

Uma aplicação de Aprendizagem de Máquina sobre a base: Drug Consumption

Autores: Manuel Ferreira Junior e
Caroline Assis de Oliveira.

Disciplina: Aprendizagem de Máquina

Professor: Telmo de Menezes e Silva Filho

<https://github.com/Manuelfjr/am-drugs-consumption>



Sumário

Introdução

Descrição dos Dados

Metodologia

- Processamento

- Análise Descritiva e Visualização dos Dados

- Modelagem

 - ADASYN

 - Árvore de Decisão para Classificação

 - Floresta Aleatória

 - Regressão Logística

 - KMeans

Resultados

- Processamento

- Análise Descritiva e Visualização dos Dados

 - Análise Descritiva e Visualização dos Dados

- Modelagem

 - Árvore de Decisão

 - Floresta Aleatória

 - Regressão Logística

 - KMeans

Conclusão

Referências



Introdução

A análise de dados se torna cada vez mais imprescindível nos tempos atuais, neste relatório será abordada diferentes formas de se analisar o banco de dados *drug consumption*, que tem como informações o consumo de drogas lícitas e ilícitas, o sexo do usuário, bem como alguns comportamentos que podem influenciar o consumo, em alguns países. Foram utilizados quatro métodos de machine learning, três com tarefas de classificação (regressão logística, árvore de decisão e floresta aleatória) e um de agrupamento (KMeans). Ao longo desta apresentação serão abordadas as técnicas utilizadas para aplicação destes métodos.



Descrição dos Dados

Para deixar nossas análises livre de preconceito, retiramos a variável **ethnicity** que indica a etnia do participante do estudo.

Após aplicados os métodos de **encoding** que serão descritos na subseção de Pré-processamento, a base tem as seguintes características.

- ▶ **age** indica a idade do participante no momento do estudo, e tem os seguintes valores
 - ▶ 0 para idades entre 18 e 24 anos;
 - ▶ 1 para idades entre 25 e 34 anos;
 - ▶ 2 para idades entre 35 e 44 anos;
 - ▶ 3 para idades entre 45 e 54 anos;
 - ▶ 4 para idades entre 55 e 64 anos;
 - ▶ 5 para idades maiores que 65 anos.



Descrição dos Dados

- ▶ **gender** indica o sexo do participante
 - ▶ 0 para Masculino;
 - ▶ 1 para Feminino.
- ▶ **country** indica qual o país que o participante vive
 - ▶ 0 Estados Unidos;
 - ▶ 1 Nova Zelandia;
 - ▶ 2 Outro;
 - ▶ 3 Austrália;
 - ▶ 4 Irlanda;
 - ▶ 5 Canadá;
 - ▶ 6 Reino Unido.



Descrição dos Dados

- ▶ **education** indica o nível educacional do participante
 - ▶ 0 para participantes que deixaram a escola antes dos 16 anos;
 - ▶ 1 para participantes que deixaram a escola aos 16 anos;
 - ▶ 2 para participantes que deixaram a escola aos 17 anos;
 - ▶ 3 para participantes que deixaram a escola aos 18 anos;
 - ▶ 4 para participantes que ingressaram em alguma faculdade ou universidade, mas não se formaram;
 - ▶ 5 para participantes que concluíram o ensino médio ou algum curso profissionalizante;
 - ▶ 6 para participantes com Ensino Superior;
 - ▶ 7 para participantes com Mestrado;
 - ▶ 8 para participantes com Doutorado.



Descrição dos Dados

- ▶ **nscore**, **escore**, **oscore**, **ascore**, **cscore**, **impulsive** e **ss** são **scores** para os respectivos comportamentos: neuroticismo, extroversão, abertura à experiência, afabilidade, consciência, impulsividade e busca de sensação.



