Understanding Drug Consumption with Machine Learning: Personality and Demographic Perspectives

By: Tiffany Hugh

Dataset Description, Selection Rationale, and Predictive Objectives

Investigating the connection between drug consumption and personality traits provides a fascinating glimpse into the nuances of human behavior. This is why the dataset titled Drug Consumption (Quantified) from the UC Irvine Machine Learning Repository was chosen. This comprehensive database comprises records for 1,885 respondents, each characterized by 12 distinct attributes.

Among the attributes are the Big Five personality traits, which encapsulate key dimensions of human behavior and emotion:

- Neuroticism (N): Describes a long-term tendency to experience negative emotions such as nervousness, tension, anxiety, and depression.
- Extraversion (E): Manifested in characteristics such as being outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation.
- Openness to Experience (O): Represents a general appreciation for art and unusual ideas, including being imaginative, creative, unconventional, and having wide interests.
- Agreeableness (A): Characterizes interpersonal relations and is marked by altruism, trust, modesty, kindness, compassion, and cooperativeness.
- Conscientiousness (C): Denotes a tendency to be organized and dependable, with traits such as strong-willed persistence, reliability, and efficiency.[1]

Additionally, the dataset includes BIS-11 (impulsivity) and ImpSS (sensation seeking) scores, as well as demographic details such as level of education, age, gender, country of residence, and ethnicity.

Respondents provided information regarding their usage of 18 different substances, both legal and illegal, including alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine, and a fictitious drug (Semeron). The frequency of usage is categorized as follows: "Never Used," "Used over a Decade Ago," "Used in Last Decade," "Used in Last Year," "Used in Last Month," "Used in Last Week," and "Used in Last Day."

To understand these nuances, the following questions are posed:

1. Is it possible to forecast the likelihood of an individual's usage or non-usage of a particular substance by leveraging their personality traits and demographic data?
2. Moreover, can clustering respondents into discernible cohorts based on their personality characteristics and demographic profiles facilitate the discernment of underlying patterns in drug consumption behaviors?

[1] E. Fehrman, A.K. Muhammad, E.M. Mirkes, V. Egan, and A.N. Gorban, "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk," June 20, 2015, Cornell University.

Data Cleaning and Transformation

There were no missing values in the dataset so all alterations were made to clearly define target and feature variables while making the data more legible on an ordinal scale. The first alteration made to the data was the creation of a binary classification to identify whether a respondent was a drug consumer or not. Based on the frequency of usage, respondents were categorized as non-consumers if they "Never Used," "Used over a Decade Ago," "Used in Last Decade," or "Used in Last Year." Those who "Used in Last Month," "Used in Last Week," and "Used in Last Day" were classified as drug consumers. The challenge was determining the threshold for what constitutes ongoing drug consumption. I decided that usage within the past year indicates a non-consumer, focusing on individuals actively using drugs rather than those who have merely experimented in the past.

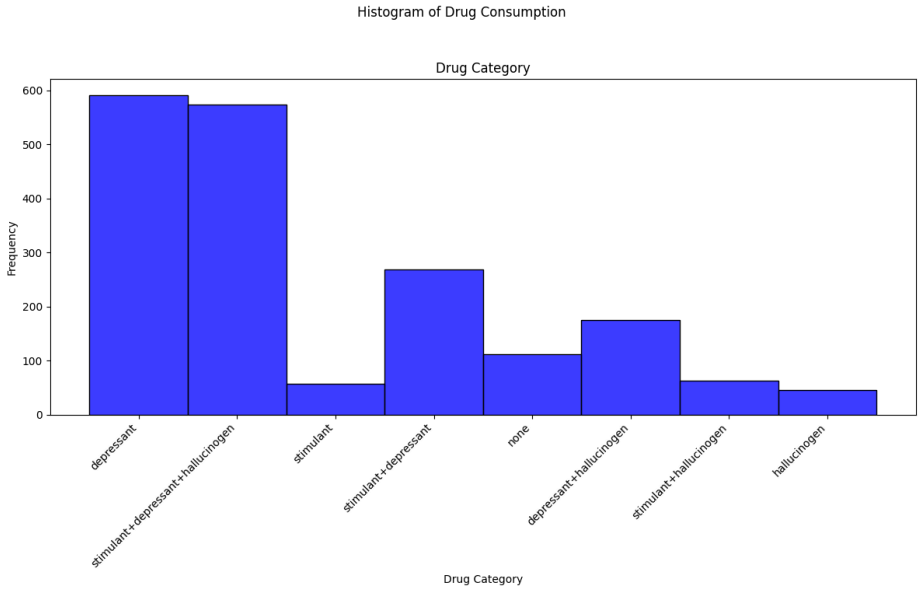
Given the specific focus on particular substances and the dataset containing 18 substances, the second alteration involved target selection. I categorized the drugs into three groups:

