U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Supervised Learning

Miguel Lima 202108659
Pedro Romão 202108660
Pedro Paixão 202008467

IA - FEUP – 25/05/2024

Definição do problema

Dependência de nicotina

- Finalidade:
  - Previsão do grau de consumo/dependência de nicotina baseada nas características físicas, culturais, psicológicas de uma pessoa.
- Abordagem:
  - Utilizar o consumo de outras substâncias para prever o consumo de nicotina.
  - Características físicas: idade, género.
  - Características culturais: país, etnia e nível de escolaridade.
  - Traços de personalidade (segundo a teoria dos Cinco Grande Fatores) :
    - Neuroticismo: Tendência a ser ansioso, depressivo, sob efeito de tensão, etc.
    - Extroversão: Preferência por estar com pessoas, participar em eventos coletivos, etc.
    - Abertura a experiências: Tendência a ser aberto a novas experiências, curioso, imaginativo, etc.
    - Agradabilidade: Tendência a ser caloroso, amigável, facilidade em relações interpessoais, etc.
    - Conscienciosidade: Tendência a seguir planos, ser metódico, pouco espontâneo, etc.

Referências Bibliográficas

- Dataset utilizado.
  (https://www.kaggle.com/datasets/mexwell/drug-consumption-classification)

- Análise de dados e Visualização da informação da mesma base de dados.
  (https://www.kaggle.com/code/kimkijun7/drug-consumption-classifier-with-python)

- Um modelo de ML criado para a mesma database.
  (https://www.kaggle.com/code/a3amat02/drug-consumption-eda-models)

- Modelo que associa os traços da personalidade à criação de vícios.
  (https://pt.wikipedia.org/wiki/Teorias_da_personalidade_da_adic%C3%A7%C3%A3o)

- Processo utilizado na implementação do projeto: CRISP-DM.
  (https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining)

Ferramentas

- Python como linguagem de programação que suporta as bibliotecas que utilizamos.
- Jupyter Notebook para descrever o trabalho realizado de maneira estruturada.
- Pandas para manipulação de dados, pré-processamento, criação de tabelas e estatísticas.
- Seaborn para uma interface apelativa de gráficos estatísticos.
- Matplotlib para criação de gráficos.
- SciKit-Learn para treinar algoritmos de Machine Learning e a avaliar esses modelos.

Abordagem

- Análise exploratória dos dados e o seu Pré-processamento.
- Definir o objetivo (grau de consumo de nicotina).
- Dividir os dados em conjunto de treino e de teste.
- Selecionar Algoritmos de Supervised ML (Árvores de decisão, k-NN, SVM, etc.).
- Treinar os modelos com o conjunto de dados de treino.
- Avaliar os modelos, comparar os resultados e tirar conclusões.
- Parametrizar o modelo com melhor exatidão (accuracy).
- Tirar conclusões dos resultados obtidos.

Algoritmos

Alguns modelos que treinamos/usamos e estudamos nas aulas:

- Decision Trees
- Support Vector Machine
- K-Nearest Neighbors
- Neural Networks

Também exploramos outros algoritmos: Logistic Regression, Extra Trees, Random Forest e Naive Bayes.

Pré-Processamento dos dados

Limpeza dos Dados e Filtragem de Outliers

Da análise que fizemos, concluímos que não havia valores em falta em nenhum atributo e não existiam itens duplicados. Todos os valores estavam uniformes e de acordo com as especificações do fornecedor do dataset. Removemos também a coluna com os IDs, pois não seriam úteis para a análise.

Codificação da Variável Alvo

As colunas relativas ao consumo de droga estavam numa escala de CL0 a CL6 (sendo CL0 "usou ontem" e CL6 "nunca usou"). Para simplificação, dividimos o consumo de nicotina em: 0 - nunca usou ou usou há mais de um ano e 1 - usou há menos de um ano. Para as restantes drogas, transformamos o consumo em um valor numérico de 0 a 6, com o mesmo significado da escala inicial.

Teste 1:
Escala de 0/1 apenas para a classe Nicotina

Classifier	Precision	Recall	Accuracy	F1 Score	Training Time
Logistic Regression	0.777778	0.803828	0.763926	0.790588	0.028445
Decision Tree	0.716895	0.751196	0.697613	0.733645	0.008405
Extra Trees	0.758621	0.842105	0.763926	0.798186	0.147902
SVM	0.758772	0.827751	0.758621	0.791762	0.040491
Neural Network	0.733032	0.775120	0.718833	0.753488	1.311188
K-Nearest Neighbors	0.731132	0.741627	0.705570	0.736342	0.000000
Random Forest	0.755459	0.827751	0.755968	0.789954	0.174182
Naive Bayes	0.775862	0.645933	0.700265	0.704961	0.000000