Descrição dos Dados

- ▶ **alcohol, amphet, amyl, benzos, caff, cannabis, choc, coke, crack, ecstasy, heroin, ketamine, legalh, lsd, meth, mushrooms, nicotine, vsa e semer** são as variáveis que indicam conforme descrição abaixo o uso das respectivas drogas: álcool, anfetaminas, nitrito de amila, benzodiazepina, cafeína, cannabis, chocolate, cocaína, crack, ecstasy, heroína, cetamina, drogas legais, LSD, metadona, cogumelos, nicotina, abuso de drogas inaláveis e uma droga fictícia (Semeron), que foi introduzida para identificar aqueles que reivindicam excessivamente. Cada droga listada acima apresentam as seguintes classes



Descrição dos Dados

- ▶ 0 indica que nunca usou;
- ▶ 1 indica que usou a mais de uma década;
- ▶ 2 indica que usou na ultima década;
- ▶ 3 indica que usou no ultimo ano;
- ▶ 4 indica que usou no ultimo mês;
- ▶ 5 indica que usou na ultima semana;
- ▶ 6 indica que usou no ultimo dia.



Descrição dos Dados

- ▶ Após essa transformação, as variáveis que indicam as drogas foram binarizadas para criar a variável **used**, que por sua vez indica o uso de drogas ilícitas
 - ▶ 0 indica que o participante nunca usou;
 - ▶ 1 indica que o participante já usou alguma droga até o momento do estudo.



Processamento

Os resultados na base original estavam normalizados, para serem vistos e analisados com mais clareza foi feita o processo de label encoding para retornar aos seus valores originais.



Análise Descritiva e Visualização dos Dados

Como parte das análises descritivas foi um mapa de calor com as **correlações** das variáveis. Para vermos as diferenças entre os sexos para as variáveis que indicam drogas foi utilizado o teste de Wilcoxon (Wilcoxon 1992), um teste estatístico não-paramétrico que mede a diferença de dois grupos mutuamente independentes.



ADASYN

A ideia essencial do ADASYN (He et al. 2008)(**adaptive synthetic sampling approach**, em tradução livre abordagem sintética adaptativa de amostragem) é usar uma distribuição ponderada para diferentes exemplos de classes minoritárias de acordo com seu nível de dificuldade de aprendizagem, onde mais dados sintéticos é gerado para exemplos de classes minoritárias que são mais difíceis de aprender em comparação com os exemplos minoritários que são mais fáceis de aprender. Como resultado, a abordagem ADASYN melhora o aprendizado com relação às distribuições de dados de duas maneiras: (1) reduzindo o viés introduzido pelo desequilíbrio de classe, e (2) adaptativamente mudando o limite de decisão de classificação para exemplos difíceis. As análises de simulação em vários dados de aprendizado de máquina conjuntos mostram que esse método tem grande eficácia.



Árvore de Decisão para Classificação

A árvore de decisão é um método que separa a base em grupos e define os aspectos que definem uma certa característica desta base, para assim poder classificar da melhor maneira



Floresta Aleatória

O algoritmo da Floresta Aleatória nada mais é do que um ensemble ou conjunto de Árvores de Decisão, com o intuito de obter melhores métricas para a tarefa de aprendizado supervisionado que for proposta para ela.



Regressão Logística

A Regressão Logística é uma técnica estatística com o intuito de prever, dado um conjunto de observações independentes (\mathbf{X}), uma variável dependente (\mathbf{Y}) costumeiramente sendo binária. Apartir disso, é possível prever a probabilidade de um evento ocorrer, dado um grupo de variáveis independentes.



KMeans

O KMeans é um algoritmo de clustering (agrupamento), sendo um método de aprendizagem não supervisionada, agrupando em grupos distintos as observações, de tal forma que entre os grupos sejam Heterogêneos entre os grupos e Homogêneos dentro dos grupos.



Processamento

A base apresentava uma padronização já pré-existente sobre os dados originais, decidimos então que seria interessante utilizarmos um método de *Label Encoder*¹, com o intuito de tirarmos a padronização dos dados; dessa forma, também foi binarizado as colunas de todas as drogas para representar se o indivíduo é usuário daquilo ou não.



