Classificação do Consumo de Drogas a partir de Características Demográficas e de Personalidade

Guilherme S.G. Brotto
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, ES, Brasil
guilherme.brotto@edu.ufes.br

Marlon Moratti do Amaral
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, ES, Brasil
marlon.amaral@edu.ufes.br

Nilo Garcia Monteiro
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, ES, Brasil
nilo.monteiro@edu.ufes.br

***Abstract*—The increasing use of drugs worldwide has raised significant concerns regarding public health consequences and the need for appropriate treatments. This study investigates the relationship between demographic and personality characteristics with substance use, whether legal or illegal, using machine learning techniques. The dataset used throughout this paper contains information on 1885 individuals, including demographic variables, personality traits, and substance use patterns. After some data preprocessing, several classification algorithms were evaluated to identify the most effective in predicting drug use. Gathered results indicate that factors such as age, gender, and certain personality traits are significant predictors of substance use. Additionally, it was also found that the performance of classifiers varies depending on the substance analyzed. This study contributes to understanding the influencing factors on drug consumption and may help in the development of more targeted prevention and treatment strategies.**

***Index Terms*—Drug use, machine learning, demographic characteristics, personality traits, classification.**

I. Introdução

Recentemente, é notável como o uso de drogas vem aumentando drasticamente no mundo inteiro. Segundo a UNODC [1], 1 a cada 17 pessoas usaram alguma droga em 2021, correspondendo a um aumento de 23% em relação à década anterior. O uso varia entre drogas lícitas como álcool e cafeína, até ilícitas como cannabis, ecstasy, heroína e diversas outras drogas sintéticas. Nesse estudo, constata-se que as consequências para um usuário podem incluir mudanças no comportamento, com traços de bipolaridade ou até agressão e danos mentais graves, causando problemas de memória e raciocínio.

De acordo com a mesma fonte, menos de 20% das pessoas com transtornos por uso de drogas estão em tratamento, o que se deve a falta de infraestruturas para mapear os usuários e encaminhá-los para cuidado. Nesse sentido, torna-se necessário compreender a relação entre características individuais com o consumo de drogas a fim de auxiliar essas pessoas antes que a situação se torne irreversível.

Diversos estudos vêm sendo realizados nessa área com o objetivo de entender o que leva indivíduos ao uso e abuso de drogas. Em [2] é estimada a associação entre o uso de drogas ilícitas durante o ensino médio e o número de anos de escolaridade concluídos. No trabalho de [3], procura-se entender se um indivíduo apresenta tendências em ficar viciado ou não em drogas com base em questões de saúde e comportamento social e familiar. Em [4], objetiva-se determinar se é ou não possível fazer uma previsão sobre o abuso de substâncias intoxicantes, o qual é frequente entre estudantes universitários em Bangladesh, utilizando aprendizado de máquina.

Desse modo, o objetivo desse trabalho é avaliar como que traços de personalidade, níveis de educação, idade, etnia e outros atributos afetam o risco de uma pessoa se tornar usuária de uma droga em específico. A princípio, será feita uma classificação simples para buscar os melhores classificadores para o problema, assim como as drogas mais fáceis de serem classificadas, ou seja, aquelas que apresentam padrões bem estabelecidos. Em seguida, serão explorados os melhores algoritmos do passo anterior, realizando uma tunagem de hiperparâmetros, se relevantes. Por fim, serão comparadas as influências de cada atributo para a predição de possíveis usuários, a fim de compreender quais características possuem maior relação com o uso de substâncias lícitas e ilícitas.

II. Dataset

Esta seção apresenta o conjunto de dados utilizado como objeto de estudo. Todos os experimentos deste trabalho utilizam o conjunto disponibilizado em [5]. No total, são disponibilizadas 1885 amostras, cada uma contendo informações sobre um indivíduo. Cada amostra apresenta 30 atributos, que podem ser divididos em características do indivíduo (12 colunas) e padrões de uso de substâncias (18 colunas).

Entre as características do indivíduo, há informações demográficas como faixa etária (*Age*), gênero (*Gender*), escolaridade (*Education*), país (*Country*) e etnia (*Ethnicity*). Além disso, incluem-se medidas de padrões comportamentais como *NEO-FFI-R* (*Nscore*, *Escore*, *Oscore*, *Ascore* e *Cscore*), *BIS-11* e *ImpSS*.

