ISEL
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Instituto Superior de Engenharia de Lisboa

Final Project A1

Aprendizagem e Mineração de Dados

Mestrado em Engenharia Informática e de Computadores

Docente: Paulo Trigo

Grupo 25

Nuno Gomes – 44021

Índice

Introdução

Um trabalho de análise de dados pode tomar várias vertentes, devido à inerente complexidade de extrair conhecimento a partir de dados. A vertente tomada neste trabalho foca-se na classificação dos dados, com recurso à métodos de classificação de dados. O objetivo desta classificação é extrair conhecimento proveniente de similaridades existentes nos dados a analisar, na qual se consegue dizer com pouca margem de dúvida que certos dados influenciam outros.

Há vários métodos de classificação de dados, mas este trabalho foca-se maioritariamente nos seguintes:1R, ID3 e Naive Bayes. Para todas estas classificações são usados dados formatados de forma a que os dados estejam organizados tuplos, que por sua vez têm os dados organizados por colunas com o seu significado.

No método 1R, estas colunas são divididas em atributos e uma classe. A classe representa o seu valor de conceito que é o que pode ser inferido através das outras colunas. As outras colunas são chamadas de atributos. Este método permite descobrir a correlação entre cada atributo e a classe, e calcular qual é a correlação entre ambos. Este método termina com a escolha do atributo que tem menos erro.

O ID3, também chamado de árvore de decisão, baseia a sua classificação em estruturas em árvore. Para executar esta classificação começa-se com um atributo como nó raiz, após isso cria-se um nó para as instâncias do atributo, e escolhe-se outro atributo e repete-se para todos os nós até todas as instâncias apresentarem a mesma classe. Variando o nó raiz podemos gerar tantas árvores quanto atributos. A forma de escolha é escolhendo a árvore que representa um maior ganho de informação, e apresenta menos entropia.

Por fim o Naive Bayes também utiliza os conceitos de classe e atributos assim como é assumido que todos os atributos são independentes entre si, mesmo que não o sejam. Este método assenta fortemente nas frequências das instâncias de um atributo. Após o cálculo de todas as frequências é aplicada a formulação de Bayes, que permite identificar qual a maior verosimilhança entre um atributo e a classe.

Para utilizar estes métodos são necessários dados, e a forma como estes são apresentados aos métodos é importante, pois impacta a integridade dos modelos. Por esta razão existem métodos para lidar com um dataset de forma a ter os melhores resultados dadas as circunstâncias, sendo que um dataset tem de ser utilizado para treinar e testar um modelo.

Existem várias formas de lidar com o dataset para estas operações, aqui são exclusivamente apresentados os métodos presentes no Orange3.

- Cross Validation – Define um número de partições nos dados, que é maior que dois. Estas partições são iteradas um certo número de vezes tanto pelo treino como pelo teste.
- Random Sampling – Retirar dados do dataset aleatoriamente. Pode trazer problemas de gerar números muito próximos e tornar a amostra não ideal.
- Leave One out – Idêntico ao Cross Validation, no entanto uma das partições é sempre deixada de fora em todas as iterações. É um processo determinístico.

Todas estas opções se aplicam a diferentes casos, e não impedem overfitting de acontecer. Overfitting é o conceito de um modelo ser demasiado treinado num modelo, não funcionando bem num caso geral, e funcionando extremamente bem nesse modelo em particular. Por esta razão é necessário conseguir avaliar um modelo, e para isso utilizam-se métricas de avaliação.

Neste trabalho focamo-nos maioritariamente nas quatro métricas que são a Accuracy, Precision, Recall e F1. Estas métricas avaliam coisas distintas, sendo que todas são úteis para determinar tais coisas. A Accuracy, ou taxa de sucesso serve para medir a proporção de classificações corretas de todos os dados classificados. A Precision indica a proporção de positivos corretamente previstos de todos os positivos previstos, tendo em conta que podem existir falsos positivos. O Recall reflete a proporção dos positivos previstos dentro de todos os positivos existentes no dataset. Por fim o F1, demonstra o equilíbrio entre a Precision e o Recall como uma média harmónica dos dois.

