* **Análise do excerto de dados e suposições sobre a atividade diária da “MedKnow”**

O ficheiro disponibilizado pela MedKnow apresenta um conjunto de dados relacionados com a prescrição de lentes oftalmológicas, onde cada registo (instância) representa uma consulta de um paciente. As variáveis disponiveis no conjunto de dados descrevem as principais caracteristicas observadas e recolhidas durante o exame oftalmológico, nomeadamente: a idade dos olhos, o tipo de visão (que seria a doença), a presença ou ausência de astigmatismo, a frequência de lágrimas e, por ultimo, o tipo de lente prescrita pelo médico com base nas variáveis anteriores.

Com base nos dados deste ficheiro assume-se que a atividade diária desta empresa seja composta pelos seguintes pontos:

* Realização de consultas oftalmológicas que servem para recolher dados sobre cada paciente;
* Registo de todos estes dados numa base de dados, permitindo o acompanhamento histórico de cada paciente.
* A análise dos dados recolhidos durante as consultas de maneira a prescrever um tipo de lentes para cada doente.

Posto isto, e com o objetivo de automatizar este ultimo ponto, o objetivo deste projeto é desenvolver um protótipo de um sistema capaz de transformar dados de consultas em conhecimento útil, permitindo predizer um tipo de lentes para um paciente de forma mais rápida e automática. Para este fim, o sistema irá recorrer a técnicas de aprendizagem automática e mineração de dados, treinando com dados existentes para aprender padrões de decisão e, assim, ser capaz de predizer o tipo de lentes com base nas informações recolhidas nas consultas.

* **Modelo de dados (EA)**

Com base na informação que a empresa utiliza uma base de dados relacional que suporta, pelo menos, as entidades *PATIENT*, *DOCTOR*,e *DISEASE*, e os atributos *birthDate* e *diseaseName*, tendo este ultimo três valores possiveis presentes no ficheiro disponibilizado - sendo estes: *myope*, *hypermetrope* e *astigmatic* - e de maneira a suportar as suposições sobre a atividade diária da empresa, desenvolveu-se o seguinte modelo de dados, neste caso Entidade-Associação.

A diagram of a company

AI-generated content may be incorrect.

Figura - Modelo EA desenvolvido

* **Criação e população da base de dados**

A screenshot of a computer program

AI-generated content may be incorrect.

Figura 2 - Ficheiro sql para criação das tabelas

A população das tabelas foi feita com dados predefinidos aleatórios para teste. Apenas a tabela das consultas foi populada em conformidade com o ficheiro disponibilizado pela empresa, ou seja, com as mesmas instâncias associadas a um paciente e doutor aleatórios, tendo ainda uma data de consulta associada.

A screen shot of a computer screen

AI-generated content may be incorrect.

Figura 3 - População da tabela de consultas

* **Exportação de dados**

A exportação de dados foi feita de maneira a ter um formato compativel com o framework do orange, sendo este formato o 3RowHeader.

Foram criadas três *views*, sendo a primeira para identificar a class, denominada lenses\_class, a segunda para determinar o tipo de dados, neste caso todos discretos, denominada lenses\_domain e a terceira para recolher os dados relevantes da consulta, como as caracteristicas do paciente e as lentes escolhidas.

Após a criação dessas três views, foram unidas em apenas uma view denominada lenses\_dataset que estará no formato 3RowHeader pronta para ser exportada.

A screenshot of a computer program

AI-generated content may be incorrect.

Figura 4 – Criação das views que compoêm o 3RowHeader

A computer screen with white text

AI-generated content may be incorrect.

Figura 5 - Criação da view com o formato 3RowHeader que será exportada

A computer screen with text on it

AI-generated content may be incorrect.

Figura - Exportação da view

A screen shot of a computer

AI-generated content may be incorrect.

Figura - Lenses\_dataset view

* **Testes**

Os testes apresentados a seguir, tanto com o OneRule, como com o ID3 e o Naive-Bayes, foram realizados utilizando o dataset apresentado na figura 7, separado em duas partes – uma maior, para treino e uma menor para testes, sendo estas as apresentadas nas tabelas 1 e 2.

Tabela - dataset de treino

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| age | prescription | astigmatic | tear\_rate | lenses |
| discrete | discrete | discrete | discrete | discrete |
|  |  |  |  | class |
| young | hypermetrope | yes | reduced | none |
| young | hypermetrope | no | normal | soft |
| young | hypermetrope | no | reduced | none |
| presbyopic | hypermetrope | yes | reduced | none |
| presbyopic | hypermetrope | yes | normal | none |
| presbyopic | hypermetrope | no | normal | soft |
| presbyopic | hypermetrope | no | reduced | none |
| pre-presbyopic | myope | yes | normal | hard |
| pre-presbyopic | hypermetrope | yes | normal | none |
| pre-presbyopic | hypermetrope | no | normal | soft |

Tabela - dataset de teste

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| age | prescription | astigmatic | tear\_rate | lenses |
| discrete | discrete | discrete | discrete | discrete |
|  |  |  |  | class |
| young | myope | yes | normal | hard |
| young | hypermetrope | yes | normal | hard |
| pre-presbyopic | hypermetrope | no | normal | soft |
| presbyopic | hypermetrope | no | reduced | none |
| presbyopic | myope | no | reduced | none |
| pre-presbyopic | myope | yes | normal | hard |

Tabela - Resultado do 1R

|  |  |  |
| --- | --- | --- |
| Atributo | Percentagem de erro | Atributo escolhido |
| Age | 0.4% |  |
| Precription | 0.3% | x |
| Astigmatica | 0.3% |  |
| Tear\_rate | 0.3% |  |

A computer screen shot of a black screen

AI-generated content may be incorrect.