Quanto ao uso de substâncias, o conjunto de dados abrange tanto substâncias legais quanto ilegais. Entre elas estão: álcool (*Alcohol*), anfetaminas (*Amphet*), nitrito de amila (*Amyl*), benzodiazepínicos (*Benzos*), cafeína (*Caff*), cannabis (*Cannabis*), chocolate (*Choc*), cocaína (*Coke*), crack (*Crack*), ecstasy (*Ecstasy*), heroína (*Heroin*), ketamina (*Ketamine*), substâncias legais (*Legalh*), LSD (*LSD*), metadona (*Meth*), cogumelos (*Mushrooms*), nicotina (*Nicotine*), Semeron - droga fictícia (*Semer*) e abuso de substâncias voláteis (*VSA*).

Com conhecimento da gama de substâncias existentes, são definidos dois subconjuntos: drogas relevantes e irrelevantes. O intuito da discriminação vêm da interferência que resultados de substâncias mais comuns possam causar na classificação geral. Assim, definimos como drogas irrelevantes o seguinte subconjunto: álcool, cafeína, chocolate, e semeron. As outras substâncias compõem o grupo complementar.

Uma descrição mais detalhada e a definição dos domínios dos dados utilizados serão feitas posteriormente.

III. Metodologia

O estudo adota uma abordagem que busca identificar o uso de substâncias com base nas características do indivíduo. Essa abordagem pressupõe que padrões comportamentais e variáveis demograficas podem ser tanto a causa quanto a consequência do uso de algumas substãncias. O problema em questão é de classificação, onde as informações sobre o indivíduo constituem a entrada e os padrões de uso de drogas constituem a saída. A metodologia adotada neste estudo é estruturada em três seções principais: *Pré-processamento*, *Características* e *Dados de saída*. Estas seções são fundamentais para compreender a preparação e transformação dos dados utilizados na análise.

A. Pré-processamento

O conjunto de dados passou por uma série de transformações. Primeiramente, variáveis categóricas com uma relação de ordem foram convertidas em variáveis numéricas ordinais, enquanto aquelas sem uma relação de ordem foram transformadas em variáveis numéricas não ordinais. Além disso, as variáveis originalmente numéricas foram normalizadas, ajustando-as para uma média de zero e um desvio padrão de um. Mais informações sobre codificação e normalização de variáveis podem ser encontradas em [6] e [7]. A seguir, são detalhadas as transformações aplicadas no conjunto de dados.

1) OrdinalEncoder: É uma ferramenta essencial para converter variáveis categóricas em números, preservando relações de ordem entre as categorias. É útil quando as categorias têm uma ordem significativa, permitindo que os algoritmos capturem essas relações.

2) OneHotEncoder: É um método utilizado para converter variáveis categóricas em vetores binários, onde cada categoria é representada como um vetor de zeros e um único valor um. Isso é útil quando as categorias não têm uma ordem natural entre elas, evitando que o modelo interprete erroneamente relações de ordem onde não existem.

3) StandardScaler: É uma técnica de pré-processamento comumente utilizada para padronizar (normalizar) características numéricas, transformando-as de modo que tenham média zero e desvio padrão um. Isso é útil quando as características têm escalas diferentes, pois coloca todas no mesmo intervalo, facilitando a comparação entre elas e melhorando o desempenho de algoritmos que são sensíveis à escala dos dados, como métodos baseados em distância.

B. Características

Nesta subseção, serão definidas as características utilizadas. Dentro das características do indivíduo, há dois conjuntos principais: medidas de personalidade e variáveis demográficas. As medidas de personalidade incluem:

- *NEO-FFI-R*, que abrange: neuroticismo, extroversão, abertura à experiência, agradabilidade e conscienciosidade, [8];
- *BIS-11*, que mede a impulsividade, [9];
- *ImpSS*, que avalia a busca de sensações, [10].

Essas medidas são contínuas e, ao todo, totalizam sete medidas de personalidade. Já as variáveis demográficas incluem:

- *Faixa etária* (ordinal), que determina um valor com base em faixas de idade;
- *Gênero* (não ordinal), que determina o gênero como masculino ou feminino;
- *Escolaridade* (não ordinal), que indica se o indivíduo completou a graduação. Originalmente o conjunto de dados apresentava 9 níveis de escolaridade, que foram agrupados nesses dois conjuntos.

