

Advanced Applied Statistics: Drug Use Based on Different Personality Traits

Course code: KSADAPS1KU

Amna ejam@itu.dk
Chrisanna Kate Cornish ccor@itu.dk
Christian Margo Hansen chmh@itu.dk
Zainab Ali Shaker Khudoir zakh@itu.dk

Date: January 4, 2024

IT UNIVERSITY OF COPENHAGEN

Drug Use Based on Different Personality Traits

Amna
ejam@itu.dk

Chrisanna Kate Cornish
ccor@itu.dk

Christian Margo Hansen
chmh@itu.dk


Zainab Khudoir
zakh@itu.dk

Abstract—Utilising a dataset including personality data and drug use, developed by Elaine Ferhman, the statistical significance of such traits as indicators of drug use was investigated.

A Gaussian Mixture Model was used to find clusters within the dataset, which aided to distinguish between users and non-users. We then applied a logistic regression model, and analysed whether personality traits provided useful information to such a model; for predicting illegal drug use as a whole, and each drug in the dataset separately.

The clustering indicated that there essentially exists two distinct personality groupings, and that one had a general lower use of illegal drugs for the majority of individuals, compared to the other group.

Finally, we conclude that personality traits are indicative of illegal drug use as a whole, and can be used as a weak, but statistically significant, indicator of specific drug use.

The code used for this report can be found at: Applied Statistics 2023 

I. INTRODUCTION

Assessing an individual’s susceptibility to drug consumption and potential misuse represents a complex challenge. In today’s society, drug-related issues are widespread, affecting various communities and age groups. Therefore, understanding people’s relationship to drugs is becoming increasingly important. A number of studies have shown statistical evidence of a link between individuals’ personalities and their drug consumption [5]. These findings suggest that certain personality traits can potential be a key factor in influencing the likelihood of engaging in drug-related behaviours.

The dataset used for this project was collected between March 2011 and March 2012 by Elaine Fehrman, and consists of 1885 entries of personality and demographic data in connection to various legal and illegal substances used by the individuals [1]. A brief overview of the dataset can be seen in Table I. We opted to exclude the demographic data present throughout our study, as there exists bodies of work that primarily examine the influence of features such as age, gender, and ethnicity ([2] [3] [4]); additionally, there was the desire to avoid risking the influence of these features from obscuring the contributions of personality traits. The personality traits in the dataset are primarily determined by the Revised NEO Five-Factor Inventory methodology, summarized by Fehrman as:

- **Neuroticism (N)**: a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression
- **Extraversion (E)**: manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics

- **Openness to experience (O)**: a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests
- **Agreeableness (A)**: a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness
- **Conscientiousness**: a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient.

In addition to these traits, **Impulsiveness (I)** was quantified using the Barratt Impulsiveness Scale, indicated by measures of ‘*acting without thinking*’, ‘*poor concentration and thought intrusions*’, and ‘*lack of consideration for consequences*’. **Sensation (S)** was measured through Impulsiveness Sensation-Seeking, representing general sensation-seeking and is considered ‘*a valid and reliable measure of high risk [behaviour]*’.

Category	Variable	Description
	ID	Record ID
Demographic	Age Gender Education Country Ethnicity	Age range of survey participant Gender of participant Highest level of education achieved Country of current residence Participant ethnicity
Personality	Neuroticism Extraversion Openness Agreeableness Conscientiousness Impulsiveness Sensation	NEO-FFI-R Neuroticism NEO-FFI-R Extraversion NEO-FFI-R Openness to experience NEO-FFI-R Agreeableness NEO-FFI-R Conscientiousness BIS-11 Impulsiveness ImpSS Sensation seeking
Drugs	Alcohol Amphet Amyl Benzos Caff Cannabis Choc Coke Crack Ecstasy Heroin Ketamine Legalh LSD Meth Mushrooms Nicotine Semer VSA	Alcohol consumption Amphetamines consumption Amyl Nitrite consumption Benzodiazepine consumption Caffeine consumption Cannabis consumption Chocolate consumption Cocaine consumption Crack consumption Ecstasy consumption Heroin consumption Ketamine consumption Legal Highs consumption LSD consumption Methadone consumption Magic Mushrooms consumption Nicotine consumption Fictitious drug Semeron consumption Volatile Substance Abuse consumption

TABLE I: Overview of Drug Consumption dataset variables

The frequency of use for each drug was recorded through seven different classifications:

- **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

The investigation follows the similar objectives of Fehrman et al [5], to examine personality traits as a predictor for drug use. Use of drugs, illegal or otherwise, is a phenomenon not isolated to any specific socioeconomic status [6] or ethnic group. Focusing specifically on personality traits, and working to identify any potential relationships between such traits and drugs use, facilitates the understanding of the internal factors that contribute towards an individual's potential drug use and abuse, regardless of their background. This can hypothetically aid in identifying individuals with a risk of drug abuse, or relapse, and encourage the development of unbiased support systems. Formally, we investigate:

Are personality traits statistically significant as indicators of drug use?

