

Aprendizagem e Mineração de Dados

Project A/A1

Mihail Ababii, 46435@alunos.isel.ipl.pt

20 dezembro de 2021

Project A

1 Analise dos dados

A MedKnow ao longo do tempo armazena a informação dos seus pacientes numa base de dados, pelo que o objectivo actual da equipa de oftalmologia, da MedKnow, baseia-se na análise dos dados relativos aos olhos dos pacientes, com o objectivo de encontrar padrões que evidenciem a necessidade de o paciente usar de lentes e a respectiva graduação. Esta análise tem em consideração dados armazenados, como por exemplo a idade do paciente, o teor das lágrimas, possíveis doenças como a miopia, a hipermetropia ou o estigmático. Os dados armazenados contêm os dados do medico, do paciente e dos problemas de visão que este possa ter. Tendo em consideração o bem-estar do paciente é importante que durante a análise se encontrem os padrões que melhor representem o estado de visão do respectivo paciente, pois uma prescrição errada pode piorar o estado do mesmo.

2 Estrutura da Base de dado Relacional

Tendo em consideração os dados apresentados pela empresa MedKnow foram criados os seguintes modelos.

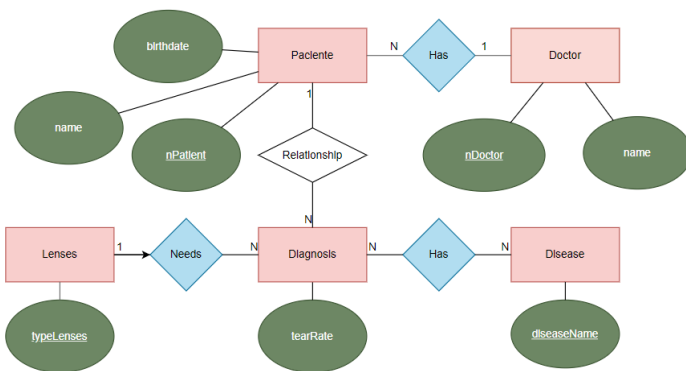


Figura 1: Entety Relationship

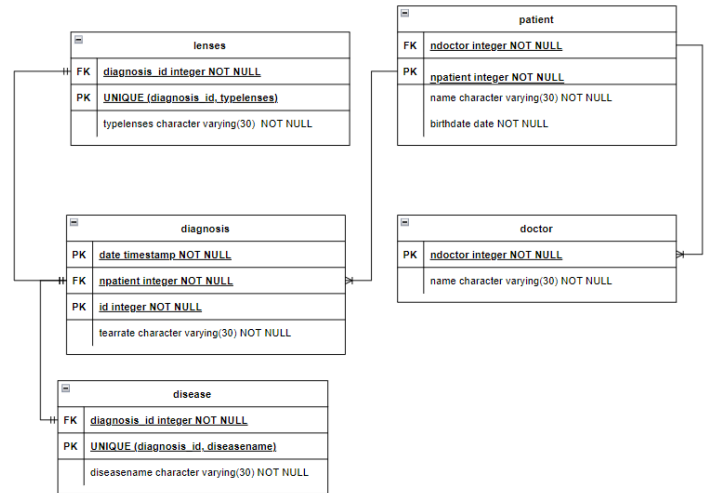


Figura 2: Entety Relationship - Modelo Conceptual

Ambos os modelos são idêntico, sendo meramente representações diferentes da estrutura de dados utilizada.

Considera-se que cada cliente está associado a 1 medico, ainda que este possa ter múltiplos pacientes. A MedKnow permite que um paciente possa realizar diversas consultas, pelo que o estado de saúde do mesmo é guardado ao longo do tempo. Em cada diagnostico é verificado o teor de lágrimas e as doenças que o doente possa ter, com o objectivo de encontrar o tipo de lentes que melhor se apropriem ao paciente. Em cada momento o doutor pode saber o estado de saúde actual do cliente observando os resultados da ultima consulta realizada. Neste modelo optou-se por guardar todos os dados obtidos em cada consulta de modo a permitir observar o progresso do paciente.