As três variáveis são categóricas e foram codificadas de acordo com a sua natureza, tanto ordinal quanto não ordinal, como explicado anteriormente. No total, são utilizadas 10 características para realizar os experimentos. As demais características foram descartadas devido ao desbalanceamento e ao risco de introduzir viés nos resultados.

C. Dados de Saída

Para cada substância citada anteriormente, a saída pode assumir os seguintes valores:

0) Nunca usou,
1) Usou na década anterior,
2) Usou nesta década,
3) Usou neste ano,
4) Usou neste mês,
5) Usou nesta semana,
6) Usou no último dia.

Esse problema pode ser tratado tanto como uma classificação binária, agrupando faixas de valores, quanto como uma classificação multiclasse. Uma série de experimentos será conduzida para explorar essas diferentes possibilidades.

IV. Metodologia Experimental

Esta seção discute os métodos utilizados durante a realização dos experimentos, esclarecendo motivações por trás de decisões tomadas em prol da performance dos modelos. Adiante, são apresentadas algumas vertentes que podem ser tomadas ao se aprofundar no trabalho.

A. Métricas Estatísticas

Devido à distribuição desbalanceada do dataset, o uso de métricas que levam em conta a quantidade de exemplos em cada classe é desejado. Em vista disso, a acurácia balanceada

é utilizada para determinar o desempenho dos classificadores durante os experimentos.

A métrica de acurácia balanceada é semelhante à média macro das revocações por classe. Para uma classe i, calculam-se seus verdadeiros positivos (TP_i) e falsos negativos (FN_i). Com isso, a acurácia balanceada para um problema com n classes é definida como

$$BA = \frac{1}{n}\sum_{j=1}^{n}\frac{TP_j}{TP_j + FN_j} \quad (1)$$

B. Classificadores

Primeiramente, devemos definir um seleto grupo de classificadores que serão utilizados durante os experimentos. Dessa forma, é possível reduzir a computação necessária para executar posteriores experimentos, assim como aprofundar em características intrínsecas dos classificadores, como hiperparâmetros.

Para isso, houve uma avaliação de desempenho médio entre todas as substâncias considerando vários classificadores. Desta forma, classificadores que, independente da substância, obtiveram resultados bons na comparação geral, serão analisados mais profundamente em sequência.

Como há uma grande diversidade de períodos possíveis do uso de substâncias, o *threshold* $t = 2$ foi definido durante a busca. Logo, as classes são explicadas como abaixo.

0) indivíduos que não usam há mais de 1 ano,
1) indivíduos que usaram dentro do perído de 1 ano.

A *framework* utilizada neste processo consiste em uma aplicação da biblioteca *LazyClassifier*[1], a qual avalia uma lista interna de classificadores e retorna seus respectivos desempenhos nos exemplos dados. Para realizar a execução do algoritmo, uma validação cruzada com 4 folds foi aplicada para evitar enviezamentos. A pontuação final de um classificador é definida como o somatório de suas médias obtidas no subconjunto de substâncias relevantes como citado na seção II.

Por fim, para auxiliar nos experimentos e na visualização de resultados, um inteiro $k = 3$ foi definido de forma que somente os primeiros k classificadores serão utilizados nos demais experimentos. Além dos classificadores encontrados durante a busca, dois classificadores clássicos foram adicionados: *Random Forest* e *K-Nearest Neighbors* (KNN). Isso possibilita uma comparação justa entre métodos, de forma a entender o impacto que um classificador pode ter nos dados apresentados. A tabela I apresenta todos os classificadores utilizados nos experimentos da seção V, assim como breve definições dos algoritmos e se há hiperparâmetros a serem testados.

C. Ajuste de Hiperparâmetros

Uma abordagem sistemática e amplamente utilizada para ajuste de hiperparâmetros é a *Grid Search*, que consiste em explorar o espaço de hiperparâmetros de forma exaustiva através de uma grade predefinida de possíveis valores. Diferente de métodos aleatórios, a *Grid Search* examina todas as combinações possíveis de hiperparâmetros, garantindo que nenhuma configuração potencial seja negligenciada. Essa abordagem meticulosa é especialmente valiosa quando é essencial maximizar o desempenho do modelo, pois permite uma avaliação completa e comparativa de cada combinação. Embora a *Grid Search* possa demandar maior tempo de computação, a precisão e abrangência na busca por configurações ótimas de hiperparâmetros podem resultar em um modelo mais robusto e eficiente. Portanto, durante a fase de validação, a *Grid Search* foi empregada para determinar o melhor conjunto de hiperparâmetros para cada modelo e experimento distinto, assegurando um processo rigoroso de otimização.