¹ <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

Processamento

Além disso, criamos uma nova coluna que será a nossa variável target para o problema, sendo ela binária e com o intuito de classificar se o usuário é ou não usuário de uma droga ilícita, obtendo assim a seguinte proporção dos dados:

| Classe | (%) |
|--------|---------|
| 1 | 84.14 % |
| 0 | 15.86 % |

Tabela: Proporção das classes *Used*

Pode-se notar que a proporção de usuários é muito maior que a de não usuários neste estudo (cerca de 84% dos participantes), mais a frente na seção **ADASYN**² será apresentado uma solução para balancear as proporções.



²https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html

Análise Descritiva e Visualização dos Dados

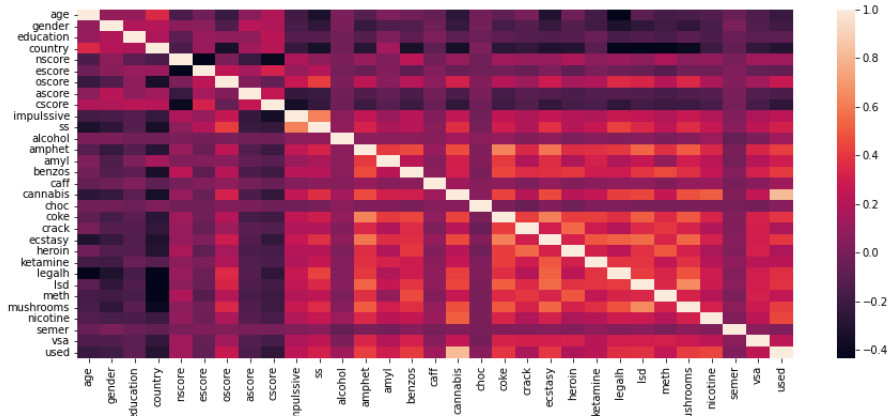


Figura: Mapa de calor da correlação entre as variáveis (ρ_{xy})



Análise Descritiva e Vizualização dos Dados

Pelo mapa de calor é perceptível que grande parte das drogas são correlacionadas entre si, podendo indicar que a utilização de uma droga pode levar ao consumo de outras drogas. Anteriormente, podemos ver a distribuição da variável used, existindo uma discrepância forte entre classes.



Análise Descritiva e Vizualização dos Dados

Além disso, ao binarizarmos as colunas de drogas ilícitas, podemos analisar a proporção de usuários de cada droga no banco, visto na tabela a baixo:

| Drogas | % |
|-----------|-------|
| cannabis | 78.09 |
| amphet | 48.22 |
| mushrooms | 47.90 |
| benzos | 46.95 |
| ecstasy | 45.84 |

Tabela: Proporção de usuários de Drogas Ilícitas

As drogas com maior consumo entre os participantes são cannabis, amfetaminas, cogumelos, benzodiazepina e ecstasy.



Análise Descritiva e Vizualização dos Dados

Tabela: Distribuição do número de usuários das 5 principais drogas ilícitas por sexo por nível educacional

| gender | education | cannabis | amphet | mushrooms | benzos | ecstasy |
|--------|-----------|----------|--------|-----------|--------|---------|
| 0 | 0 | 13 | 9 | 10 | 10 | 8 |
| | 1 | 46 | 36 | 25 | 27 | 27 |
| | 2 | 16 | 13 | 14 | 9 | 11 |
| | 3 | 60 | 42 | 40 | 34 | 46 |
| | 4 | 325 | 216 | 241 | 195 | 232 |
| | 5 | 110 | 70 | 70 | 66 | 63 |
| | 6 | 149 | 110 | 107 | 95 | 89 |
| | 7 | 84 | 49 | 46 | 46 | 43 |
| | 8 | 23 | 17 | 16 | 21 | 17 |
| 1 | 0 | 9 | 9 | 5 | 9 | 7 |
| | 1 | 19 | 12 | 7 | 17 | 9 |
| | 2 | 9 | 7 | 6 | 7 | 7 |
| | 3 | 26 | 11 | 13 | 14 | 11 |
| | 4 | 148 | 79 | 85 | 91 | 89 |
| | 5 | 78 | 46 | 40 | 56 | 37 |
| | 6 | 198 | 109 | 107 | 105 | 106 |
| | 7 | 118 | 50 | 52 | 59 | 43 |
| | 8 | 41 | 24 | 19 | 24 | 19 |