- Stimulants: caffeine, nicotine, amphetamines, cocaine, crack, meth.
- Depressants: alcohol, benzodiazepines, heroin.
- Hallucinogens: LSD, mushrooms, ketamine, ecstasy, cannabis.

A challenge that arose was that respondents could have used multiple drugs, making the data multicategorical. This complicates the process when running precision, recall, and F1 tests.

For the demographic data, I re-coded values assigned on an ordinal scale (0, 1, 2) to be more legible. Through exploratory analysis, I determined that the demographic variables of country and ethnicity were not representative of the sample due to unequal distribution of respondents. Therefore, I focused on age, gender, and level of education for further analysis.

Data Visualization



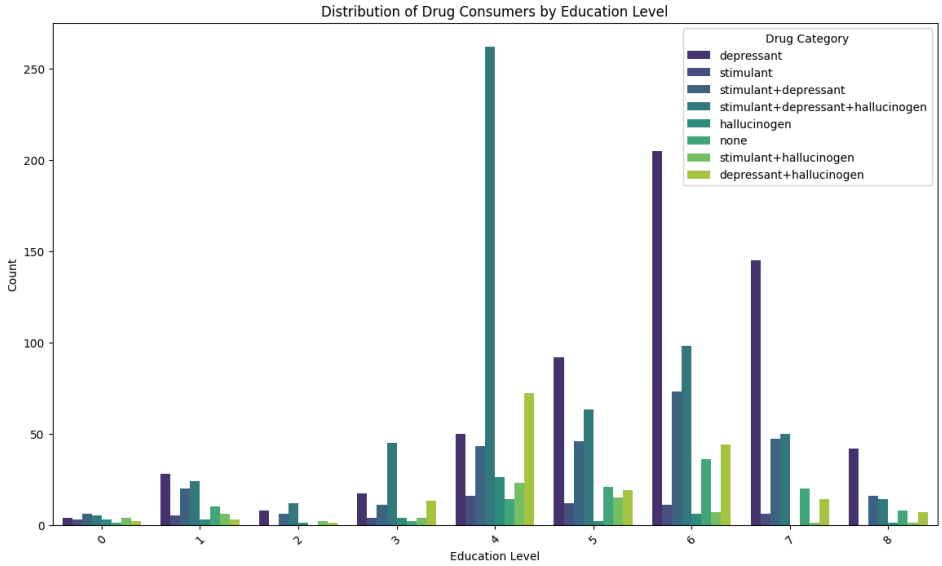
Histogram 1: Drug Category Frequency

This histogram shows the frequency distribution of drug categories, revealing insights into drug consumption patterns among the respondents. A significant portion of the dataset, 591 individuals, consumes depressant drugs, indicating a high prevalence of substances like alcohol, benzodiazepines, and heroin, which can be attributed to the legality and wide accessibility of alcohol. The second largest group, comprising 573 individuals, engages in polydrug use, combining stimulants, depressants, and hallucinogens, suggesting complex patterns of drug dependence and higher risks of adverse effects. Additionally, 268 individuals consume both stimulants and depressants, while 175 individuals use depressants and hallucinogens. The non-consumers comprise 112 individuals, followed by those who use both stimulants and hallucinogens (63 individuals), only stimulants (57 individuals), and only hallucinogens (46 individuals). This distribution underscores the diversity and complexity of drug use behaviors.