II. METHODS

A. Gaussian Mixture Models

To begin the exploration of our dataset, we will use a clustering technique to investigate potential groupings/clusters underlying the data. More specifically, as we are focusing our study to personality traits we want to investigate patterns and groupings of individuals with similar personality types. To do this we employ Gaussian Mixture Models (GMM) as a clustering technique. This approach was chosen due to its suitability for datasets with complex relationships and overlapping clusters, where traditional methods might fall short. GMM allows for pattern discovery in the data by accommodating varying shapes and capturing nuanced relations among variables [10]. Considering that personality traits' distributions seemed to mostly follow normal curves (see Fig. 1), GMM was a natural choice.

GMM presents a flexible approach to probability distribution modelling, surpassing the limitations of single Gaussian distributions [10]. Defined by the parameters μ_k (mean), Σ_k (covariance matrix), and π_k (weight), the distribution is a blend of Gaussian components:

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

- $\mathcal{N}(x|\mu_k, \Sigma_k)$: Distribution of k 'th Gaussian. The probability of x given that we are in class k .
- π_k : The weight of the k 'th Gaussian, the probability of being in class k .
- $p(x)$: The marginal distribution of x as a Gaussian mixture, i.e. sum of all weighted Gaussian's.

The GMM assumes each cluster as a different Gaussian distribution. The Expectation-Maximisation (EM) algorithm is used in the GMM to find the best fit by iteratively optimising parameters [10].

1) *GMM Evaluation and Hyper Parameter Selection:* We used the Silhouette score ¹ to evaluate cluster quality, where a score closer to one indicates better clustering. The hyper parameter $n_{clusters}$, number of clusters, is selected based on this evaluation. The optimal number of clusters is found to be 2 when evaluating different² GMM on personality traits, as it maximises the Silhouette score [11] [12].

B. Principal Component Analysis

To visualise the clusters obtained from GMM analysis, we performed Principal Component Analysis (PCA) to reduce the data from 7 dimensions to 2 dimensions. This reduction provided a 2-dimensional representation of the feature space.

C. Multiple Hypotheses Testing

Furthermore, the data consists of a wide range of drugs. We want to investigate how individuals' drug behaviour are different in the personality clusters obtained from the GMM. We do this by conducting Multiple Hypothesis testing on types of drugs, using Chi-square test for independence for comparing distributions of a specific categorical variable (drug) in the different clusters [13].

Our hypotheses are as follows:

H_0 : There is no difference in the use of a specific drug in the personality clusters

$H_{1,...,18}$: There is a difference in the use of a specific drug in the personality clusters

We test our hypotheses on all the drugs except for 'Semer', the fictitious drug, employing an initial significance value (α) of 0.05.

We either reject or fail to reject the null hypothesis about each drug based on the comparison of its p-value with the fixed alpha value of 0.05. We reject the null hypothesis when the drug's p-value is less than 0.05 and we fail to reject the null hypothesis when the drug's p-value is greater than 0.05. Multiple Hypothesis testing with the basic alpha value of 0.05 gives rise to increasing the Family-wise Error Rate (FWER).

1) *Bonferroni Correction:* To account for multiple hypotheses, we applied Bonferroni correction, which involves reducing the significance level (α) to minimise the FWER. The corrected α value was calculated as $\frac{\alpha}{k}$, where k is the number of hypotheses (drugs) in our dataset. A new α_{new} value of 0.00278 was derived, resulting in a stricter criterion for statistical significance [9].

α by k :

$$\alpha_{new} = \frac{\alpha}{k} = \frac{0.05}{18} = 0.00278$$

¹Silhouette score checks how much the clusters are compact and how well they are separated.

²With random.seed(42)

D. Logistic Regression

Logistic regression was utilised to classify drug use, utilising the seven personality traits as the independent variables. Binary classification experiments were conducted for a super-class of ‘any illegal drug use’ and for each drug individually, making a total of 19 tests. The divide between users and non-users was established following the results from GMM clustering (see section III-B: *GMM Clustering*), with groups CL0-CL2 considered non-users, and CL3-CL6 as users. We based the ‘any illegal drug’ category on what was legal in the UK at the time of the data collection, as most of the participants were from the UK, and the original study was authored there.

