Kaden N. Nguyen

DS 3001

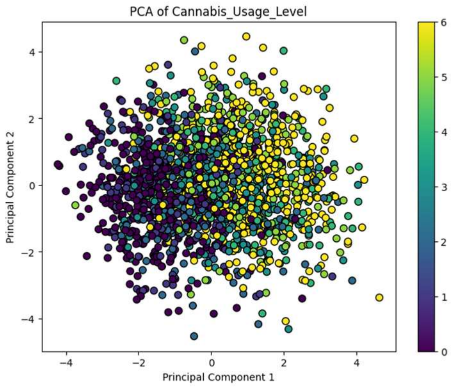
7 May 2025

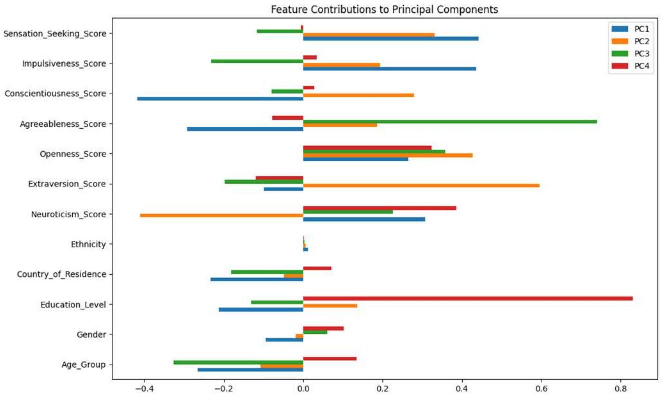
Code: https://github.com/kadennguyen0329/DS-3001-Final

Introduction

This report investigates the prediction of cannabis usage levels using the "Drug Consumption (Quantified)" dataset from the UCI Machine Learning Repository. The dataset is composed of self reported data from 1,885 respondents, including 12 features: demographic variables (age, gender, country, ethnicity, education) and personality traits (neuroticism, extraversion, openness, agreeableness, conscientiousness, impulsivity, sensation seeking). The original dataset includes 18 target variables representing drug use frequencies for 18 various substances, with usage categorized on a 7 point scale from never used (CL0) to used in the last day (CL6). The chosen target variable to classify was cannabis, as cannabis legalization expands worldwide. Given the ordinal nature of the target variable, the problem was formulated as a classification task.

Among the exploratory data analysis techniques, including feature heatmaps, scatter plots, and statistical analysis of individual features, principal component analysis yielded critical insights: visualization of the first two principal components revealed significant class overlap, particularly between adjacent usage levels. While extreme categories (e.g. CL0 versus CL6) formed distinct clusters, intermediate usage levels showed substantial overlap. Furthermore, analysis of feature contributions to the first four principal components demonstrated that all predictors influenced classification outcomes, with varying magnitudes of positive/negative impact.





Data Pipeline

The feature variables within the original dataset were already well preprocessed upon acquisition. Categorical demographic features were encoded, standardized, and normalized, while numeric personality scores underwent standardization and normalization to ensure comparability across features in model training. The standardization and normalization was kept since it is critical for distance based models like K Nearest Neighbors to prevent features with larger scales from dominating. No missing values were present, simplifying preprocessing by eliminating the need for imputation. Furthermore, to improve data quality, rows corresponding to individuals who reported using Semeron, a fictitious drug included in the survey to detect over claimers, were removed.

The target variable was originally labeled 'CL0 CL6' as strings, so label encoding was used to cast the classes into integer representation. The target variable's original scale was transformed into three broader classes: nonusers (level 0), light users (levels 1–3, representing use over a decade ago to within the last year), and heavy users (levels 4–6, representing use within the last month to last day). This was done to address class imbalance and, more importantly, the overlap observed in EDA. PCA visualizations showed distinct clusters for extreme cannabis usage groups but significant overlap between adjacent classes. This motivated the grouping strategy and justified retaining all features for modeling, as feature contribution analysis showed that all variables meaningfully influenced the principal components.

Model Overview

Three models were developed and evaluated: two supervised learning models, K Nearest Neighbors and Random Forest, and one unsupervised model, PCA. F1 was chosen as the metric for evaluating supervised models due to its robustness to class imbalance and consideration of both precision and recall. Hyperparameter tuning for the supervised models followed an iterative two stage approach. First, parameters from scikit learn's documentation were tested individually over narrow ranges to assess their impact on the F1 score. Parameters causing minimal F1 score changes were excluded from the next step, combined parameter testing. After the parameters were chosen, grid searches were used to optimize combinations of the influential parameters.

The KNN model was selected first due to the clustering patterns observed in PCA, which suggested that proximity based classification would be effective. The number of neighbors, k, was tuned over the range 1 to 30. The optimal value was found to be k=18, yielding the highest F1 score of 61.97% compared to the default model with 57.91%. Other parameters such as weighting schemes and distance metrics (e.g. Manhattan vs. Euclidean) were tested but did not improve performance, so only the number of neighbors were included in parameter tuning.