Gender	Stimulant + Depressant + Hallucinogen	Depressant	Stimulant + Depressant	Depressant+ Hallucinogen	None	Stimulant	Hallucinogen	Stimulant + Hallucinogen
Male(0)	384	197	105	107	35	34	31	50
Female(1)	189	394	163	68	77	23	15	13

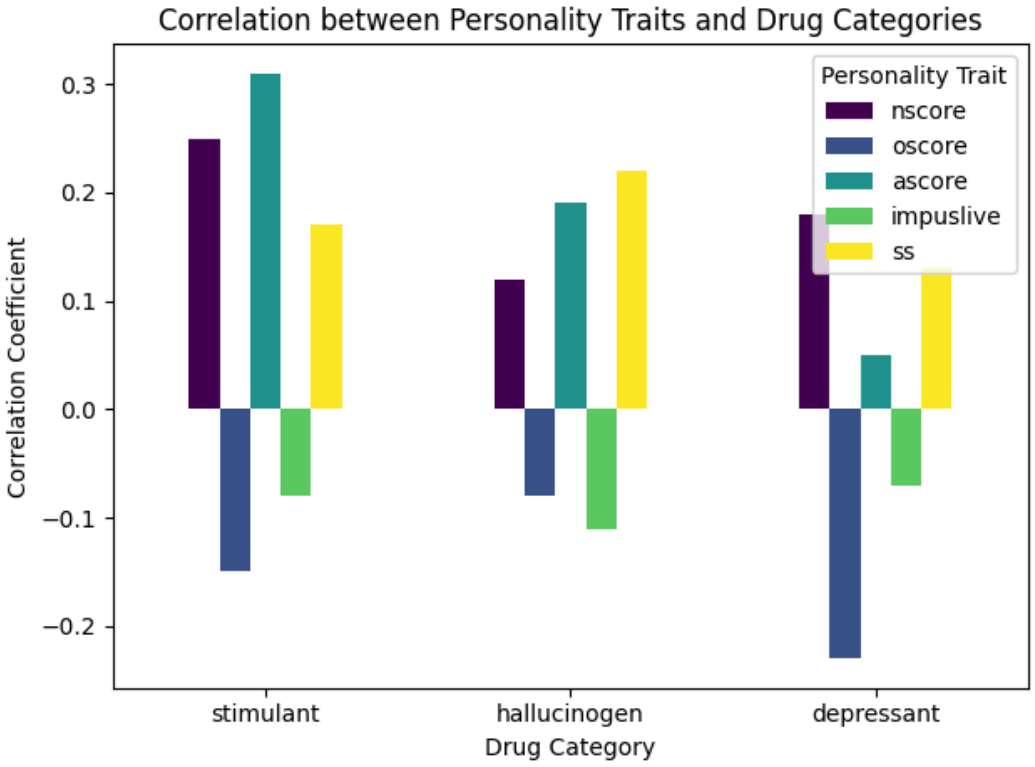
Table 1: Distribution of Drug Consumers by Gender

The table illustrates the distribution of drug categories among respondents categorized by gender, where '0' represents males and '1' denotes females. Among males, 394 individuals are reported as consuming depressants, while 197 females fall into this category. Females show a higher prevalence of stimulant and depressant combinations, with 163 individuals compared to 105 among males. Both genders also exhibit instances of non-consumption, with 77 females and 35 males abstaining from drug use. These findings highlight distinct gender-specific patterns in drug consumption behaviors.



Countplot 1: Distribution of Drug Consumption by Education Level (0 - Left school before 16 years, 1 - Left school at 16 years, 2 - Left school at 17 years, 3 - Left school at 18 years, 4 - Some college or university, no certificate or degree, 5 - Professional certificate/diploma, 6 - University degree, 7 - Masters degree, 8 - Doctorate degree.)