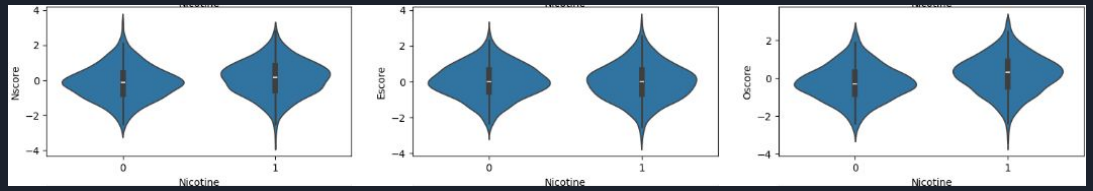
Teste 2:
Escala de 0/1 para todas as drogas

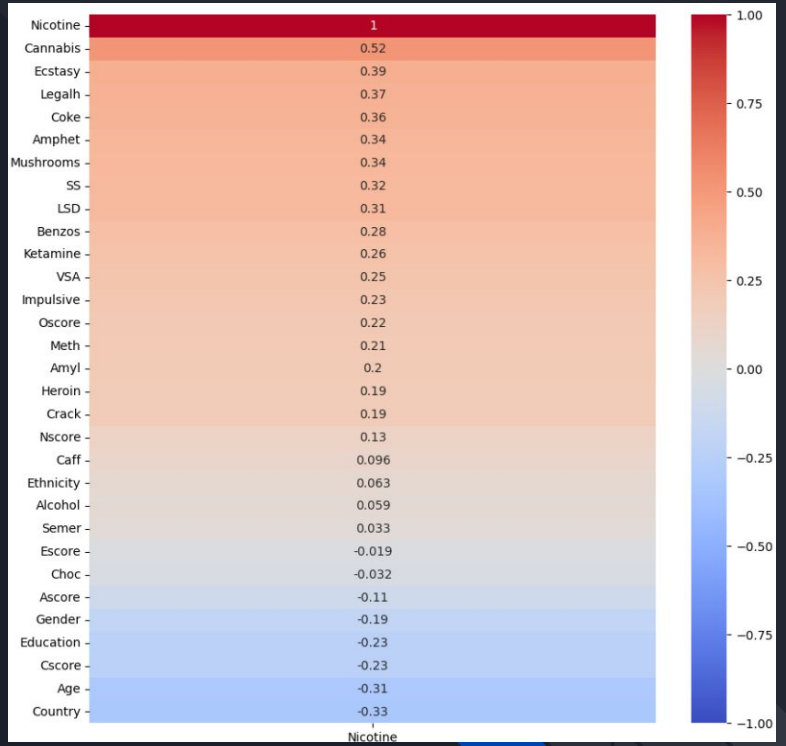
Classifier	Precision	Recall	Accuracy	F1 Score	Training Time
Logistic Regression	0.785047	0.803828	0.769231	0.794326	0.024189
Decision Tree	0.668293	0.655502	0.628647	0.661836	0.007978
Extra Trees	0.757848	0.808612	0.750663	0.782407	0.132066
SVM	0.773756	0.818182	0.766578	0.795349	0.041010
Neural Network	0.750000	0.775120	0.732095	0.762353	2.195397
K-Nearest Neighbors	0.702439	0.688995	0.665782	0.695652	0.000000
Random Forest	0.758772	0.827751	0.758621	0.791762	0.211494
Naive Bayes	0.762195	0.598086	0.673740	0.670241	0.000000

Análise dos dados

Analisando os dados, percebemos que havia pouca correlação entre o consumo de nicotina e as outras variáveis. As características psicológicas apresentaram um coeficiente de correlação muito baixo. Mesmo considerando o consumo de outras drogas, a correlação com a nicotina não era muito elevada. O consumo de cannabis apresentou o coeficiente mais elevado (52%), mas ainda assim um valor baixo.

Da análise dos dados, concluímos também que o dataset não é totalmente equilibrado (209 que consumiram no último ano/168 que não consumiram no último ano).





Modelos desenvolvidos

Estes foram os modelos que utilizamos e os respetivos resultados para uma divisão de 80/20 entre treino e teste, o que nos permitiu comparar algumas métricas de desempenho entre os diferentes modelos.

```
(training_inputs,
 testing_inputs,
 training_classes,
 testing_classes) = train_test_split(features, labels, test_size=0.2, random_state=1)
```