3 Criação da base de dados através de scripts

Neste passo foram criados diversos scripts, de modo possibilitar a simulação de uma situação real, permitindo trabalhar com uma base de dados e as respectivas funcionalidade. Numa primeira fase é criado uma base de dados vazia, por modo a possibilitar a criação das tabelas, sendo que estas implementam as tabelas seguindo os modelos das figuras 1 e 2. É também importante popular as tabelas criadas de modo a possibilitar o uso da estrutura de dados criada, assim foram inserido dados de maneira a estes representarem os dados atribuídos pelo professor no documento *d01_lenses.xls* Para finalizar este passo do projecto realizou-se um script que exporta os dados armazenados na base de dados.

age	prescription	astigmatic	tear_rate	lenses
young	myope	yes	normal	hard
young	myope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	no	reduced	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	hypermetrope	yes	normal	none
pre-presbyopic	hypermetrope	no	normal	soft

Figura 3: Documento d01.lenses.xls

	A	B	C	D	E	F
1	patient_number	age	tear_rate	prescription	astigmatic	lenses
2	discrete	discrete	discrete	discrete	discrete	discrete
3	ignore					class
4		1 young	normal	myope	yes	hard
5		2 young	normal	myope	no	soft
6		3 young	reduced	hypermetrope	yes	none
7		4 young	normal	hypermetrope	no	soft
8		5 young	reduced	hypermetrope	no	none
9		6 presbyopic	reduced	myope	yes	none
10		7 presbyopic	normal	myope	yes	hard
11		8 presbyopic	reduced	hypermetrope	yes	none
12		9 presbyopic	normal	hypermetrope	yes	none
13		10 presbyopic	normal	hypermetrope	no	soft
14		11 presbyopic	reduced	hypermetrope	no	none
15		12 pre-presbyopic	reduced	hypermetrope	yes	none
16		13 pre-presbyopic	normal	myope	yes	hard
17		14 pre-presbyopic	normal	myope	no	soft
18		15 pre-presbyopic	normal	hypermetrope	yes	none
19		16 pre-presbyopic	normal	hypermetrope	no	soft

Figura 4: Dados exportados

Como podemos observar os dados exportados representam os todos dados originais, com adição do numero do paciente. Podemos também observar que foram adicionados cabeçalhos, tendo estes o propósito de possibilitar o uso dos dados e fases seguintes do desenvolvimento do projecto. Tal como anteriormente o cabeçalho ainda contem o nome das variáveis utilizadas, no entanto temos dois novos elementos no cabeçalho, sendo o segundo utilizado para indicar o tipo de dados que as variáveis poderiam ser e o terceiro tem o propósito de indicar algumas flags como colunas a ignorar ou indicar a classe.

4 Metodo 1R

O método 1R é um método bastante simplista de tratamento de dados, que tem em consideração apenas um dos atributos para obter a classe mais adequada, sendo este o mais o que apresenta maior informação. Este método cria para cada atributo uma matriz de contingência, que é uma matriz que contem a frequência dos diversos valores possíveis na matriz tendo em consideração a classe, que será utilizada para calcular a probabilidade de erro da precisão dos valores. Com os erros anteriormente calculados é seleccionado o atributo que apresenta o menor erro, que será utilizado para criar uma matriz confusão.

```
>> Contingency Matrix <<
age & lenses :
[[1. 2. 2.]
 [0. 3. 1.]
 [0. 1. 1.]]
erro total: 0.45454545454545453

tear_rate & lenses :
[[1. 2. 4.]
 [0. 4. 0.]]
erro total: 0.2727272727272727

prescription & lenses :
[[0. 5. 3.]
 [1. 1. 1.]]
erro total: 0.45454545454545453

astigmatic & lenses :
[[0. 1. 4.]
 [1. 5. 0.]]
erro total: 0.18181818181818182
```

Figura 5: Matrizes de Contingências

Como podemos observar o neste caso a matriz de contingência que apresenta o menor erro é a matriz *Astigma & Lenses* assim sendo, foi criada uma matriz de erro para o atributo seleccionado, figura 6. Assim sendo a partir da matriz de erro é seleccionado, para cada valor do atributo, a classe com menor erro, obtendo assim as regras da figura 7