The results highlight distinct patterns of drug category distribution across different education levels. Individuals with lower educational attainment, such as those who left school before 16 years or at 16 years (levels 0 and 1), generally show lower counts across most drug categories compared to those with higher education levels. For instance, at the highest educational levels (6, 7, and 8—representing university degrees and beyond), there's a notable increase in the prevalence of depressant and stimulant use, as well as polydrug use involving combinations of depressants, stimulants, and hallucinogens. This suggests a correlation between higher educational achievement and more complex patterns of drug consumption. This could be attributed to the fact that in higher education people are exposed to more substances, or even the pressure of achievement causes drug usage. The presence of non-consumers is also observed across all education levels, with varying frequencies. These findings underscore the influence of educational attainment on drug consumption behaviors.



Plot 2: Correlation Between Personality Traits and Drug Category

The plot shows in general that neuroticism, agreeableness, and sensation seeking have moderate to positive correlations across the drug categories whereas openness personality and impulsivity have slight negative correlations. For stimulant consumers, higher levels of neuroticism (r =

0.25) and agreeableness (r = 0.31) are positively correlated, indicating that individuals scoring higher on these traits are more likely to consume stimulant drugs. Conversely, openness to experience (r = -0.15) shows a negative correlation, suggesting that those with lower openness scores may lean towards stimulant use. Hallucinogen consumption shows a mild positive correlation with neuroticism (r = 0.12) and agreeableness (r = 0.19), implying a nuanced association where these traits might influence experimentation with hallucinogens. In contrast, depressant users exhibit a stronger negative correlation with openness to experience (r = -0.23) and conscientiousness (impulsiveness, r = -0.07), hinting that individuals lower in these traits may be more inclined towards depressant substances. Overall, these findings underscore the complex interplay between personality dimensions and drug consumption patterns.

Supervised Machine Learning

Supervised Learning is when a machine learning model is trained on a labeled dataset, where the features in this dataset include the five personality traits, age, gender, education level, impulsivity, and attention-seeking behavior. The desirable outcome, known as the label, is the drug category.

The first step in using machine learning is splitting the data into the training and testing sets. The drug category target distribution results reveal a closely aligned pattern between the two sets, indicating a well-stratified sampling. In the training set, the predominant category is "stimulant+depressant," comprising 43.44% of the data, closely mirrored in the test set at 43.50%. Similarly, "stimulant+depressant+hallucinogen" follows with 38.79% in the training set and 38.73% in the test set. The other categories also show minimal discrepancies between the two sets: "stimulant" (8.02% train, 7.96% test), "stimulant+hallucinogen" (5.31% train and test), "depressant" (2.12% train and test), "none" (0.93% train, 1.06% test), "depressant+hallucinogen" (0.93% train, 0.80% test), and "hallucinogen" (0.46% train, 0.53% test). These consistent distributions suggest that the data was split effectively, maintaining the proportion of each category and thus ensuring the representativeness and reliability of the test set in evaluating model performance.

Supervised Machine Learning: Random Forest Classifier

Random Forest Classifier is a form of decision tree where multiple decision trees are trained on different subsets of the data and their outputs are combined to improve overall performance and robustness. Since the dataset was re-coded as multicategorical and random forest is best equipped to run with categorical variables and interactions between features it was selected. Another advantage would be that this algorithm can capture non-linear relationships and perform well without much hyperparameter tuning and is less sensitive to outliers.[2]

[2] Sarah Guido and Andreas C. Muller, *Introduction to Machine Learning with Python* (Sebastopol, CA: O'Reilly Media, Inc., 2016), 85-89.

With default parameters the Random Forest Classifier results provide valuable insights into its effectiveness in predicting drug consumption behaviors based on personality traits and demographic data. With a precision of 0.585, the model demonstrates that when predicting whether an individual consumes a particular substance, it is accurate approximately 58.5% of the time on the test set. The recall of 0.674 indicates that the model successfully identifies about 67.4% of all instances of drug consumption. The F1 score, which harmonizes precision and recall, stands at 0.614, reflecting a balanced performance overall. These metrics collectively underscore the model's ability to leverage personality traits and demographic profiles to forecast drug usage likelihood.