Análise Descritiva e Visualização dos Dados

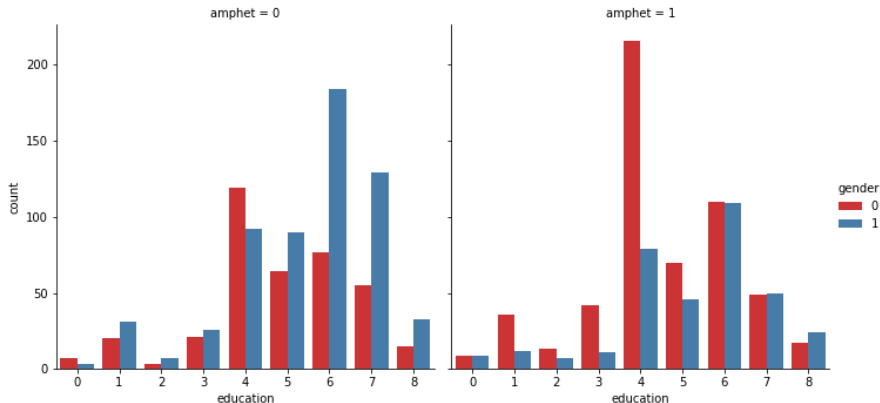


Figura: Distribuição do número de usuários de anfetamina por nível educacional separado por sexo



Análise Descritiva e Visualização dos Dados

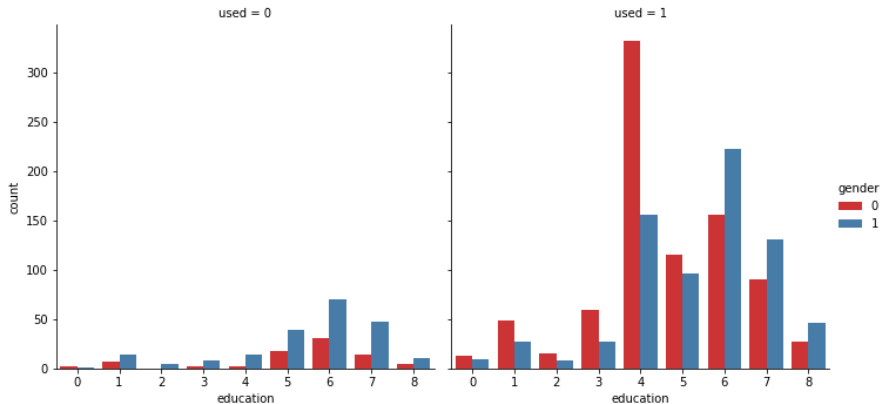


Figura: Distribuição do número de usuários de drogas ilícitas por nível educacional separado por sexo



Análise Descritiva e Vizualização dos Dados

É possível ver que, de forma geral, para níveis educacionais mais baixos existem mais usuários homens para as drogas, quando chegamos aos níveis mais altos (graduação, mestrado e doutorado) existem mais usuárias do sexo feminino. Aplicando o teste de Wilcoxon entre os sexos, é possível medir se há ou não diferença significativa entre o consumo das diferentes drogas ilícitas.



Análise Descritiva e Vizualização dos Dados

| | p-value |
|-----------|---------|
| amphet | 0.0793 |
| amyl | 0.0499 |
| benzos | 0.3726 |
| cannabis | 0.5533 |
| coke | 0.1614 |
| crack | 0.0654 |
| ecstasy | 0.0929 |
| heroin | 0.0378 |
| ketamine | 0.0176 |
| legalh | 0.0180 |
| lsd | 0.0177 |
| meth | 0.0251 |
| mushrooms | 0.0499 |
| semer | 0.4795 |
| vsa | 0.1088 |
| used | 0.5940 |

Tabela: Teste de Wilcoxon para comparar se existe diferença entre os sexos, quanto a nível educacional por droga.