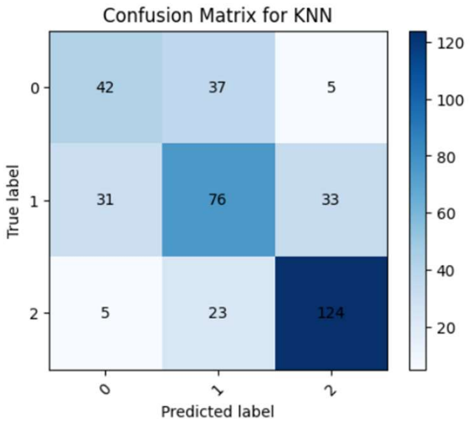
Random Forest was chosen as the second supervised model for its ability to consider feature importances, which proved valuable during EDA. The hyperparameters selected for tuning included the splitting criterion, maximum tree depth, and number of estimators, as each demonstrated a meaningful impact on the F1 score during individual parameter testing. The default Random Forest configuration yielded an F1 score of 59.10%. However, subsequent grid search optimization resulted in a slightly worse F1 score of 59.07%, with the best parameters identified as a Gini splitting criterion, a maximum tree depth of 11, and 50 estimators.

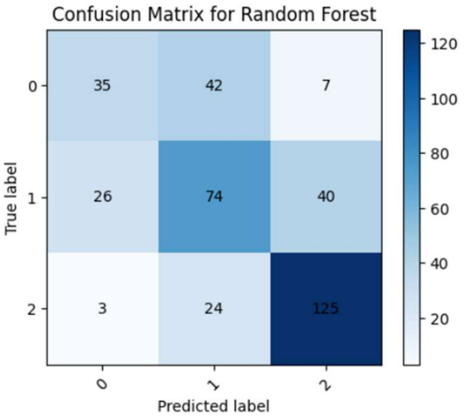
For unsupervised analysis, PCA was applied again, but with the cannabis usage target grouped into the three classes as opposed to 6 classes in the original dataset. Six principal components explained over 80% of the variance, indicating a complex feature space. Clusters became more distinct between nonusers and heavy users, although overlap persisted between adjacent groups. Lastly, feature weight analysis identified impulsiveness, sensation seeking, and openness as the most influential traits contributing to the first four principal components.

Discussion

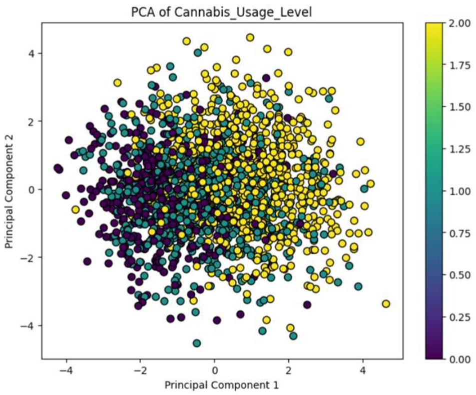
Analysis

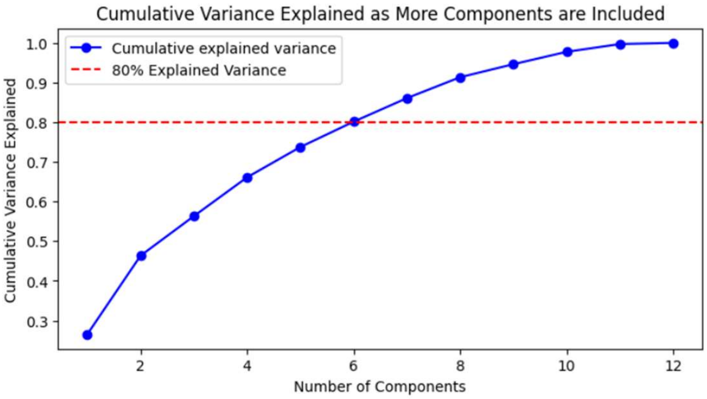
The KNN model outperformed Random Forest, achieving an F1 score of 61.97% compared to Random Forest’s best score of 59.10%. This likely reflects KNN’s advantage in leveraging the natural clustering structure of the data. Both models, however, struggled to accurately classify individuals in adjacent usage categories, confirmed with further evaluation using confusion matrices. These matrices revealed that most misclassifications, for both KNN and Random Forest were one level off, confirming the challenge posed by class overlap. However, very few misclassifications occurred with identifying level 0 as level 2, or vice versa, which supports the fact that distinguishing between extreme usage levels is easier.