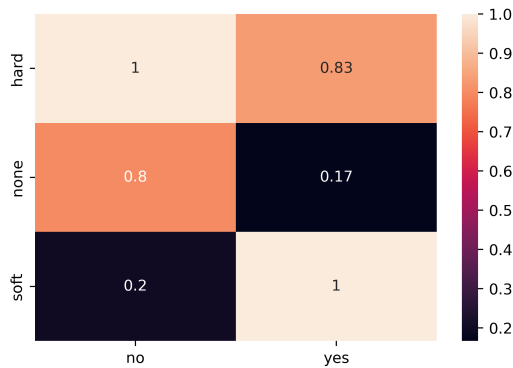


Figura 6: Matriz Erro 1R

```
--> so, the rule, and error, for the astigmatic feature are:
(astigmatic, no, soft) : 0.200
(astigmatic, yes, none) : 0.167
```

Figura 7: Regras 1R

Posteriormente a partir matiz de erro obtemos as regras a utilizar no deployment da aplicação. Como podemos observar no exemplo da figura 7, as regras são constituídas pelo atributo e os respectivos valores, bem como a classe que será predita. A componente deploy neste momento está a ser realizada assim que a regra é obtida, sendo que esta não será actualizadas com a introdução de novos dados na base de dados.

5 Metodo ID3 e Naive Bayes

O desenvolvimento deste exercício foi bastante similar ao método 1R, mudando apenas o Métodos utilizados. Devido á mudança de métodos foi alterado também o código, por forma a suportar a leitura de dados do tipo *.csv*. Esta alteração deve-se ao facto destes métodos apresentarem um código mais simplista para este tipo de dados. Neste modulo ambos os modelos são treinados em simultâneo, pois em tudo São idênticos. A lógica do código mantém-se do 1R, ou seja, os dados são divididos em dados de treino e em dados de teste, os modelos são treinados e testados e por fim é realizado o deploy. Estes modelos são um modelo mais complexos que o 1R, pois têm em consideração múltiplos atributos dos dados, pelo que é esperado que as métricas de teste seja consideravelmente melhores.

5.a ID3

O modelo ID3 é um modelo que forma uma *Decision Tree*, com base nos diversos atributos. Ainda que seja mais complexo, este apresenta uma lógica consideravelmente mais simplista, pois este apenas cria um nó para cada atributo e calcula a respectiva entropia (incerteza) seleccionando o nó com maior ganho de informação e os respectivos ramos utilizando o mesmo processo para ir construindo a árvore, até que se converja ou que se chegue a um dado threshold. Este processo é realizado utilizando uma estratégia greedy, pelo que os cálculos realizados são todos, no scope local, ou seja sem considerar nós futuros ou passados.

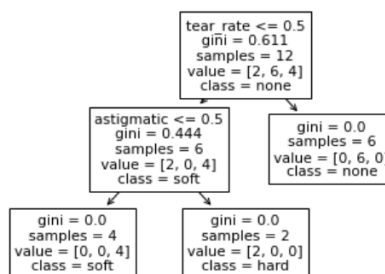


Figura 8: Decision tree Example

Um problema deste modelo é a facilidade de acontecer overfitting ao set de treino, assim sendo este necessita de ser limitado, colocando um limite maximo de niveis de ramificação ou retirando alguns ramos apos a criação da arvore completa.

5.b Naive Bayes

O método Naive Bayes é um método que considera que todos os atributos são independentes e que apresentam a mesma importância na predição da classe. Motivo pelo qual este calcula a probabilidade condicional $P(A|B)$ para todos os atributos, ou seja a probabilidade de A sabendo que B. Assim, a probabilidade de uma dada classe dá-se pelo somatório das diversas probabilidade condicionais nas quais esta foi utilizada. Um problema deste método deve-se ao facto de todos atributos serem contabilizados como tendo a mesma importância, o que muitas vezes não é o caso. Como este método considera todos os dados estes têm que ser filtrados antes de serem utilizados.