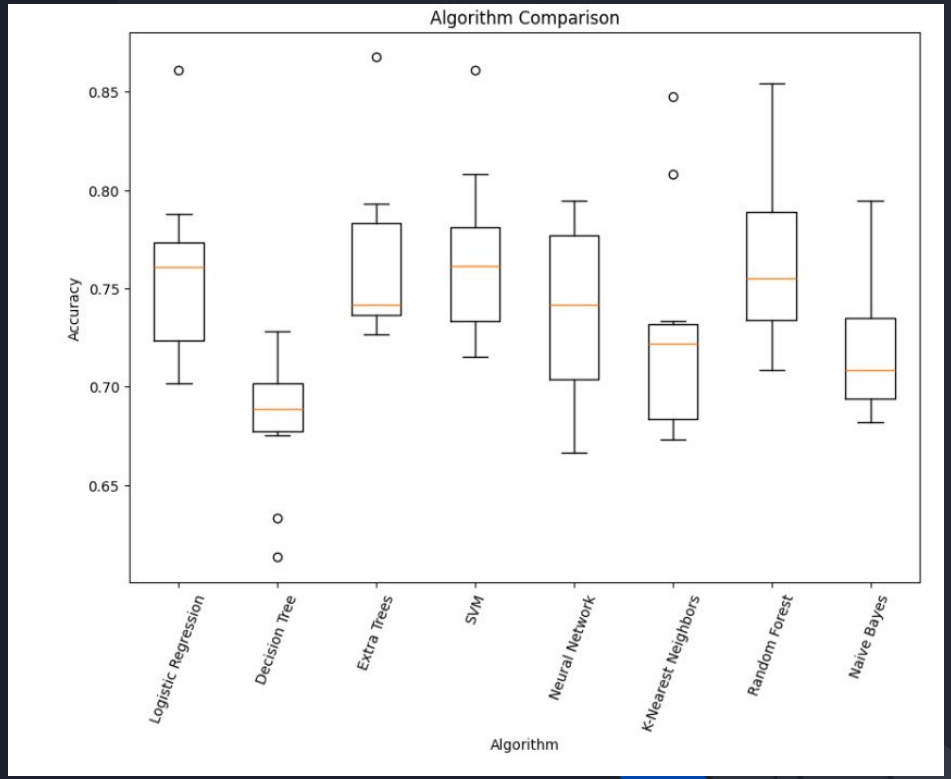
Classifier	Precision	Recall	Accuracy	F1 Score	Training Time
Logistic Regression	0.777778	0.803828	0.763926	0.790588	0.028445
Decision Tree	0.716895	0.751196	0.697613	0.733645	0.008405
Extra Trees	0.758621	0.842105	0.763926	0.798186	0.147902
SVM	0.758772	0.827751	0.758621	0.791762	0.040491
Neural Network	0.733032	0.775120	0.718833	0.753488	1.311188
K-Nearest Neighbors	0.731132	0.741627	0.705570	0.736342	0.000000
Random Forest	0.755459	0.827751	0.755968	0.789954	0.174182
Naive Bayes	0.775862	0.645933	0.700265	0.704961	0.000000

Resultados

O desempenho do modelo pode variar se os conjuntos de treinamento e teste forem diferentes.

Para evitar overfitting (quando o modelo aprende a classificar o conjunto de treino tão bem que não generaliza e não tem um bom desempenho em dados que não viu antes), devemos aplicar K-fold validation de 10-folds nos modelos.

Além disso, para evitar uma proporção desequilibrada entre o número de consumidores de nicotina no último ano e o número de consumidores de nicotina que consumiram pela última vez antes do último ano, utilizamos stratified K-fold validation, mantendo as proporções das classes semelhantes em todas os folds, de acordo com a proporção original do dataset.



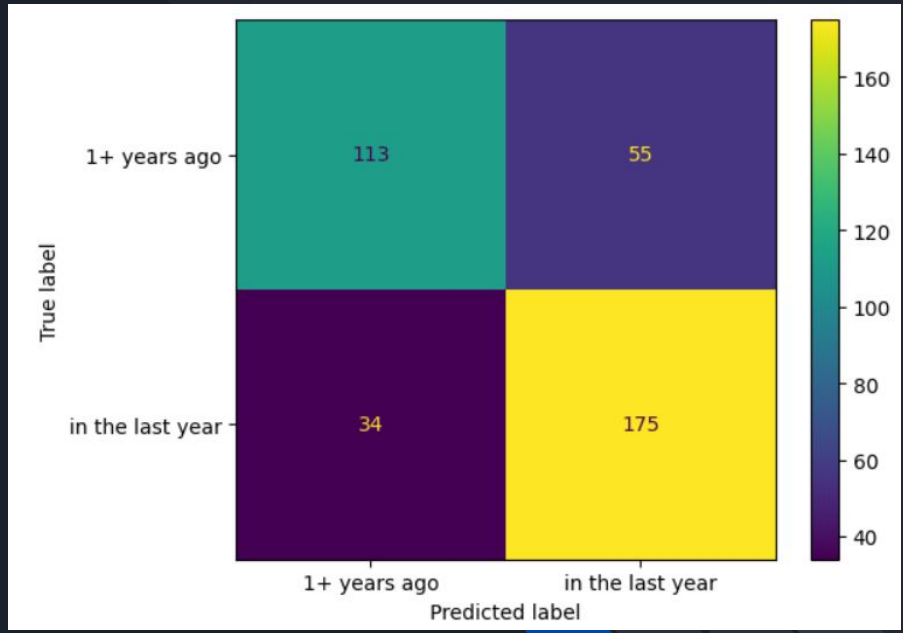
Concluímos que o algoritmo Random Forest tem a melhor exatidão, analisando o diagrama de caixas.