The dataset was split into three parts. First, 200 samples of the data was randomly selected and set aside as a holdout test set³. For each drug, the remaining 1685 samples were split 80/20 into training and validation sets, using a stratified splitting method⁴ due to the unbalanced nature of the dataset to create a more representative training set. It was necessary to make separate splits for each drug classifier, as some drugs were highly used (for example, Chocolate, Caffeine), and others very rarely (see Appendix A for percentage of drug usage present in the dataset).

We tuned the threshold individually for each experiment to maximise the F1 score of the classification. A null model was also created to randomly assign a class according to the proportions present in the training dataset. This was repeated 10,000 times and an F1 score calculated for each run. With the assumption under central limit theorem that the large number of samples means that the generated F1 scores are normally distributed, we could then compare the F1 score obtained from the logistic model against the generated distribution. This allowed us to examine if there was a statistically significant result from the classifier, and verify whether or not utilising personality trait information in predicting drug use was more useful than random assignment.

To account for the multiple hypotheses, we will again apply a Bonferroni correction. To calculate the new α value, we divide α by k :

$$\alpha_{new} = \frac{\alpha}{k} = \frac{0.05}{19} = 0.00263$$

1) *Hypotheses*: We intend to test whether personality trait data is able to predict for some particular drug use. Formally:

H_0 : Personality traits are not useful to predict drug use

$H_{1,...,19}$: Personality traits are useful for predicting specific drug use

III. RESULTS AND ANALYSIS

A. Exploratory Data Analysis

Our data consists of a mixture of continuous numerical variables (personality traits) and categorical variables (drugs).

³pd.sample(200, replace=False, random_state=42)

⁴sklearn: train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

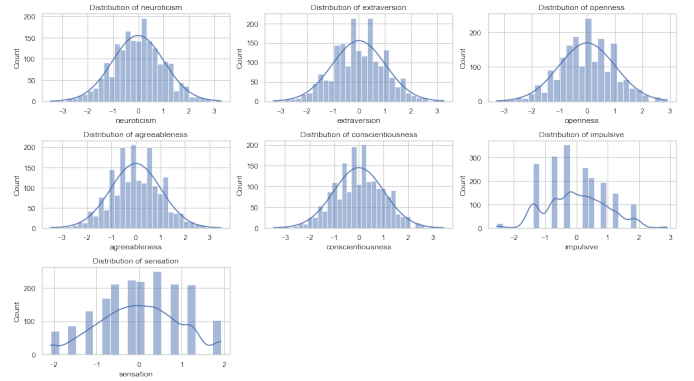


Fig. 1: Distributions of the different personality traits

Fig. 1 illustrates the distribution of each personality trait. It seems, from the plots, that the distributions generally follow the distinctive normal bell-curves.



Fig. 2: Correlations between variables

Correlations between variables were also looked at, and we found that whilst generally there were small or weak correlations, some variables did exhibit stronger ties. In Fig. 2, we can see traits such as sensation-seeking and impulsiveness have a fairly strong correlation. As we intend to use logistic regression which assumes independence between variables, we checked the multicollinearity between traits by calculating the Variance Inflation Factor⁵, shown in Table II. A score higher than 1 shows some multicollinearity, but in this case they are not high enough to significantly effect model interpretation.

Another notable attribute is that in all but 2 drugs (Nicotine and Cannabis), the dataset is imbalanced. Chocolate and Caffeine have almost all participants as users ($\approx 97\%$), and most of the illegal drugs have fewer users (10-30%). A few drugs, such as Crack and VSA, have an extremely small proportion of users (4-5%) (see Appendix A).

⁵<https://www.statology.org/multicollinearity-regression/>

Trait	VIF
Neuroticism	1.40
Extraversion	1.49
Openness	1.28
Agreeableness	1.15
Conscientiousness	1.40
Impulsiveness	1.79
Sensation	1.91

TABLE II: Multicollinearity between personality traits

B. GMM Clustering

Results from the GMM are illustrated in Fig. 3.

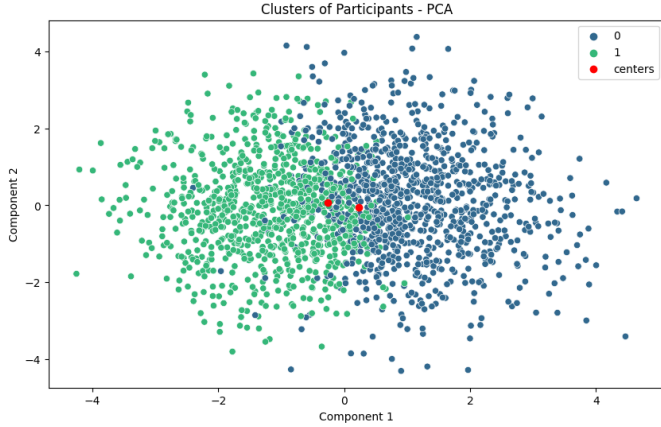


Fig. 3: Plot of the two different clusters of individuals based on their personality traits. Each point represents an individual. The different clusters are color-coded.