To optimize the performance of the model, GridSearchCV was utilized with a parameter grid consisting of: 'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10], and 'min_samples_leaf': [1, 2, 4]. The worst-performing parameters were identified as max_depth of 20, min_samples_leaf of 1, min_samples_split of 2, and n_estimators of 100, resulting in a lower F1 score of 0.570. This suggests that increasing the tree depth and the number of estimators beyond the optimal values may lead to reduced model performance on this dataset due to overfitting. Conversely, the best parameters were a max_depth of 10, min_samples_leaf of 4, min_samples_split of 2, and n_estimators of 50. These parameters yielded the highest F1 score of 0.585 on the validation set, indicating a good balance between precision and recall for the model's predictions.

The model shows promising performance in predicting the target labels based on the test set. The precision of 0.631 indicates that when the model predicts a positive label, it is correct approximately 63.1% of the time. The recall of 0.671 suggests that the model correctly identifies about 67.1% of all actual positive instances. The F1 score, which harmonizes precision and recall, is 0.609, reflecting an overall balanced measure of the model's accuracy in predicting both positive and negative labels.

Reflecting on the reasons for differences in performance between models and parameter settings, it is evident that deeper trees (higher max_depth) and a higher number of estimators (n_estimators) can lead to overfitting, which negatively impacts the model's ability to generalize to new data. On the other hand, limiting the max_depth and the number of estimators while increasing min_samples_leaf helps in reducing overfitting and improves the model's generalization performance.

The adjustments made through GridSearchCV with the best-performing parameters improved the model's precision, recall, and F1 score compared to the default settings. This indicates that fine-tuning hyperparameters is crucial for achieving better performance in predicting drug consumption behaviors based on personality traits and demographic data.

Supervised Machine Learning: Support Vector Machine (SVM)
The second algorithm used was a support vector machine which is well-suited for binary classification tasks and complex relationships through the use of kernel functions. In this case, a MultiOutputClassifier with SVM was used to accommodate the multi-label nature of the target variable. Different kernels—radial basis function (rbf), linear, and polynomial—were evaluated to determine the best performance.

In the context of predicting the likelihood of an individual's substance usage based on their personality traits and demographic data. The rbf kernel showed the best performance, achieving a precision of 0.561, a recall of 0.682, and an F1 score of 0.615 on the test set. These metrics indicate balanced performance in identifying true positives while minimizing false positives.

Conversely, the linear kernel SVM yielded slightly lower metrics with a precision of 0.554, recall of 0.674, and an F1 score of 0.608. The polynomial kernel SVM performed the worst among the kernels tested, with a precision of 0.527, recall of 0.637, and an F1 score of 0.575.

Reflecting on the reasons for performance differences, the rbf kernel's ability to map data into higher-dimensional spaces efficiently and capture complex relationships between features likely contributed to its superior performance. In contrast, the linear kernel, while simpler and computationally efficient, may struggle with datasets that are not linearly separable, leading to slightly lower performance metrics. The polynomial kernel, while capable of handling non-linear relationships, may have been less suitable for this specific dataset due to overfitting or underfitting concerns.