Como método de validação escolhemos utilizar a matriz de confusão, que permite comparar os valores previstos com os valores reais de dada classe. Assim sendo com uma avaliação das percentagens esperadas podemos avaliar se um método é melhor que outro ou não.

Análise de dados

No documento que nos foi providenciado para analisar (d01_lenses.xls), analisámos os dados para inferir qual o trabalho diário realizado nesta empresa. Com base na nossa análise, compreendemos que o trabalho se foca maioritariamente na área da oftalmologia, e conseguimos tirar as seguintes conclusões mais específicas:

- Medição de indicadores de saúde ocular, de forma a posteriormente auxiliar uma escolha mais informada de um tratamento, no caso de ser necessário.

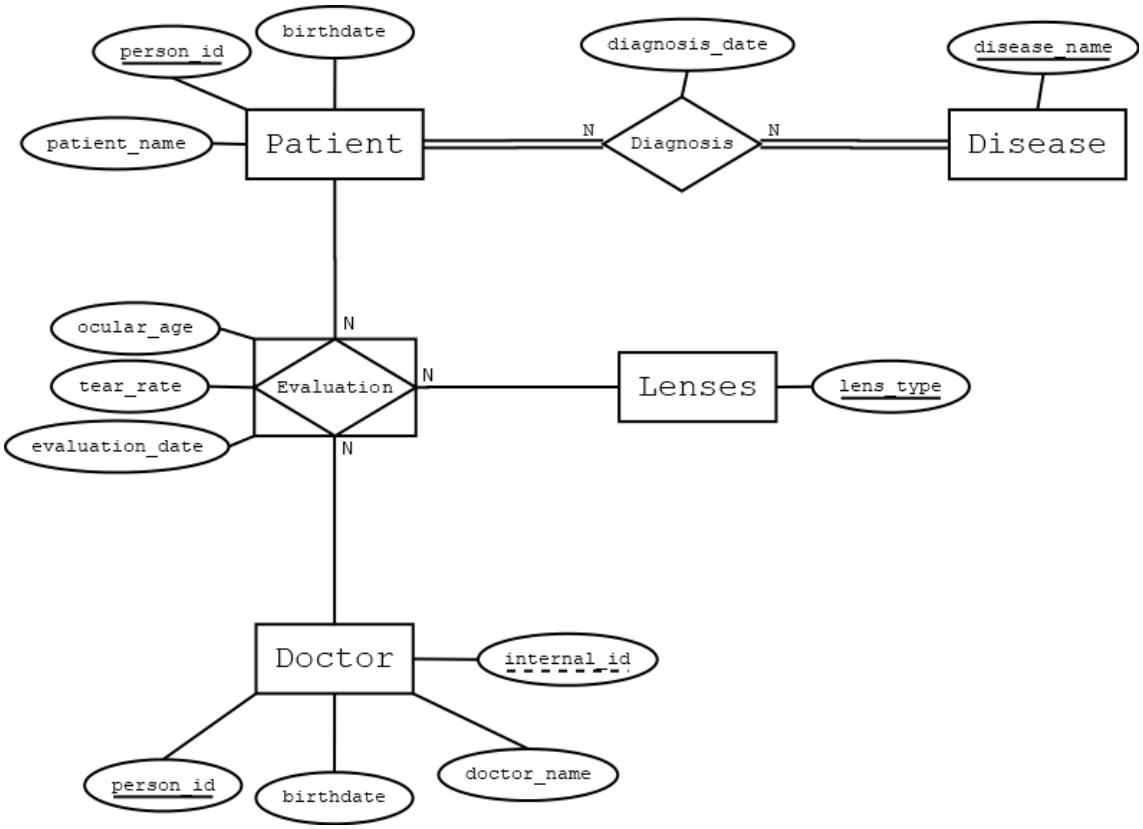
- Diagnóstico de doenças oculares, especificamente as três doenças oculares mais comuns no geral da população que são hipermetropia, miopia e astigmatismo.