Comparar resultados

Depois de escolher o melhor modelo, utilizámos grid search para selecionar os parâmetros para uma melhor exatidão. (77.6%)

O gráfico da Confusion Matrix é importante, pois evidencia as classificações corretas e incorretas do modelo.

Isso comprova que o modelo apresenta um recall ligeiramente maior do que a precisão, pois existem mais falsos positivos do que falsos negativos.



Conclusões

Em geral, drogas pesadas acabam por ter uma exatidão melhor, pois a maioria das pessoas presentes no dataset nunca experimentou a droga ou não a usa tão recorrentemente, o que leva o modelo a ficar enviesado para prever que a maioria não experimentou (tendência para a classe maior em um dataset não equilibrado). Escolhemos a nicotina para analisar por uma questão de equilíbrio, apesar da menor precisão.

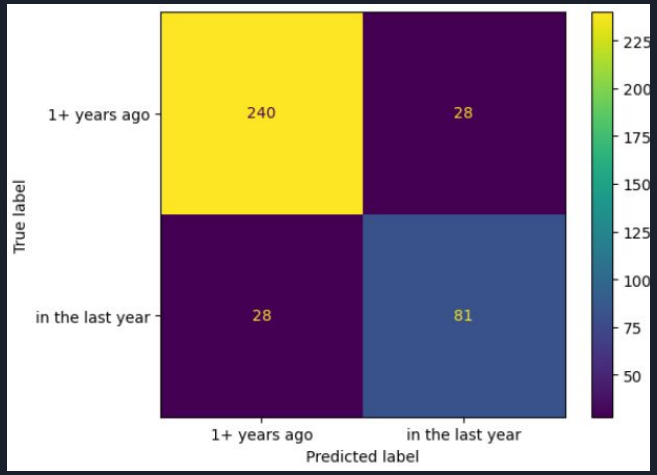
Em geral, as drogas mais comuns (álcool, nicotina, cannabis, etc.) não tem correlações elevadas com outras variáveis, possivelmente por serem consumidas por mais pessoas em geral e serem um pouco “normalizadas” pela sociedade.

Já drogas menos comuns e consideradas pesadas, algumas podem ter correlações ligeiramente mais elevadas com outras variáveis. Após aplicar o mesmo processo para outras drogas para além da nicotina, detectamos uma maior precisão e exatidão, porém acreditamos que isso se pode dever a um desequilíbrio no dataset, levando a uma tendência pela classe maior, como já mencionado.

Em conclusão, admitimos que a exatidão obtida não foi a desejada, mas acreditamos que não existe uma clara correlação entre as features do dataset e o consumo de drogas, o que acaba por justificar os resultados de certa forma.

Ecstasy

Classifier	Precision	Recall	Accuracy	F1 Score	Training Time
Logistic Regression	0.747664	0.733945	0.851459	0.740741	0.770196
Decision Tree	0.631068	0.596330	0.782493	0.613208	0.111318
Extra Trees	0.761905	0.733945	0.856764	0.747664	2.480235
SVM	0.761062	0.788991	0.867374	0.774775	0.647046
Neural Network	0.730435	0.770642	0.851459	0.750000	14.393543
K-Nearest Neighbors	0.712871	0.660550	0.824934	0.685714	0.015667
Random Forest	0.732143	0.752294	0.848806	0.742081	1.569551
Naive Bayes	0.635714	0.816514	0.811671	0.714859	0.023250



Mushrooms

Classifier	Precision	Recall	Accuracy	F1 Score	Training Time
Logistic Regression	0.735632	0.666667	0.854111	0.699454	0.724366
Decision Tree	0.648936	0.635417	0.819629	0.642105	0.059519
Extra Trees	0.797619	0.697917	0.877984	0.744444	2.440984
SVM	0.760870	0.729167	0.872679	0.744681	0.265390
Neural Network	0.694737	0.687500	0.843501	0.691099	12.375015
K-Nearest Neighbors	0.747126	0.677083	0.859416	0.710383	0.162746
Random Forest	0.795181	0.687500	0.875332	0.737430	5.048519
Naive Bayes	0.614173	0.812500	0.822281	0.699552	0.023719