This plot helps visualise the separation between clusters in reduced dimensions. Our results can be interpreted as follows; people in the dataset are divided into two main personality types, namely those belonging to Cluster 0 and those belonging to Cluster 1.

Personality Trait	Cluster 0 Median Score	Cluster 1 Median Score
Neuroticism	0.313	-0.246
Extraversion	0.004	-0.046
Openness	0.293	-0.318
Agreeableness	-0.301	0.288
Conscientiousness	-0.527	0.585
Impulsiveness	0.530	-0.712
Sensation	0.765	-0.526

TABLE III: Cluster Characteristics Summary. For each cluster, the median value of the different personality scores are calculated.

1) *Cluster Characteristics*: We can consider the following cluster characteristics for each cluster:

- Cluster 0: Higher median scores for neuroticism, openness, sensation, and impulsivity. Higher percentage drug use across various substances, including, Amphetamines,

Drug	p-value	Significant
Amphet	6.42e-56	Yes
Amyl	3.39e-11	Yes
Benzos	8.14e-38	Yes
Cannabis	9.55e-82	Yes
Coke	7.69e-44	Yes
Crack	3.67e-18	Yes
Ecstasy	2.75e-55	Yes
Heroin	5.85e-21	Yes
Ketamine	1.749e-22	Yes
LSD	1.40e-52	Yes
Meth	2.49e-26	Yes
Mushrooms	5.56e-55	Yes
VSA	3.25e-27	Yes
Alcohol	0.00013	Yes
Caffeine	0.032	No
Chocolate	0.048	No
LegalH	5.81e-60	Yes
Nicotine	7.49e-40	Yes

TABLE IV: Results after conducting Multiple Hypothesis testing with Chi-square test and Bonferroni Correction ($\alpha = 0.00278$)

Ecstasy, etc. See Appendix B for a comprehensive overview of percentage of drug users in each cluster.

- Cluster 1: Higher median scores for agreeableness and conscientiousness. Lower drug consumption across various substances compared to Cluster 0. Interestingly, for any illegal drug more than 90% of individuals in Cluster 1 are spread from CL0, CL1 and CL2 (except for Cannabis (69.3%) and Ecstasy (88.2%)). This indicates none to very low drug consumption for vast majority of people belonging to Cluster 1.

These cluster characteristics provide a high-level overview of how individuals within each cluster differ in terms of personality traits. In summary, individuals in Cluster 0 have higher aggregated scores for Neuroticism, Openness, Impulsiveness, Sensation (highest) and lower scores for Conscientiousness and Agreeableness. In contrast, individuals in Cluster 1 have lower aggregated scores for Neuroticism, Openness, Sensation, Impulsiveness (lowest) and higher scores for Conscientiousness and Agreeableness. Median scores for Extraversion seems to be more or less centered around 0 for both clusters.

2) *Multiple Hypothesis Testing with Bonferonni Correction*: Table IV represents results of performing chi-squared testing on each of the 18 drugs, respectively comparing their distributions in each cluster. It follows, whenever a null hypothesis with respect to a certain drug is rejected (it is statistically significant) or fails to be rejected (it is statistically insignificant).

The chi-square results for Caffeine and Chocolate between Cluster 0 and Cluster 1, are statistically insignificant. Contrarily, all other drugs showcases a statistical significance, indicating a statistical difference between drug distributions in the two distinct personality clusters. This aligns with the our

expectation, that drug consumption patterns differ in different personality clusters, as we previously observed significantly lower drug use behaviour in Cluster 1 compared to Cluster 0.

C. Logistic Regression

As mentioned earlier, we created a model for each of the 18 drugs and the overall Illegal Drug category. The resulting F1 score from the holdout test dataset are shown in Table V. The results indicate we are able to accept 14 of our 19 hypotheses; that we are able to use personality traits to predict whether an individual is likely to use particular drugs, or at least to a better degree than random assignment.