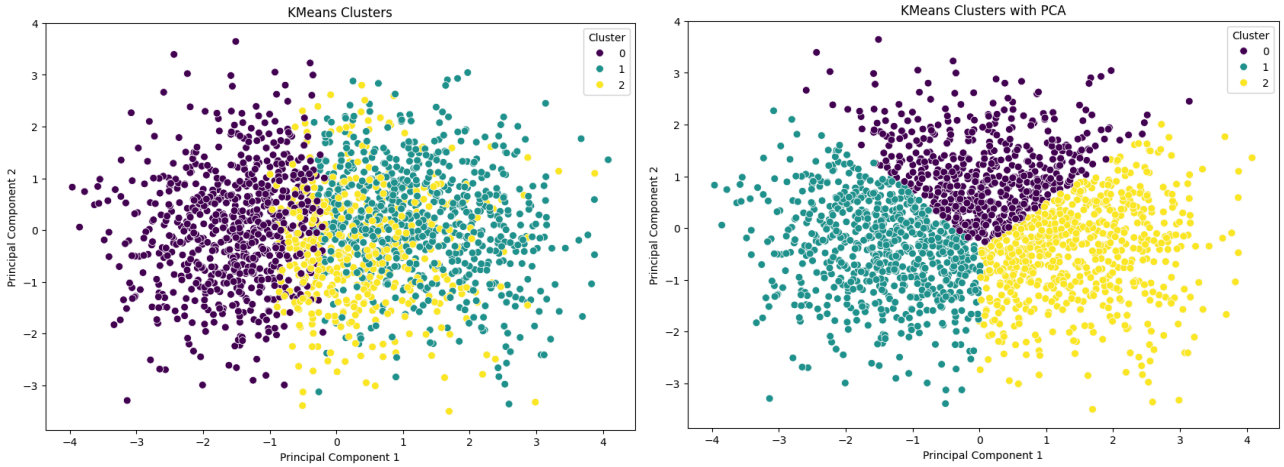
Additionally, the approach of using a reduced parameter grid focusing on C, gamma, and kernel values ([0.1, 1] for C, [1, 0.1] for gamma, and 'linear' and 'rbf' for kernel) aimed to balance model complexity with computational feasibility. This grid allowed practical exploration of the SVM's effectiveness without overwhelming computational resources, leading to reasonable predictive capability across multiple classes of substance usage. Overall, SVMs, particularly with the rbf kernel, show promise in leveraging personality traits and demographic data to forecast substance usage patterns.

Unsupervised Machine Learning
Unsupervised Learning is using the training data that is unlabeled, and the system tries to learn patterns and make predictions without a teacher. Principal Component Analysis (PCA) is a technique for feature selection that would result in the best variance of the dataset through standardization. PCA was applied on the best-performing supervised learning algorithm, the Random Forest Classifier, which had previously achieved a best F1 score of 0.4083. Initially, the dataset was standardized before fitting PCA to identify the number of components required to retain 95% of the variance. This resulted in a significant reduction in dimensionality. Using the PCA-transformed features, the Random Forest model was trained with the best hyperparameters identified through GridSearchCV (max_depth=10, min_samples_leaf=1, min_samples_split=10,

n_estimators=50). The model trained on the PCA-transformed data achieved a testing F1 score of 0.6054, reflecting an improvement of 0.1971 over the previous best score. Although the overall accuracy and F1 scores improved, a detailed classification report revealed that the model still struggled with certain classes, likely due to class imbalance. However, the improvement in the F1 score demonstrates that PCA effectively enhanced the model's performance, especially for the more frequent classes, thereby validating the utility of PCA for feature selection in this context.

PCA: KMeans

KMeans is an unsupervised machine learning algorithm designed for clustering. Clusters are formed by the data points features, the primary objective of KMeans is to reduce the variance within each cluster while increasing the variance between different clusters.[3]



Scatterplot 1&2: KMeans Without and With PCA

Looking at the scatterplot, it is apparent that without PCA, the clusters are mixed together, whereas with PCA, there is a clear division between the three types of clusters. KMeans without PCA produced clusters with an Adjusted Rand Index (ARI) of 0.1576 and a Silhouette Coefficient of 0.1499. The ARI reflects how well the clusters align with the true drug consumption behaviors, while the Silhouette Coefficient indicates the cohesion and separation of these clusters. These metrics suggest that while some structure is present, the clusters may not be very distinct or well-separated.