Figura 8 - Resultado do 1R

Após implementar e testar o onerule, testou-se o ID3 e o Naive-Bayes, para comparar as três percentagens de erro.

A screenshot of a computer program

AI-generated content may be incorrect.

Figura 9 - Resultado do ID3 e do Naive-Bayes

Tabela 4 - Percentagem de erro dos classificadores

|  |  |
| --- | --- |
| Algoritmo | Percentagem de erro |
| 1R | 66.67% |
| ID3 | 33.33% |
| Naïve-Bayes | 33.33% |

Como podemos ver pela tabela 4, o 1R é o que apresenta piores resultados, ou seja, uma percentagem de erro maior, em comparação com os outros dois algoritmos. Isto acontece porque o 1R não tem muitas regras de classificação, e não deve ser utilizado diretamente neste tipo de treino e testes, apesar deste algoritmo ser muito eficiente noutro tipo de aplicações como para encontrar as melhores variaveis (que contêm informação mais valiosa) num dataset.

Após realizarmos estes testes, tentámos melhorar o treino, utilizando o Leave-One-Out, que é técnica de validação cruzada. Esta tecnica é um caso especial da validação cruzada k-fold, onde o numero de folds k é igual ao número total de instancias do conjunto de dados.

Esta ténica funiona da seguinte maneira:

* Em cada iteração, é deixada de fora uma instancia do dataset.
* Usa as n – 1 instancias restantes como conjunto de treino.
* Treina o modelo e testa com a instancia deixada de fora, fazendo uma previsão.
* No final, calcula-se o erro médio sobre todas as previsões.

Assim sendo, já não precisamos de segmentar o dataset e utilizamos o dataset total, apresentado na figura 7.

Utilizou-se esta técnica em todos os algoritmos (1R, ID3 e Naive-Bayes) e obteve-se os resultados apresentados na figura 10.

**Nota:** Até este ponto tinhamos utilizado o orange framework para o ID3 e o Naive-Bayes, mas quando se implementou o Leave-One-Out, começámos a utilizar a biblioteca scikit-learn, que contem estes algoritmos, uma vez que, com esta biblioteca é mais simples preparar as regras para exportar para um ficheiro e podermos utilizar o mesmo num modelo de previsão sem precisarmos de estar sempre a treinar.

A computer screen with white text

AI-generated content may be incorrect.

Figura 10 - Erro dos algoritmos com o leave one out

Como o dataset é muito reduzido (poucas instancias), é de se esperar que os resultados não sejam os mais realistas. Posto isto tentou-se utilizar outro dataset, com umas instancias a mais, para ver o efeito que teria nos resultados. Os resultados obtidos foram os apresentados na figura 11.

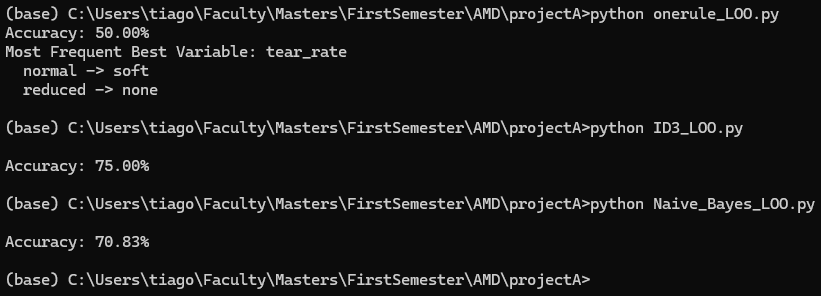


Figura 11 - Resultados com novo dataset

Após realizar todos estes testes, implementámos uma aplicação em python que recebe os dados de um novo paciente e prediz com base no treino que fez com o algoritmo 1R e outra aplicação que prediz com base no treino que fez com o algoritmo ID3 ou Naive-Bayes tendo em conta a escolha do utilizador.

Para isto funcionar, foi necessário guardar as regras de predição que foram geradas a partir do treino com os algoritmos em ficheiros, para poderem ser utilizadas na aplicação.

Como o algoritmo 1R é o mais simples, apenas foi necessário guardar num ficheiro a melhor variavel e as respetivas regras.

Já para os restantes algoritmos, optou-se por guardar as regras num ficheiro .pkl, utilizando a biblioteca joblib presente no python, que depois será utilizado na aplicação para predizer um tipo de lente para o paciente.

Todos os scripts .py estão documentados, comentando todas as opções feitas e explicando o código na generalidade.

Abaixo, nas figura 12 e 13, são demonstradas as aplicações em funcionamento.

A screen shot of a computer code

AI-generated content may be incorrect.

Figura 12 - Aplicação de predição utilizando o 1R

A computer screen with white text

AI-generated content may be incorrect.

Figura 13 - Aplicação de predição utilizando o ID3 ou o Naive-Bayes

Como podemos ver, o 1R apresentou um resultado diferente dos outros algoritmos mas, tendo em conta as percentagens de erro apresentadas na figura 11, podemos confiar mais facilmente no resultado que nos deu os algoritmos id3 e naive-bayes.