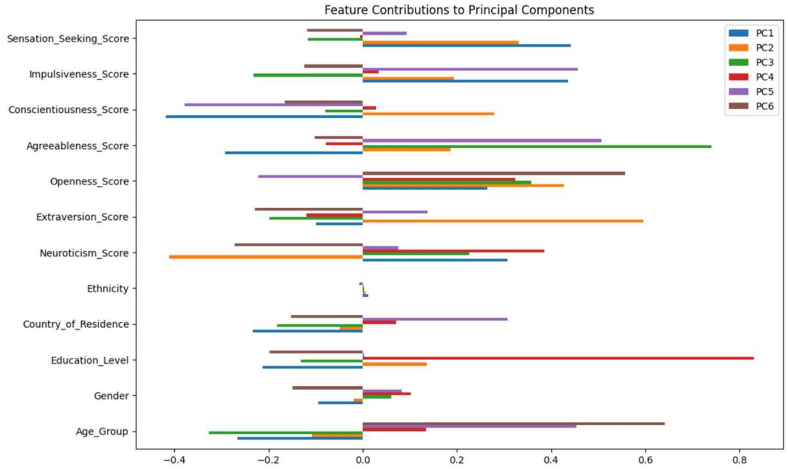




The unsupervised PCA results supported these findings by demonstrating clear separability only between extreme usage groups, with residual overlap among adjacent the classes. Feature weight analysis of the first 6 PC's, since around 80% of the data's variance can be contributed to the first 6 PCs, highlighted that all features contributed meaningfully to classification, indicating no single dominant predictor.







Ethical Integration

The development and deployment of predictive models for cannabis usage classification, as explored in this report, present significant ethical challenges that deal with with data privacy, algorithmic bias, and social implications. First, the inclusion of ethnicity as a feature could reinforce harmful stereotypes if applied in contexts such as law enforcement or employment screening. Historical biases in drug related arrest rates, which disproportionately target marginalized communities, could be amplified if models trained on data including ethnicity are used to profile individuals (Rosenberg). This underscores the need to critically evaluate whether demographic features like ethnicity should be excluded from real world uses to prevent algorithmic discrimination.

Second, the reliance on self-reported data introduces privacy vulnerabilities. Predictive models extract sensitive behavioral patterns, creating risks of reidentification from unique profiles even with anonymized datasets. The mere act of generating predictions about individuals, such as classifying cannabis use levels, could enable unequal treatment of individuals based on inferred rather than observed behaviors. This is particularly concerning given the legal and social stigma still associated with cannabis in many regions (King). These unwarranted predictive analytics could expose reported anonymous users to punitive measures, such as employment discrimination, solely based on algorithmic inferences (King).

The model's emphasis on personality traits like impulsivity and sensation seeking also raises ethical questions. While these traits contribute to evidence-based risk factors for substance use, their use in predictive tools could stigmatize individuals flagged as high risk form such an algorithm based on personality (Delibas). This could put an emphasis on innate characteristics rather than addressing systemic reasons of addiction. Furthermore, the dataset's exclusion of contextual factors, such as socioeconomic status or access to healthcare, limits its ability to distinguish between recreational use and dependency, potentially leading to unjustified interventions.

Ultimately, the ethical value of such models depends on their ability to advance harm reduction rather than unwarranted, biased analysis. By prioritizing privacy, equity, and systemic change, one can harness predictive analytics to address root causes of substance use without adding onto existing disparities.

Conclusion

The K Nearest Neighbors (KNN) model is recommended for predicting cannabis usage levels in this dataset, achieving a higher F1 score compared to Random Forest. Its strength lies in leveraging the natural clustering of non-users and heavy users observed in PCA, which aligns with its distance based calculations. While both models struggled with adjacent class overlap, KNN demonstrated slightly better heavy user identification, but with fewer misclassifications in adjacent categories. These predictive results could inform public health interventions, such as tailoring education campaigns for high-risk groups identified by the model. The unsupervised PCA results, which highlighted impulsivity and sensation seeking as key drivers of usage, provide a foundation for future research into targeted behavioral therapies. By integrating supervised predictions with unsupervised clustering, one can develop better strategies that address both individual risk factors and systemic causes to substance use.

References

Delibaş, D. H., Akseki, H. S., Erdoğan, E., Zorlu, N., & Gülseren, Ş. (2018). Impulsivity, Sensation Seeking, and Decision-Making in Long-Term Abstinent Cannabis Dependent Patients. Noro psikiyatri arsivi, 55(4), 315–319. https://doi.org/10.5152/npa.2017.19304 Fehrman, E., Egan, V., & Mirkes, E. (2015). Drug Consumption (Quantified) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TC7S.

King, D. D., Gill, C. J., Cadieux, C. S., & Singh, N. (2024). The role of stigma in cannabis use disclosure: an exploratory study. Harm reduction journal, 21(1), 21. https://doi.org/10.1186/s12954-024-00929-8

Rosenberg, A., Groves, A. K., & Blankenship, K. M. (2017). Comparing Black and White Drug Offenders: Implications for Racial Disparities in Criminal Justice and Reentry Policy and Programming. Journal of drug issues, 47(1), 132–142. https://doi.org/10.1177/0022042616678614

Scikit-Learn. (2018). sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

scikit-learn. (2019). sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.22.1 documentation. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html