Drug	F1	p-value	Lower CI	Upper CI	Significant
Any illegal	0.834	7.15e-12	0.553	0.677	Yes
Amphet	0.496	2.70e-07	0.106	0.326	Yes
Amyl	0.207	2.00e-02	0.000	0.222	No
Benzos	0.526	3.62e-06	0.189	0.391	Yes
Cannabis	0.761	5.19e-10	0.453	0.602	Yes
Coke	0.512	9.46e-08	0.103	0.327	Yes
Crack	0.133	5.78e-02	0.000	0.200	No
Ecstasy	0.603	7.06e-10	0.170	0.383	Yes
Heroin	0.421	3.82e-08	0.000	0.214	Yes
Ketamine	0.340	1.19e-04	0.000	0.245	Yes
LSD	0.585	4.21e-11	0.098	0.322	Yes
Meth	0.359	5.93e-04	0.058	0.286	Yes
Mushrooms	0.613	1.99e-11	0.132	0.355	Yes
VSA	0.145	7.96e-02	0.000	0.211	No
Alcohol	0.958	2.62e-03	0.907	0.947	Yes
Caff	0.985	8.09e-03	0.953	0.980	No
Choc	0.992	1.48e-02	0.966	0.990	No
Legalh	0.601	1.81e-09	0.187	0.393	Yes
Nicotine	0.694	5.48e-05	0.481	0.623	Yes

TABLE V: F1 scores and p-values from logistic regression. Significance after Bonferroni correction $\alpha = 0.263$

To visualise an example from Table V, Fig. 4 shows the null model histogram for Cannabis use. The average F1 score, and the 95% confidence interval are marked, and we can see the F1 score from the logistic regression model lies far outside this, suggesting personality traits are useful to predict an individual’s cannabis use. If the result was possible by random chance, we would expect it to lie within the bounds.

We can also use the coefficients from each logistic regression model as a way to consider how much each trait is adding to the model. One way to do this is to calculate the odds-ratio (OR) [7]. We can then interpret the OR⁶ as follows:

- OR=1: Exposure does not affect odds of outcome
- OR>1: Exposure associated with higher odds of outcome
- OR<1: Exposure associated with lower odds of outcome

For example, for the ‘any illegal drug’ class, shown in Table VI, we can see ‘Neuroticism’ has an OR of 0.91, which is close to 1. We can interpret this as being a single unit increase in neuroticism decreases the likelihood of drug use by 9%, so it does not have a great deal of influence. Compared to ‘Openness’, with an odds ratio of 2.20, indicating a unit increase increases the likelihood of drug use by 120%, suggesting this is an important factor in predicting drug usage.

⁶code adapted from: Logistic Regression in Python with statsmodels

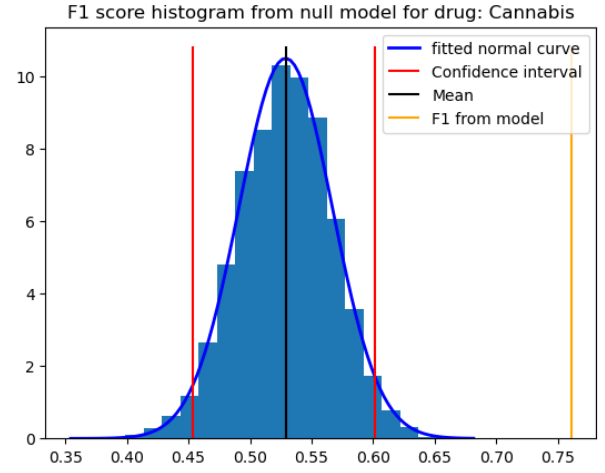


Fig. 4: Null model histogram for Cannabis

Personality Trait	OR	Lower CI	Upper CI
Neuroticism	0.91	0.78	1.07
Extraversion	0.67	0.57	0.80
Openness	2.20	1.88	2.59
Agreeableness	0.86	0.75	0.99
Conscientiousness	0.64	0.55	0.76
Impulsive	1.12	0.93	1.34
Sensation	2.40	1.97	2.92

TABLE VI: Odds Ratios for ‘any illegal drug’ usage

Fig. 5 shows a boxplot summary for the ‘usefulness’ of each trait to the logistic model, calculated from the distance from 1 of the OR for each trait.

In general, we can see that Sensation (S) is the most useful predictor, with the other traits showing some influence in most cases.

IV. DISCUSSION

The results of our analysis showcase notable patterns within the dataset. The identification of distinct personality clusters through GMM clustering provides a valuable framework for understanding how individuals with similar personality traits tend to group together. The association between these personality clusters and drug usage patterns illustrates potential links between specific combinations of personality traits and drug-related behaviour.

The logistic regression models further emphasise the influence of personality traits in predicting drug usage. Sensation emerges as a particularly influential trait, aligning with prior research suggesting its connection to risk-taking behaviours, including substance use [8].