Análise Descritiva e Visualização dos Dados

Adotando um nível de significância de 5% ($\alpha = 0,05$), podemos notar que só há diferença (rejeitamos H_0) no consumo das drogas ketamina, lsd, drogas legalizadas, heroína, nitrito de amila e cogumelos. Para as outras drogas e também o indicador de usuário de drogas ilícitas, o consumo entre os gêneros é estatisticamente iguais.



Separação dos Dados pelo método ADASYN

Nesta seção aplicamos o método ADASYN na variável resposta (used) para balancear as proporções, e também é feita a separação das bases de treino e teste para os modelos que serão feitos a seguir.



Separação dos Dados pelo método ADASYN

| used | Sem ADASYN | Com ADASYN |
|------|------------|------------|
| 1 | 1586 | 1586 |
| 0 | 299 | 1560 |

Tabela: Distribuição da variável target antes e depois do ADASYN

Como é possível observar, nossa variável resposta tem proporções de ocorrência bem mais próximas.



Separação dos Dados pelo método ADASYN

Por fim, a base foi separada em 20% de teste e 80% de treino, essa separação será utilizada em todos os modelos feitos aqui.



Árvore de Decisão

Será utilizado para o modelo de Árvore de Decisão e os próximos, um dicionário de possibilidades de parâmetros e logo em seguida será aplicado Grid Search com o intuito de encontrar os melhores parâmetros para o modelo. Dessa forma, Através da seleção dos melhores parametros para a árvore de decisão, pode-se ver que a entropia cruzada é o melhor critério, fazendo com que o modelo performe melhor.



Feature Importance

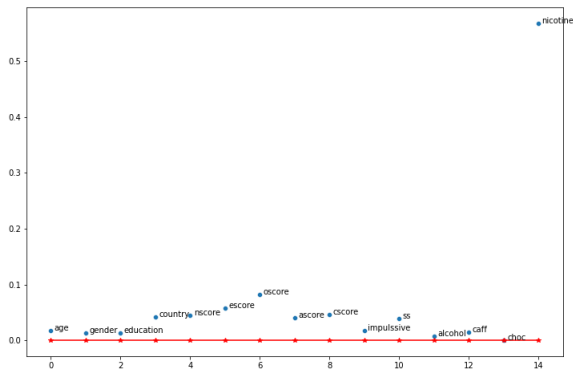


Figura: Feature Importance para a Arvore de Decisao



Validação Cruzada

Obtivemos uma acuracia de 0.8506

| | 0 | 1 |
|-----------|--------|--------|
| Recall | 0.8609 | 0.8406 |
| Precision | 0.8435 | 0.8621 |

Tabela: Resultados da validação cruzada



Floresta Aleatoria

Assim como na árvore de decisão, foi feita a seleção dos melhores parametros para a floresta aleatória, e o índice de gini é o melhor para medir a impureza nesta situação.



Feature Importance

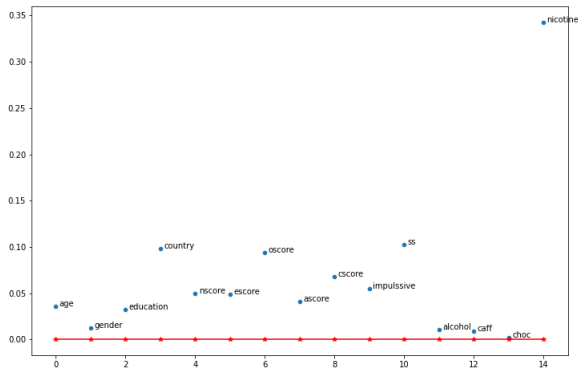


Figura: Feature Importance para a Floresta Aleatoria



Validação Cruzada

Obtemos uma acuracia de 0.8802

| | 0 | 1 |
|-----------|--------|--------|
| Recall | 0.8699 | 0.8904 |
| Precision | 0.8865 | 0.8756 |

Tabela: Resultados da validação cruzada



Regressão Logística

Como para os outros modelos, também foi realizado um grid de parâmetros para encontrar a melhor combinação dos mesmos, de tal forma que otimize a performance do modelo.