[1]Disponibilizado em: https://github.com/shankarpandala/lazypredict

TABLE I
CLASSIFICADORES UTILIZADOS NOS EXPERIMENTOS

Classificador	**Algoritmo**	**HP**[a]
Nearest Centroid	Cada classe do problema é relacionada a um centroide. Predições são definidas com base na proximidade à um centroide.	Não
Gaussian Naive Bayes	Naive Bayes gaussiano.	Não
Bernoulli Naive Bayes	Naive Bayes especializado em características binárias/booleanas	Não
Random Forest	Ensemble de árvores de decisão com subsets do dataset.	Sim
KNN	Define predições com base nos exemplos de treino ao seu redor.	Sim

[a]Define se há hiperparâmetros a serem testados.

D. Configuração Experimental

Com a grande coleção de informações fornecidas pelo conjunto de dados, há diversas possibilidades de pré-processamento, análise e classificação. De um lado, é possível obter modelos que predizem o uso de uma substância com base na personalidade do indivíduo. De outro, treinar um modelo que determina um padrão no consumo considerando somente indivíduos mais novos também parece viável. Portanto, cinco experimentos foram realizados visando percorrer várias vertentes do problema.

Todos os experimentos foram conduzidos utilizando de validação cruzada estratificada aninhada com 3 repetições, 10 subconjuntos no ciclo externo e 4 subconjuntos no ciclo interno, se necessário, a fim de evitar sobreajuste nos dados.

O primeiro experimento, denominado "Classificação Geral" na subseção de resultados V-A, tem como objetivo fazer uma breve análise do desempenho dos classificadores selecionados para compor o modelo.

O segundo experimento, denominado "Impacto da Personalidade" na subseção de resultados V-B, busca identificar se os padrões comportamentais do indivíduo exclusivamente possuem implicação no uso de substâncias. Neste caso, assume-se que o comportamento é indiferente das outras características.

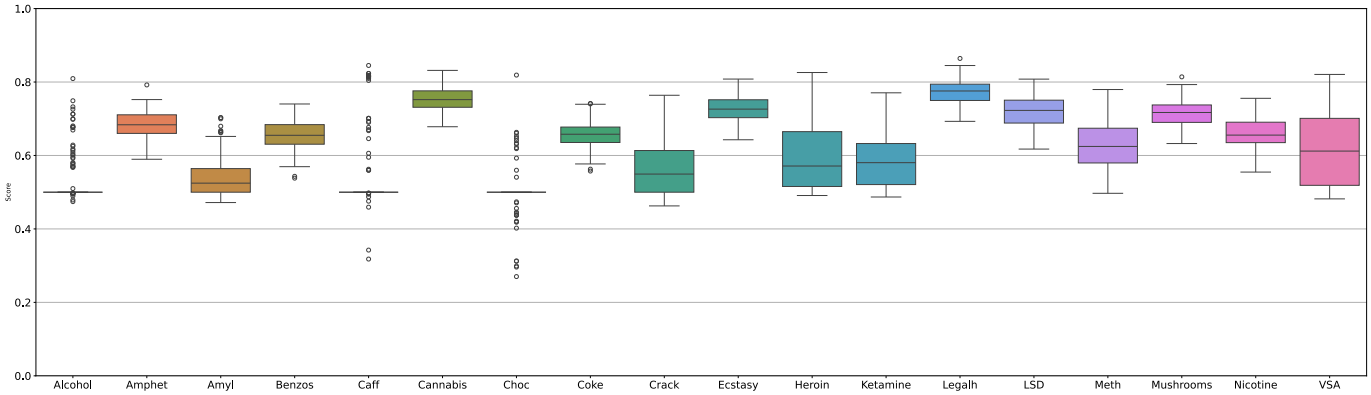

Fig. 1. Boxplot da média da acurácia balanceada em todos os classificadores selecionados, em relação a cada substância.