A. Reflections on the Dataset

While the dataset offers valuable insights, we acknowledge some major limitations upon careful reflection. The imbalance in drug usage categories, as discussed earlier, poses a challenge to the generalisability of the findings. The hypotheses we fail to support are the drugs that were at the extremes of usage, with either very few or almost every participant using

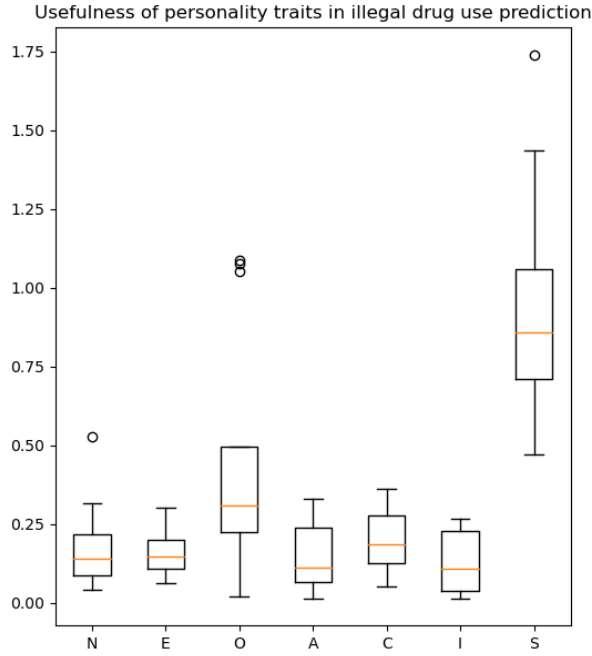


Fig. 5: Boxplot for trait usefulness for predictions

them. This may suggest that our failure to find evidence to support has more to do with the unbalanced dataset. Methods that address this problem may produce stronger results. Some sources⁷ suggest that the size of the dataset (1348 samples in the training data) is too small and that given the unbalanced data and the number of features, a dataset closer to 3000⁸ samples would be an appropriate minimum size.

Moreover, the self-reported nature of the data introduces inherent biases and subjectivity. Participants may under- or over-report their drug consumption due to a number of reasons such as societal pressure, forgetfulness, or mistrust of the researchers motivations. Despite attempts to mitigate these issues [5], the reliability of self-reported data remains a concern. The dataset relies on a single survey, which may overlook temporal variations in personality traits and drug usage. This limits our ability to establish direct causation of our findings. While we identify associations between personality traits and drug usage, understanding the direction and magnitude of influence is more complicated. Studies capturing changes over time could offer a more dynamic perspective.

Additionally, participants belong to a small number countries, and the sample of participants does not seem to be particularly diverse, which risks the resulting data being biased towards countries that are culturally similar to each other (all of the countries are primarily English-speaking countries). Hence adding participants from other different countries could have given interesting insights about personalities of people from different parts of the world and their drug consumption.

B. Reflections on Personality Traits

We show personality traits do add useful information to the problem of whether drug usage can be predicted. It would be interesting to further this by including interactions between specific traits that others have found to co-exist in drug users ([15], [16]), with the hypothesis that this would produce a stronger model.

Additionally, the potential influence of external factors, such as education level, economic conditions, or age could shape drug usage patterns independently or interact with personality traits. Acknowledging and accounting for these factors in future research may strengthen the understanding of this topic.

C. Reflections on the Statistical Modelling

Reflecting on the suitability of the dataset for our statistical modelling task, it becomes evident that while the data provides valuable insights and generates hypotheses, it might not be sufficient for predicting drug behaviour from personality traits with high precision. The connection between an individual's personality traits and drug consumption involves more complex dynamics that extend beyond the scope of this dataset. Thus, our statistical models should be considered as exploratory rather than concluding. To improve the predictions and understanding of the link between personality traits and drug-related behaviour, future studies may benefit from incorporating longitudinal datasets, and employing more sophisticated modelling such as survival analysis.

Furthermore, the Bonferroni corrections can be overly conservative and increase the likelihood of type 2 errors [9], where we fail to reject a null hypothesis when we ought to, thus might be too restrictive in this case.

V. CONCLUSION

Ultimately, the project indicates that there exists two distinct groupings of people based on the examined personality traits, and that these traits are somewhat indicative of drug use. This is supported by the ability to use these traits in predicting if an individual is likely to use any illegal drug at all ($F1 = 0.834$, $p\text{-value} = 7.15e^{-12}$).

