGE	1	0.11520	0.11200	0.19780	0.08121	0.09556	0.08157	0.20290	0.07916	0.09656	0.09115	0.09030	1.00000	0.07607	0.13370	0.09095	0.06642
RACE	1	0.35070	0.04325	0.06110	0.08053	0.04995	0.06940	0.08144	0.19890	0.09667	0.07850	0.10470	0.07607	1.00000	0.12740	0.07115	0.05151
SUB1	1	0.16220	0.07203	0.10940	0.13600	0.12220	0.10630	0.10130	0.21960	0.12470	0.14720	0.10290	0.13370	0.12740	1.00000	0.19200	0.09720
FREQ1	1	0.27970	0.05027	0.05838	0.23020	0.07442	0.10750	0.03347	0.02871	0.09279	0.03808	0.03434	0.09095	0.07115	0.19200	1.00000	0.09723
FREQ_ATND_SELF_HELP	1	0.20240	0.02959	0.03982	0.09129	0.08466	0.07831	0.06073	0.06044	0.08733	0.02986	0.04563	0.06642	0.05151	0.09720	0.09723	1.00000

Upon examining the correlation matrix, we observed strong associations between the STFIPS column and VET, SERVICES, PSYPROB, RACE, and METHUSE. Consequently, we made the decision to exclude this column from our dataset.

Next, we constructed logistic regression models for each of the four mental diseases. Logistic regression is a suitable modeling technique for binary classification tasks, where the dependent variable represents the presence or absence of a particular mental illness. We chose logistic regression as our modeling approach because it allows us to estimate the probability of having a specific mental illness based on the given predictors.

The performance of the logistic regression models was evaluated using several metrics, including AUC-ROC, F1 score, recall, precision, and the number of false negatives (FN), true negatives (TN), false positives (FP), and true positives (TP). These metrics provide insights into the model's accuracy, precision, and ability to correctly identify individuals with the respective mental illnesses.

The key results of our analysis, as shown in the provided table, are as follows:

	TP	FP	TN	FN	Precision	Recall	F1	AUC-ROC
Schizophrenia	58	106	633	53	0.3536	0.5225	0.4218	0.7472
Anxiety	118	289	570	56	0.2899	0.6781	0.4061	0.6311
Bipolar	75	229	606	63	0.2467	0.5434	0.3393	0.6049
Depressive	315	406	422	6	0.4368	0.9813	0.6046	0.5783

Based on the results provided, we can observe that the predictive performance of the logistic regression models varies across different mental diseases. Specifically, the model for schizophrenia demonstrates relatively good predictive results, while the model for depression yields less accurate predictions.

One notable characteristic of our approach is the emphasis on minimizing false negatives (FN), which are cases where individuals have the mental illness but are incorrectly classified as not being sick.

To achieve this, our models may have a relatively higher number of false positives (FP), where individuals are classified as having a specific mental illness even if they do not actually

have it. This cautious approach is deliberate, as it allows for further tests and assessments to be conducted to confirm the presence of the illness. By erring on the side of classifying individuals as potentially being sick, we reduce the risk of missing genuine cases and ensure that those who require additional care and treatment receive the necessary attention.

This approach aligns with the diagnostic process followed in the medical field, where comprehensive evaluations and multiple tests are often employed to confirm the presence of various diseases.

The AUC-ROC values for schizophrenia (0.7472), anxiety (0.6311), and bipolar disorder (0.6049) indicate relatively good predictive performance, while the AUC-ROC value for depressive disorder (0.5783) is lower. This discrepancy may suggest that the model's ability to accurately classify depressive disorder is relatively weaker compared to the other mental illnesses. Further investigation is necessary to identify the factors contributing to this lower performance and to refine the model to improve its predictive capabilities for depressive disorder.

In conclusion, our logistic regression models provided valuable insights into the correlation between drug use and mental illness. While the models demonstrated relatively good predictive performance for schizophrenia, anxiety, and bipolar disorder, the model for depressive disorder showed lower accuracy. Our cautious approach prioritized minimizing false negatives, leading to a higher number of false positives. Further investigation is needed to enhance the model's performance for depressive disorder and refine it for more accurate .predictions in the future.

Limitations and Future Work:

Despite the valuable insights gained from our project, there are several limitations to consider when generalizing beyond our sample. One limitation is the reliance on reports filled in by social workers, which may introduce potential bias as we cannot verify their accuracy. Additionally, our models cannot predict if a person has multiple mental illnesses, which is a common occurrence.

Given more time, we would pursue several future directions. First, we would explore alternative algorithmic models to gather more data and improve our predictions. We would also seek expert input to identify the specific characteristics associated with each mental illness and collect additional data accordingly. Additionally, we would investigate the connection between the ingredients of different drugs and mental triggers to gain a deeper understanding of their relationship.

While our project provides valuable insights into the correlation between drug use and mental illness, it is important to acknowledge the limitations of our approach and findings. With additional time, we would address these limitations by refining our models, gathering more data, and our further investigating the underlying factors contributing to the predictive performance of models.