O terceiro experimento, denominado ”Filtro de Idades” na subseção de resultados V-C, visa observar os dados em um escopo menor, de forma a determinar se há padrões entre um grupo com idades mais baixas, especificamente entre 18 e 34 anos. Da mesma forma, o *threshold* t é equivalentemente menor, considerando um uso válido (1) como aquele que foi feito dentro de um período mensal. A hipótese levantada é que, em idades menores, há menos variação de tempo no consumo de substâncias, justificando a aplicação de classificação binária ao problema. Como somente dois blocos de idades estão presentes, a característica não é utilizada pelos classificadores.

O quarto experimento, denominado ”Redução de Dimensionalidade” na subseção de resultados V-D, explora o uso de métodos para reduzir a complexidade intrínseca de dados exuberantes. Com isso, o *Principal Component Analysis* (PCA) é utilizado para diminuir a distância espacial.

O quinto experimento, denominado ”Problema Multiclasse” na subseção de resultados V-E, estuda a aplicação de mais classes no problema. A presunção existente é que, com uma menor distância entre os períodos desde o último uso de uma substância, os classificadores consigam ter um desempenho maior. Assim, os *thresholds* foram definidos de forma que três classes são definidas: (0) para períodos de mais de 1 ano, (1) dentro de 1 ano e (2) dentro de 1 mês.

Os experimentos foram realizados utilizando Python e a biblioteca de código aberto *scikit-learn*.

V. Experimentos e Resultados

Esta seção apresenta os resultados e conclusões obtidas nos experimentos.

A. Classificação Geral

Para obter melhor compreensão sobre que algoritmos possam interpretar melhor os dados apresentados, foi executada uma busca de classificadores. Cada substância foi testada para cada algoritmo e os 3 melhores classificadores gerais foram selecionados para compor o conjunto inicial. Ademais, como citado previamente, houve a adição de *Random Forest* e *KNN* ao conjunto. A performance média geral dos classificadores por substância é mostrada na figura 1. É inferível que há

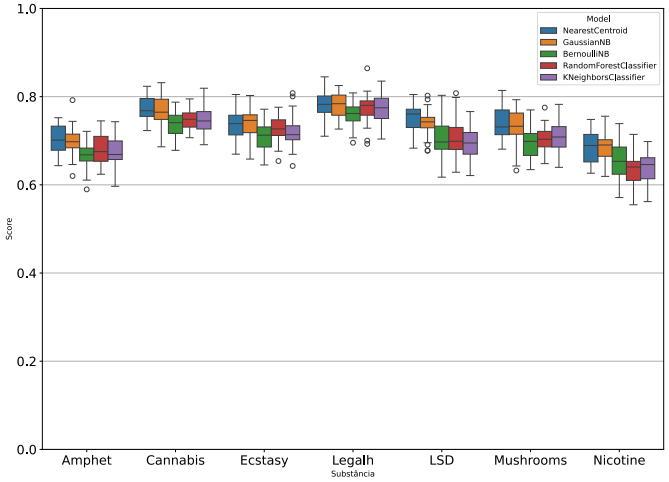

Fig. 2. Boxplot das acurácias balanceadas em todos os classificadores selecionados, nas 7 substâncias selecionadas.

substâncias em que o grupo possui desempenho estável, como, por exemplo, cannabis, substâncias legais, ecstasy, LSD e cogumelos.

Considerando a visualização e comparação de desempenho, foram selecionadas as 7 substâncias com maior média geral no grupo. Experimentos subsequentes, portanto, focarão na comparação das performances ao utilizar diferentes métodos no ramo de inteligência artificial. A figura 2 apresenta o desempenho dos classificadores individualmente para as substâncias mais bem colocadas. No total, não há destaques entre os classificadores ou substâncias, mas nota-se que substâncias como cannabis e substâncias legais (*legalh*) são mais bem representadas.

B. Impacto da Personalidade

Levando em conta que mudanças de comportamento são notáveis entre usuários [1], procura-se entender como o padrão comportamental de um indivíduo pode indicar o uso de substâncias. Portanto, esse experimento analisa somente as métricas de personalidade, buscando encontrar padrões entre usuários e substâncias. Em geral, a figura 3 mostra que não