- Elaboração de tratamento. No caso de ser diagnosticada uma doença ocular a um paciente será proposto um tratamento. O tratamento atribuído serão lentes oculares para correção de visão. Estas são identificadas pela sua rigidez, das quais as mais populares são as rígidas e as moles.

Foram feitas algumas assunções em relação aos dados.

- Sempre que há contacto entre um médico da Medknow e um paciente, é gerada uma avaliação do paciente com os dados atualizados das suas condições oftalmológicas.
- Um paciente pode ser diagnosticado com mais que uma doença.
- É assumido que as idades oculares correspondem a: jovem, desde os 0 aos 34 anos; pré-presbiópico dos 35 aos 44 anos; presbiópico com mais de 45 anos.
- Foi assumido que há mais de três tipos de doença, bem como mais de dois tipos de lente
- Não há uma relação de acompanhamento entre médico e doente explicita.

Modelo EA


person_id
birthdate
patient_name
Patient
diagnosis_date
N
Diagnosis
N
disease_name
Disease
ocular_age
tear_rate
evaluation_date
N
Evaluation
N
N
Lenses
lens_type
Doctor
internal_id
person_id
birthdate
doctor_name

Modelo Relacional

Legenda

CE = Chaves Estrangeiras
CC = Chaves Candidatas
RI = Regras de Integração

Tabelas

***Patient* (person_id, birthdate, patient_Name)**

RI: birth_date é escrito na forma 'yyyy-mm-dd'

***Disease* (disease_name)**

***Diagnosis* (patient_id, disease_name, diagnosis_date)**

CE: (patient_id) referencia *Patient*(person_id),
(disease_name referencia *Disease*(disease_name)

***Doctor* (person_id, internal_id, birthdate, doctor_name)**

CC: internal_id
RI: internal_id é um valor único.

***Lenses*(lens_type)**

***Evaluation* (doctor_id,patient_id,lens_type, ocular_age, tear_rate, evaluation_date)**

CE: (patient_id) referencia *Patient*(person_id),
(doctor_id) referencia *Doctor*(doctor_id),
(lens_type) referencia *Lenses*(lens_type)
RI: ocular_age apenas toma por valores [young, pre-presbyopic, presbyopic]

Regras de Negócio:

O atributo person_id utilizado nas tabelas *Doctor* e *Person* representa o número de identificação pessoal de cada uma destas pessoas. Na tabela *Doctor*, o internal_id é o número do médico dentro da empresa "Medknow".
Na tabela *Lenses*, o atributo lens_type foi intencionalmente deixado com os valores em aberto, pois após uma pesquisa foi descoberto que há mais tipos de lentes do que aqueles utilizados nos dados que nos foram passados.
Por fim, a tabela *Evaluation*, não costringe o valor de tear_rate a [normal, reduced], pois o grupo também acha que há mais valores de tear_rate que possam ser admitidos.

Procedimento Experimental

Projeto A

Doravante, este trabalho foca-se maioritariamente nos exercícios 6, 7, 8 e 9 do trabalho, devido a serem os que ainda não foram aboradados.

Exercício 6

Para este exercício fizemos uma implementação no IDE PyCharm recorrendo à linguagem Python para modelar a regra 1R. Esta implementação foca-se em descobrir o atributo que possui uma melhor precisão para identificar a classe do dataset recebido. Por fim apresentamos qual o melhor atributo para tal.

Exercício 7

Esta implementação foi feita com recurso ao Orange3. Para tal apenas utilizámos o atributo apresentado anteriormente na implementação do 1R, que foi a tear_rate. Após isso foram alimentados esses dados ao previsor constante que tendo em conta esse atributo e a classe previu as saídas com base nesse valor.

Exercício 8

Esta implementação foi feita com recurso ao Orange3. Este foi bastante mais simples porque o Orange implementa estes métodos. No ID3, o Orange3 escolhe automaticamente a árvore que apresenta uma menor entropia e que assume ser a melhor. Após isso utilizámos o widget de Naive Bayes que também apresentou automaticamente a solução.