6 Métodos de teste

No método 1R não foi possível utilizar métricas de teste, no entanto é de esperar que estas não sejam elevadas pois só é tido em consideração um atributo, para realizar a predição da classe, que não é muito viável, visto que cada vez mais nos deparamos com problemas mais e mais complexos.

Em contrapartida ao utilizar os métodos ID3 e Naive Bayes, foi possível utilizar as diversas avaliações. Para realizar estes testes utilizou-se um StratifiedShuffleSplit, com tamanho de 20% para teste, pelo que se obteve resultados, predominantemente positivos. Um resultado positivo é um resultado que apresente uma baixa elevada taxa de erro, sendo possível observar que a variação máxima do erro dificilmente ultrapassa os 20%. As métricas utilizadas foram o accuracy_score, o precision_score, recall_score, f1_score e cohen_kappa_score.

```
..VW-----id3-----VW..
::accuracy_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 100.00% | 50.00% | 100.00% | 100.00% | 50.00% | 100.00% | 100.00% | 75.00% |
87.50% (+/- 20.16%)
::precision_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 100.00% | 50.00% | 100.00% | 100.00% | 100.00% | 87.50% | 100.00% | 33.33% |
87.08% (+/- 23.31%)
::recall_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 75.00% | 100.00% | 100.00% | 50.00% | 50.00% | 100.00% | 100.00% | 100.00% |
87.50% (+/- 20.16%)
::f1_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 37.50% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
93.75% (+/- 18.75%)
::cohen_kappa_score::
::all-evaluated-datasets::
 20.00% | 63.64% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 33.33% | 33.33% |
75.03% (+/- 32.21%)
```

Figura 9: ID3

Tal como previsto os resultados são bastante positivos, pois quase todos os resultados foram 100%, podendo observar certas variações ocasionais. Tal como mencionado anteriormente, o método ID3 é bastante propenso a overfitting, e devido ao baixo número de dados, este pode ser um dos motivos pelo qual os resultados do score apresentam variações, ocasionalmente.

```

..VV-----naive bayes-----VV..
::accuracy_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 75.00% | 100.00% | 100.00% | 100.00% | 100.00% |
97.50% (+/- 7.50%)
::precision_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 87.50% | 33.33% | 100.00% | 87.50% | 100.00% | 100.00% | 100.00% | 87.50% |
89.58% (+/- 19.57%)
::recall_score::
::all-evaluated-datasets::
 75.00% | 75.00% | 50.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 75.00% |
87.50% (+/- 16.77%)
::f1_score::
::all-evaluated-datasets::
 100.00% | 100.00% | 75.00% | 75.00% | 100.00% | 75.00% | 100.00% | 100.00% | 100.00% | 100.00% |
92.50% (+/- 11.46%)
::cohen_kappa_score::
::all-evaluated-datasets::
 100.00% | 63.64% | 100.00% | 100.00% | 63.64% | 63.64% | 33.33% | 100.00% | 63.64% | 63.64% |
75.15% (+/- 22.09%)

```

Figura 10: Naive bayes

Tal como no ID3 é possível observar que os resultados obtidos são bastante positivos, no entanto também é possível observar avaliações inferiores a 100%. Isto pode ser justificado pelo facto que alguns dos atributos poderem ser mais relevantes que outros ou pelo facto de no código realizado não ser feita a selecção dos atributos mais relevantes.

Outro factor que pode influenciar os resultados obtidos, pode ser a divisão dos dados em dados de treino e dados de teste, pois sendo este um dataset relativamente pequeno, os dados podem não ser devidamente representados nos dados de treino.