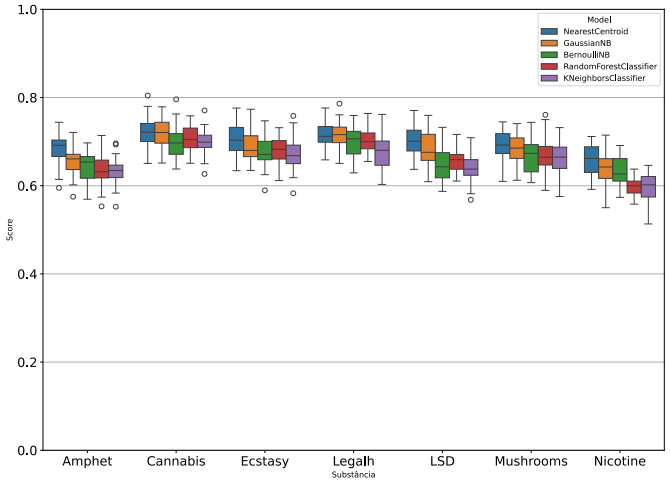

Fig. 3. Boxplot das acurácias balanceadas em todos os classificadores selecionados, nas 7 substâncias selecionadas, com uso exclusivo das características de personalidade.

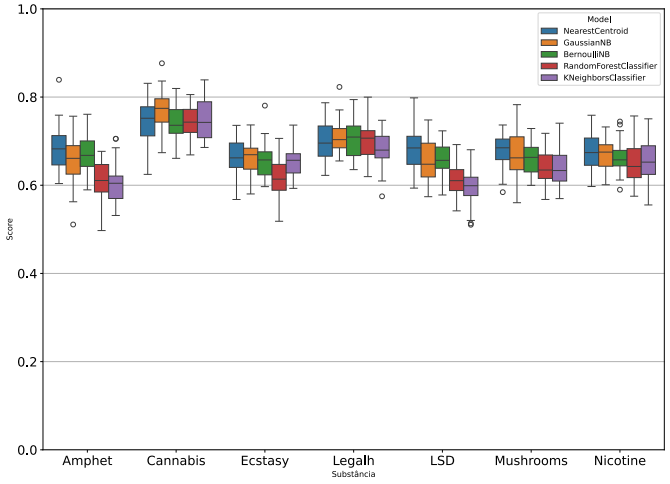

Fig. 4. Boxplot das acurácias balanceadas em todos os classificadores selecionados, nas 7 substâncias selecionadas, com aplicação do filtro de idades.

há melhora significativa ao comparar os resultados deste com o uso de todas as informações fornecidas pelo conjunto de dados na figura 2.

C. Filtro de Idades

Com a grande gama de informações oferecidas pelo dataset, é possível que padrões sejam ocultados por outras relações enganosas entre as caracterísiticas. Ao filtrar os dados, é possível reduzir o escopo e estudar o problema mais afundo. O pressuposto é que, com um range menor de idade, indivíduos sejam mais parecidos e, logo, mais comparáveis. A figura 4 mostra que, principalmente, houve mais variação nos resultados, deixando predições menos confiáveis. Assim, é inferido que a idade tem grande interferência na classificação, considrando que sua filtragem causou grande variação nos resultados.

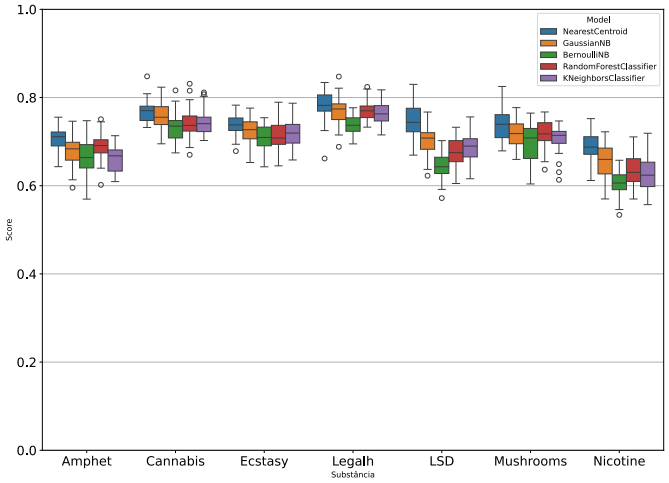

Fig. 5. Boxplot das acurácias balanceadas em todos os classificadores selecionados, nas 7 substâncias selecionadas, com redução de dimensionalidade via PCA.