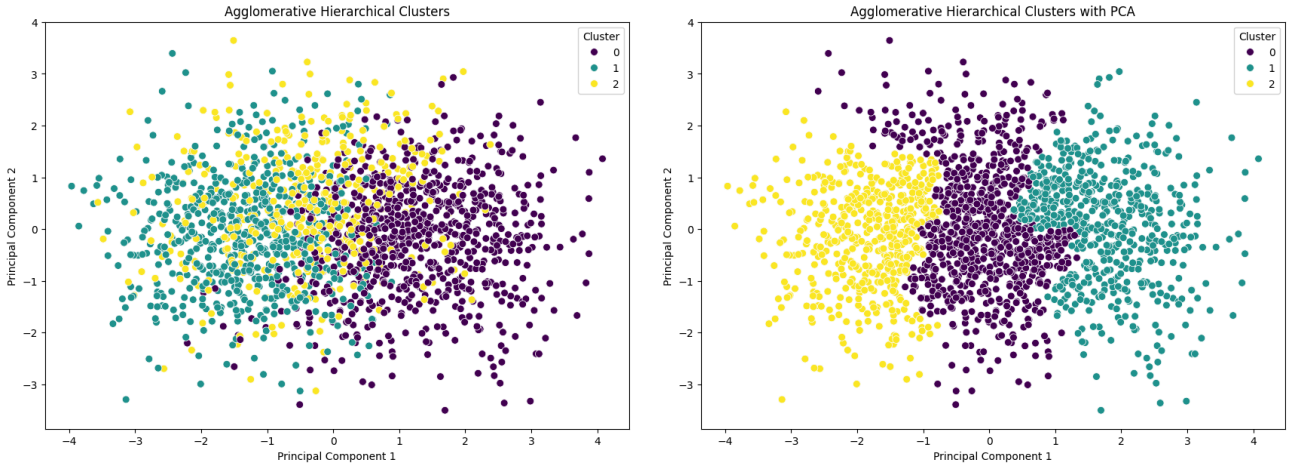
When PCA was applied before clustering, the ARI slightly decreased to 0.1331, indicating a minor drop in the alignment with true labels. However, the Silhouette Coefficient significantly increased to 0.3320, showing that the clusters became more compact and distinct in the lower-dimensional space. This improvement in the Silhouette Coefficient suggests that PCA helps in creating clearer and more defined clusters, even if the alignment with the true labels is slightly compromised.

[3] Guido and Muller, *Introduction to Machine Learning with Python*, 170-173.

Thus, clustering respondents into discernible cohorts using PCA can indeed facilitate the discernment of underlying patterns in drug consumption behaviors, making it a valuable preprocessing step in such analyses.

PCA: Agglomerative/Hierarchical

Agglomerative clustering creates hierarchical clustering by merging similar pairs of clusters, starting with each data point as its own cluster, and continuing until all data points are grouped into a single cluster or a specified number of clusters is achieved.[4]



Scatterplot 3&4: Agglomerative Hierarchical Without and With PCA

Examining the scatterplot without PCA, the clusters are less organized and appear scrambled together, whereas with PCA, there is a clear division and organization of the clusters. For agglomerative clustering without PCA, the Adjusted Rand Index (ARI) is 0.1429 and the Silhouette Coefficient is 0.1231. These metrics indicate that the clusters have some structure but are not very distinct or well-separated. When PCA is applied before agglomerative clustering, the ARI slightly decreases to 0.1159, showing a minor reduction in alignment with the true labels. However, the Silhouette Coefficient significantly increases to 0.2576, indicating that the clusters become more compact and distinct in the lower-dimensional space.