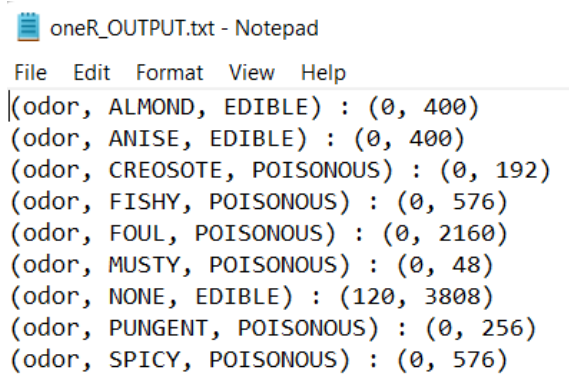
Project A1

7 Transformação do dataset

O objectivo deste exercício é criar um programa que transforme um ficheiro `.csv` num ficheiro `.tab`, sendo a maior diferença a divisão entre os atributos sendo que ficheiros `.csv` dividem os dados através de virgulas enquanto o `.tab` separa através de um `"tab"`. É também importante referir que o documento `.csv` não está definido seguindo a estrutura de três headers, e que a classe está definida com o nome `"class"`. Assim sendo para além de alterar o delimitados foi também necessário adicionar o header de tipos e o header de flags.

8 Output 1R

Para realização deste exercício, foi necessário alterar o código previamente realizado, ainda que seja apenas a adição da escrita no ficheiro.



```
oneR_OUTPUT.txt - Notepad
File Edit Format View Help
(odor, ALMOND, EDIBLE) : (0, 400)
(odor, ANISE, EDIBLE) : (0, 400)
(odor, CREOSOTE, POISONOUS) : (0, 192)
(odor, FISHY, POISONOUS) : (0, 576)
(odor, FOUL, POISONOUS) : (0, 2160)
(odor, MUSTY, POISONOUS) : (0, 48)
(odor, NONE, EDIBLE) : (120, 3808)
(odor, PUNGENT, POISONOUS) : (0, 256)
(odor, SPICY, POISONOUS) : (0, 576)
```

Figura 11: ID3

Tal como podemos ver, a regra 1R foi bem seleccionada pois, à excepção de não existir odor, a probabilidade de erro é 0, para todos os outros valores de odor.

9 Orange Canvas

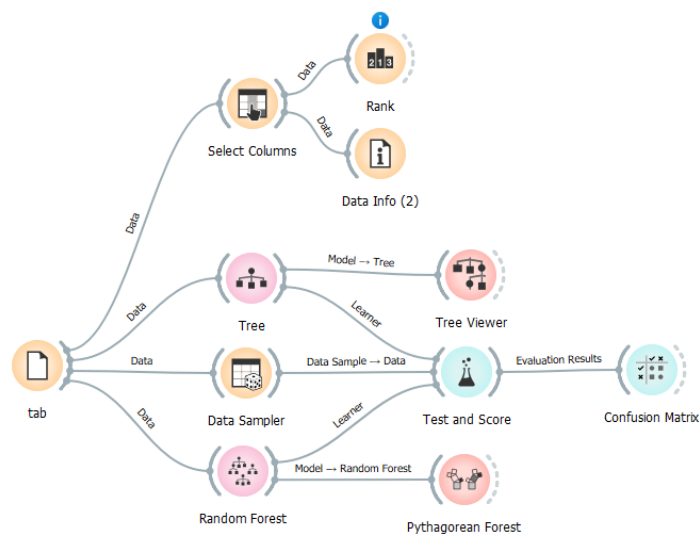


Figura 12: Orange

Test and Score						
Sampling		Evaluation Results				
<input type="radio"/> Cross validation		Model	\hat{AUC}	CA	F1	Precision
Number of folds: 5		asdasd	0.999	0.999	0.999	0.999
<input checked="" type="checkbox"/> Stratified		Random Forest	1.000	0.999	0.999	0.999
<input type="radio"/> Cross validation by feature						

Figura 17: Orange

		Predicted		
		EDIBLE	POISONOUS	Σ
Actual	EDIBLE	28722	8	28730
	POISONOUS	56	25084	25140
Σ		28778	25092	53870

Figura 18: Tree Error Matrix

		Predicted		
		EDIBLE	POISONOUS	Σ
Actual	EDIBLE	28727	3	28730
	POISONOUS	29	25111	25140
Σ		28756	25114	53870

Figura 19: Forest Error Matrix

Tal como podemos observar estes métodos apresentam resultados bastante bons, aproximando-se bastante de 1. Em adição apresentamos a matriz de erro em ambas as situações e podemos observar que em ambas as situações o numero de predições erradas não chega aos 100, o que é bastante bom visto que foram realizadas 53870 predições.