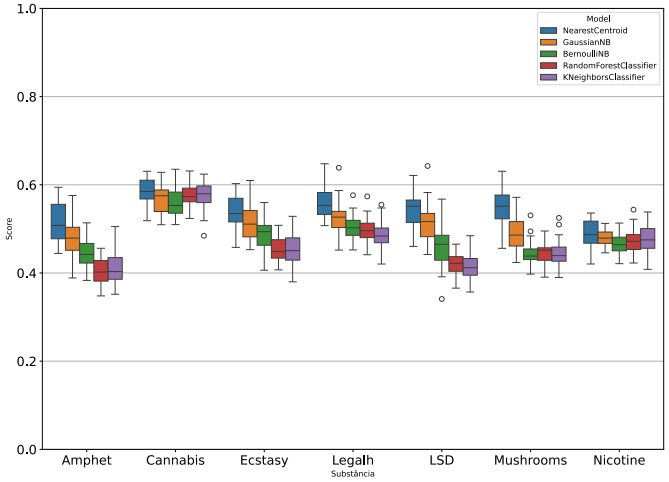

Fig. 6. Boxplot das acurácias balanceadas em todos os classificadores selecionados, nas 7 substâncias selecionadas, considerando um problema multiclasse.

D. Redução de Dimensionalidade

Muita informação pode ser, muitas vezes, solicitada pelo modelo, entretanto, na maior parte do tempo, um aumento na dimensionalidade dos dados introduz cada vez mais problemas. Sejam eles problemas computacionais ou de desempenho, geralmente procuram-se maneiras de reduzir a informação sem perder suas contribuições. Dessa forma, levanta-se a experimentação do uso de PCA para reduzir a dimensionalidade dos dados, possivelmente facilitando a classificação. Como apresentado na figura 5, a redução não obteve impactos majoritariamente negativos, mas não introduziu nenhuma interpretação nova para as características. Em alguns casos, tal adição ocasionou na queda de performance, como em alguns classificadores com nicotina e LSD.

E. Problema Multiclasse

Por fim, analisa-se a interpretação do problema como multiclasse. Previamente, utilizando de uma classificação binária, nota-se que há uma distância não desejada entre classes por conta do salto temporal extremo. Nesta configuração, pressupõe-se que as características tenham mais informações quando distribuídas entre mais classes. A figura 6 demonstra que, ao menos, o uso de cannabis tem um padrão ligeiramente reconhecível com tal configuração. Entretanto, outras substâncias não se beneficiam da mudança.

VI. Conclusões

Este trabalho apresentou, ao todo, diferentes interpretações do problema de classificação do uso de substâncias. Tais análises podem ser utilizadas posteriormente para possíveis aprofundamentos em problemas relacionados a grupos mais seletos de substâncias.

Os estudos presentes neste trabalho demonstraram que o problema de classificação do consumo de substâncias é bastante complexo, logo, podem ser necessárias mais informações sobre os participantes. Apesar disso, ainda há padrões a serem descobertos ao se aprofundar nos dados em questão. Entretanto, os classificadores apresentados podem não conseguir capturar tais nuâncias. Em geral, um dos destaques entre os classificadores é o *Nearest Centroid*, que, pela construção de centroides, consegue aproximar melhor os indivíduos, aplicando padrões mais parecidos e resultando em pontuações mais sensatas.

References

[1] "World drug report 2023."

[2] P. Chatterji, "Illicit drug use and educational attainment," *Health economics*, vol. 15, pp. 489–511, 05 2006.

[3] A. Shahriar, F. Faisal, S. Uddin Mahmud, A. Chakrabarty, and M. G. R. Alam, "A machine learning approach to predict vulnerability to drug addiction," pp. 1–7, 12 2019.

[4] M. A. Ismail and A. Islam, *Study on the Analysis and Prediction of Drug Addiction Among University Students of Bangladesh Using Machine Learning*, pp. 181–196. 03 2024.

[5] "Drug consumption dataset." Accessed: 2024-06-10.

[6] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International journal of computer applications*, vol. 175, no. 4, pp. 7–9, 2017.

[7] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 729–735, 2020.

[8] R. R. McCrae and P. T. Costa, "A contemplated revision of the neo five-factor inventory," *Personality and Individual Differences*, vol. 36, no. 3, pp. 587–596, 2004.

[9] M. S. Stanford, C. W. Mathias, D. M. Dougherty, S. L. Lake, N. E. Anderson, and J. H. Patton, "Fifty years of the barratt impulsiveness scale: An update and review," *Personality and Individual Differences*, vol. 47, no. 5, pp. 385–395, 2009.

[10] M. Zuckerman, *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge University Press, 1994.