Feature Importance

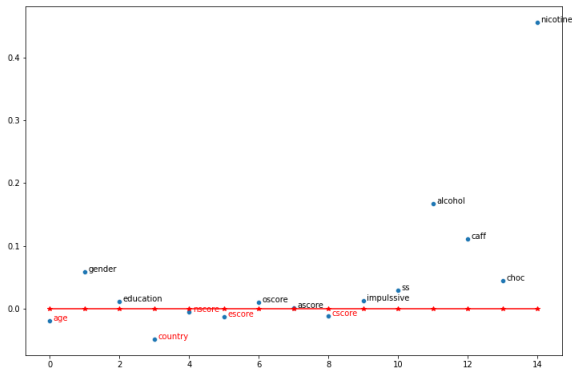


Figura: Feature Importance para a Regressão Logística



Validação Cruzada

Obtemos uma acuracia de 0.8583

| | 0 | 1 |
|-----------|--------|--------|
| Recall | 0.8468 | 0.8696 |
| Precision | 0.8649 | 0.8536 |

Tabela: Resultados da validação cruzada



KMeans

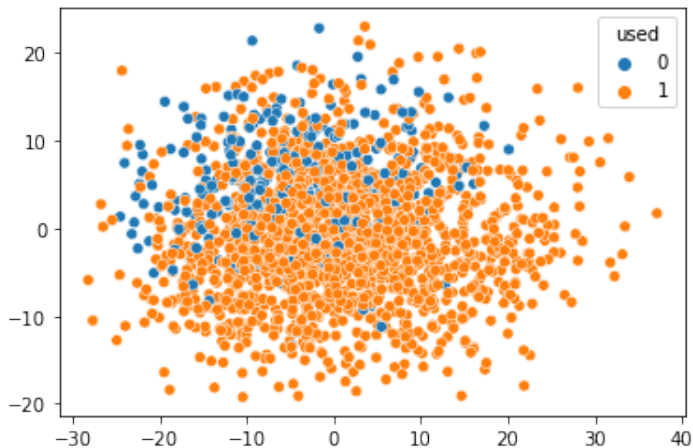


Figura: PCA aplicado sobre o conjunto de dados (ncomponents=2)



KMeans

Como podemos ver no gráfico, as classes estão dispostas de forma sobrepostas, dificultando a tarefa de agrupamento e tornando assim inviável de ser utilizado algum método de agrupamento. Foi aplicado o método KMeans para ilustrar que para esse objetivo (identificar usuários de não usuários de drogas ilícitas) essa tarefa não é a mais recomendada, obtendo um índice de rand médio de 0.0444, considerando um método de avaliação similar a validação cruzada



Conclusão

Com base nos resultados descritos acima, podemos observar que a droga de maior importância nos modelos de classificação é a nicotina, uma droga legalizada e de alto teor viciante, que causa problemas de saúde extremamente graves, podendo ser a porta de entrada para outras drogas que são consumidas da mesma forma. O algoritmo de **floresta aleatória** apresentou a melhor performance dentre os modelos de classificação, apresentando altos índices de precisão.





Conclusão

Um fato intrigante que pôde ser observado é que o agrupamento não é uma boa técnica para esta análise dada a característica de sobreposição dos grupos. O modelo proposto pode ser utilizado também para predizer a probabilidade de um individuo ter usado qualquer tipo de droga ilícita, podendo ser interessante para empresas que trabalham com testagem anti-dopping.



Referências

-  He, Haibo et al. (2008). “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. *Em: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, pp. 1322–1328.
-  Wilcoxon, Frank (1992). “Individual comparisons by ranking methods”. *Em: Breakthroughs in statistics*. Springer, pp. 196–202.