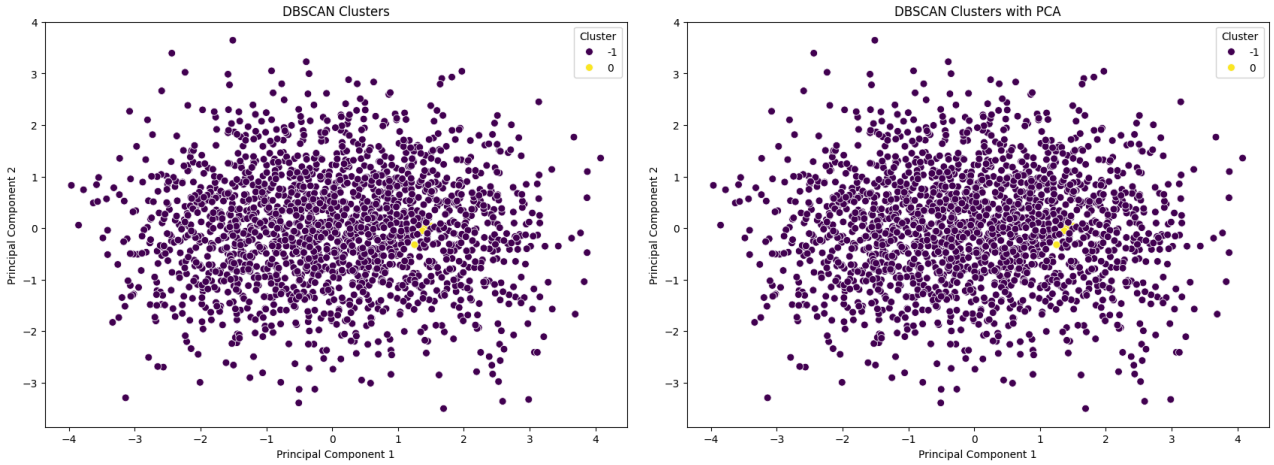
This improvement in the Silhouette Coefficient with PCA suggests that dimensionality reduction helps create clearer and more defined clusters, even if the alignment with true labels is slightly compromised. This ties into the question of whether clustering respondents into discernible cohorts based on their personality characteristics and demographic profiles can facilitate the discernment of underlying patterns in drug consumption behaviors. By enhancing the clarity and definition of clusters, PCA enables the clustering algorithm to from more meaningful and

[4]Guido and Muller, *Introduction to Machine Learning with Python*, 186-188.

interpretable cohorts, thereby making it easier to identify and understand the underlying patterns in drug consumption behaviors.

PCA: Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise(DBSCAN) is another clustering technique; however, instead of partitioning data into a predefined number of clusters, DBSCAN identifies clusters based on the density of data points in the feature space.[5]



Scatterplot 5&6: DBSCAN Without and With PCA

There is no discernible difference between the scatterplots; however, looking at DBSCAN without PCA, the Adjusted Rand Index (ARI) is -3.688024331725701e-05, and the Silhouette Coefficient is -0.1610343682919878. This suggests that the clustering is poor, with clusters being neither well-defined nor cohesive. In contrast, DBSCAN with PCA shows a noticeable improvement: the ARI increases to 0.004372581701798384, and the Silhouette Coefficient significantly improves to 0.4386995767693842. These metrics indicate that applying PCA before DBSCAN results in more coherent and distinct clusters.But when plotted it does not show that difference. Maybe this was due to the data points being multilabel causing the clustering to be different compared to the other methods.

Summary

This assignment was the first time I used Python, and I have learned a myriad of information. During the first project, I had the chance to import the dataset and recode it so the data would be more legible. In some ways, it was similar to R Studio. However, if I had a better understanding of supervised and unsupervised techniques, I would have chosen a different way to recode the target values. Creating categories for stimulants, depressants, and hallucinogens caused the data to become multicategorical, which led to difficulties along the way. In the future, I could classify the drugs based on legality or their modular structure to create a clearer division between each target category.

---

[5] Guido and Muller, *Introduction to Machine Learning with Python*, 189.

In the second project, the major issues arose because the target variables were multicategorical. The error I frequently encountered was that multiclass-multioutput is not supported, making it challenging to run precision, recall, and F1-score using a Random Forest classifier. I was able to use the MultiOutputClassifier, but I had to aggregate the target values. This also caused the .fit test to run extremely slowly, given that my dataset is not very large. Perhaps Random Forest Classifier was not the best technique, but it is said to be the best technique for multi-label data.

Lastly, in the third project, complications arose because the clusters only aligned on -1 and 1 as a straight line. This issue could be due to various reasons: incorrect parameters, scaling issues, or the data distribution not having clear clusters. I had to preprocess the data using StandardScaler and create separate data frames for visualizations. Overall, the biggest change I would make in this project would be the classification of the target values. Not making them multicategorical would have made this project smoother, especially considering it was my first time working on such an assignment.

Bibliography

Fehrman, E., A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk." arXiv, 2017. https://arxiv.org/abs/1506.06297

Müller, Andreas C., and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Incorporated, 2018.