Additionally, while predicting the use of specific illegal drugs other than Cannabis is not particularly impressive, it has been shown that using personality as a predictor is statistically significantly more accurate than random classification. So, while there may exist other contributing factors, it can be argued that personality has a significant influence on the drug habits of an individual.

These findings should encourage the understating of substance use and abuse to include an individual-centric focus, and not rely solely on socioeconomic status, ethnicity, and other such demographic attributes.

⁷<https://www.statology.org/assumptions-of-logistic-regression/>

⁸calculated as $\frac{10 \times \text{number of features}}{\text{smallest proportion}} = \frac{10 \times 7}{0.024} = 2917$

REFERENCES

- [1] Fehrman, E, et al. "Drug consumption (quantified)" UCI Machine Learning Repository, 2016, <https://doi.org/10.24432/C5TC7S>.
- [2] Lamptey, JJ. "Socio-Demographic Characteristics of Substance Abusers Admitted to a Private Specialist Clinic." *Ghana Medical Journal*, vol. 39, no. 1, 11 July 2006, <https://doi.org/10.4314/gmj.v39i1.35973>
- [3] Wallace, John M., et al. "Tobacco, Alcohol, and Illicit Drug Use: Racial and Ethnic Differences among U.S. High School Seniors, 1976-2000." *Public Health Reports* (Washington, D.C.: 1974), vol. 117 Suppl 1, no. Suppl 1, 2002, pp. S67-75
- [4] McCabe, Sean Esteban, et al. "Race/Ethnicity and Gender Differences in Drug Use and Abuse among College Students." *Journal of Ethnicity in Substance Abuse*, vol. 6, no. 2, 17 Dec. 2007, pp. 75–95, https://doi.org/10.1300/j233v06n02_06.
- [5] Fehrman, E, et al. "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk." 20 June 2015, <https://doi.org/10.48550/arxiv.1506.06297>
- [6] S. E. Bergen, C. O. Gardner, S. H. Aggen, and K. S. Kendler, "Socioeconomic Status and Social Support Following Illicit Drug Use: Causal Pathways or Common Liability?," *Twin Research and Human Genetics*, vol. 11, no. 3, pp. 266–274, Jun. 2008, doi: <https://doi.org/10.1375/twin.11.3.266>
- [7] Szumilas, Magdalena. "Explaining Odds Ratios." *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie Canadienne de Psychiatrie de l'Enfant et de L'adolescent*, vol. 19, no. 3, Aug. 2010, pp. 227–9
- [8] W. Pedersen, S.-E. . Clausen, and N. J. Lavik, "Patterns of drug use and sensation-seeking among adolescents in Norway," *Acta Psychiatrica Scandinavica*, vol. 79, no. 4, pp. 386–390, Apr. 1989, doi: <https://doi.org/10.1111/j.1600-0447.1989.tb10274.x>.
- [9] Renesh Bedre. Reneshbedre.com: "Multiple hypothesis testing problem in Bioinformatics" - <https://www.reneshbedre.com/blog/multiple-hypothesis-testing-corrections.html>
- [10] Bishop, Christopher M. "Pattern Recognition and Machine Learning". Chapter 9. New York :Springer, 2006.
- [11] Amin, Yara. "Gaussian Mixture Model". medium, 2020. - <https://medium.com/@yara.ahmed.amin/gaussian-mixture-model-4c71342b67d3>
- [12] Amy. "How To Decide The Number Of Clusters — Data Science Interview Questions And Answers". 2022. - <https://grabngoinfo.com/how-to-decide-the-number-of-clusters-data-science-interview-questions-and-answers/>
- [13] Lee, Wei-Meng. "Statistics in Python — Using Chi-Square for Feature Selection". 2021. - <https://towardsdatascience.com/statistics-in-python-using-chi-square-for-feature-selection-d44f467ca745>
- [14] DataTechNotes.com: "Clustering Example with Gaussian Mixture in Python" - <https://www.datatechnotes.com/2022/07/clustering-example-with-gaussian.html>
- [15] Turiano NA, Whiteman SD, Hampson SE, Roberts BW, Mroczek DK. "Personality and Substance Use in Midlife: Conscientiousness as a Moderator and the Effects of Trait Change". *J Res Pers*. 2012 Jun 1;46(3):295-305. doi: 10.1016/j.jrp.2012.02.009. PMID: 22773867; PMCID: PMC3388488.
- [16] Sherry H. Stewart, Heather Devine, "Relations between personality and drinking motives in young adults, Personality and Individual Differences", Volume 29, Issue 3, 2000, Pages 495-511, ISSN 0191-8869, [https://doi.org/10.1016/S0191-8869\(99\)00210-X](https://doi.org/10.1016/S0191-8869(99)00210-X). (<https://www.sciencedirect.com/science/article/pii/S019188699900210X>)

APPENDIX A

Drug	Users	Non-users	User%
Choc	1840	45	97.61
Caff	1824	61	96.76
Alcohol	1749	136	92.79
<i>Any Illegal</i>	<i>1174</i>	<i>711</i>	<i>62.28</i>
Nicotine	1060	825	56.23
Cannabis	999	886	53.00
Legalh	564	1321	29.92
Benzos	535	1350	28.38
Ecstasy	517	1368	27.43
Amphet	436	1449	23.13
Mushrooms	434	1451	23.02
Coke	417	1468	22.12
LSD	380	1505	20.16
Meth	320	1565	16.98
Ketamine	208	1677	11.03
Amyl	133	1752	7.06
Heroin	118	1767	6.26
VSA	95	1790	5.04
Crack	79	1806	4.19

TABLE VII: User proportions in whole dataset, CL0-CL2 are designated as non-users. Blue highlighted drugs indicate those we were able to support our hypthoses in the Logistic Regression section II-D1

APPENDIX B

Drug	Cluster	CL0 %	CL1 %	CL2 %	CL3 %	CL4 %	CL5 %	CL6 %
Amphet	0	36.5	10.9	16.0	15.8	6.7	5.6	8.4
	1	68.2	13.6	10.0	4.7	0.9	0.7	2.2
Amyl	0	62.9	11.3	15.2	7.0	1.9	1.3	0.3
	1	76.1	10.9	9.7	2.6	0.6	0.1	0.0
Benzos	0	39.9	5.5	13.6	17.3	9.3	7.2	7.3
	1	67.3	6.8	11.1	7.4	3.2	1.5	2.6
Cannabis	0	8.7	5.7	11.8	13.8	10.7	14.4	34.8
	1	36.2	16.6	16.5	8.4	3.9	4.9	13.6
Coke	0	40.5	8.3	17.8	19.9	8.3	3.7	1.5
	1	70.8	8.7	10.6	6.9	2.0	0.5	0.4
Crack	0	79.1	4.7	9.0	5.3	0.8	0.8	0.2
	1	94.0	2.3	2.6	0.7	0.1	0.1	0.0
Ecstasy	0	37.6	5.9	14.5	22.9	11.8	5.6	1.6
	1	72.0	6.1	10.1	5.8	4.5	0.8	0.5
Heroin	0	77.5	4.2	8.0	5.3	2.2	1.5	1.2
	1	93.4	3.0	1.8	1.4	0.2	0.1	0.1
Ketamine	0	70.1	2.7	9.5	10.6	3.6	3.2	0.3
	1	88.6	2.1	5.4	2.7	0.7	0.2	0.1
LSD	0	40.3	15.3	13.2	16.9	8.5	4.5	1.2
	1	74.4	12.0	5.3	5.3	1.5	1.3	0.1
Meth	0	65.3	2.0	7.1	11.8	4.0	4.2	5.6
	1	87.1	2.1	3.1	3.7	1.2	0.7	2.0
Mushrooms	0	35.3	11.3	18.5	21.7	9.4	3.5	0.3
	1	70.2	10.8	8.7	6.9	2.5	0.6	0.1
VSA	0	67.3	12.7	11.7	5.3	1.2	1.3	0.5
	1	87.9	8.4	2.3	0.1	0.1	0.1	0.2
Alcohol	0	1.2	0.7	4.1	11.6	16.0	37.7	28.7
	1	2.4	3.0	3.1	9.4	14.4	43.0	24.7
Caffeine	0	0.8	0	0.9	3.4	5.9	15.7	73.2
	1	2.1	1.1	1.7	3.0	5.3	13.1	73.8
Chocolate	0	1.6	0.1	0.6	3.1	18.4	35.3	40.9
	1	1.8	0.2	0.4	2.6	12.8	37.3	44.9
LegalH	0	40.0	1.8	12.4	25.2	9.3	5.5	5.6
	1	77.4	1.2	8.4	8.5	2.1	1.1	1.3
Nicotine	0	13.4	0.61	8.6	11.8	6.9	10.8	41.9
	1	32.2	14.7	13.2	7.7	4.5	5.6	22.1

TABLE VIII: Drug Consumption Patterns for each drug within the different clusters. All numbers above 50.0% are highlighted.

APPENDIX C

Contribution Statement:

Amna, Zainab: Gaussian Mixture Methods and Results

Chrisanna, Christian: Logistic Regression Methods and Results

Shared responsibility: Introduction, EDA, Discussion, Conclusion