Escolha de método de avaliação

O modelo escolhido foi o leave-one-out. Esta decisão foi feita pois pareceu o método de avaliação mais completo, pois utiliza a maior quantidade de dados possível para treino, pelo que consegue avaliar bem o desempenho do método de classificação. Para além disso, o método é determinístico garantindo resultados consistentes, que é importante para fazer comparação, tendo em conta que o aleatório apresentar um "bias" pode ser problemático.

Projeto A1

Exercício 1

O dataset utilizado para este projeto é relacionado com cogumelos, ao contrário do anterior, que era relacionado com lentes. Através da análise de dados percebemos que este possui valores omissos, ao contrário do anterior. Para além disso possui 8416 instâncias e todos os seus atributos e classe são de domínio discreto, que abre a porta à utilização de algoritmos sem restrições por essa razão.

Exercício 2

Para este exercício foi feita uma solução em Python, que fosse capaz de receber um ficheiro *.csv e conseguisse traduzir para um compatível com o Orange (*.tab).

O algoritmo desenvolvido inicialmente vai buscar o ficheiro real com base no nome do mesmo, e carrega os conteúdos do mesmo. Após isso é feita a normalização do ficheiro de forma a não ter a informação de troca de linha, pois não é necessária. Concluído isso é feita formatação da informação, que recebe como base os separadores do ficheiro, neste caso vírgulas, o ficheiro em si. Após isso o algoritmo apresenta 4 passos:

- Normalizar nomes das colunas – Aqui simplesmente trocam-se os separadores antigos por Tabs.

- Criação de informação de domínio – Através dos índices, são passados quais as colunas com domínios contínuos e discretos.

- Criação da linha de indicação de classe – Através de um índice único, é indicado qual das colunas é a classe.

- Normalização dos tuplos – Simplesmente com base no separador antigo, é feita a troca para tabs de forma a ser compatível com o formato *.tab.

Por fim, são escritos os conteúdos deste trabalho para um novo ficheiro com o formato .tab, e o utilizador a partir daí pode utilizar o mesmo no Orange3.

Exercício 3

Nesse exercício, foi reaproveitada a implementação do modelo 1R do projeto anterior, que foi feito exatamente para poder ser genérico para qualquer dataset.

O formato apresentado já tinha o formato pedido, pelo que foi escolhido não mexer no código do mesmo de forma a evitar criar uma segunda versão.

Exercício 4

Este exercício foi feito exclusivamente em Orange3 com a utilização de operadores visuais. Foram utilizados todos os operadores sugeridos no enunciado do trabalho.

Resultados

Projeto A

Para a avaliação de resultados foram escolhidas as métricas apresentadas na introdução, por parecerem as mais apropriadas, que são a Accuracy, o F1, a Precision e o Recall. Por fim são apresentadas as matrizes de confusão de todos os dados de forma a ter informação mais detalhada.

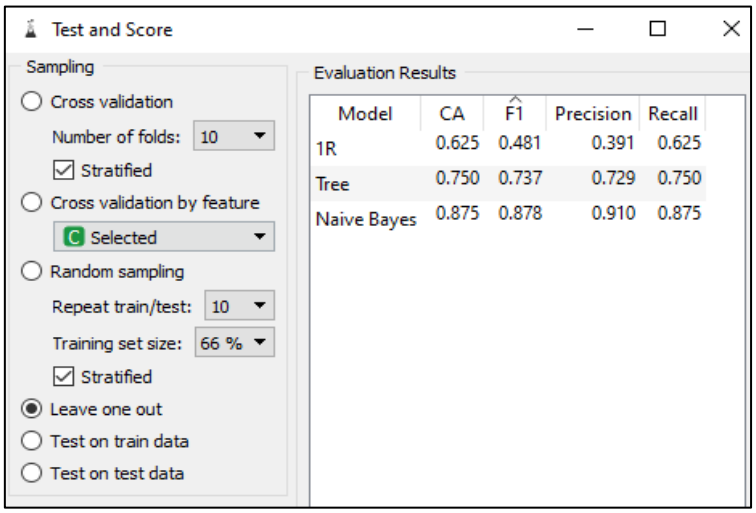


Figura 1 - Avaliação dos métodos de classificação

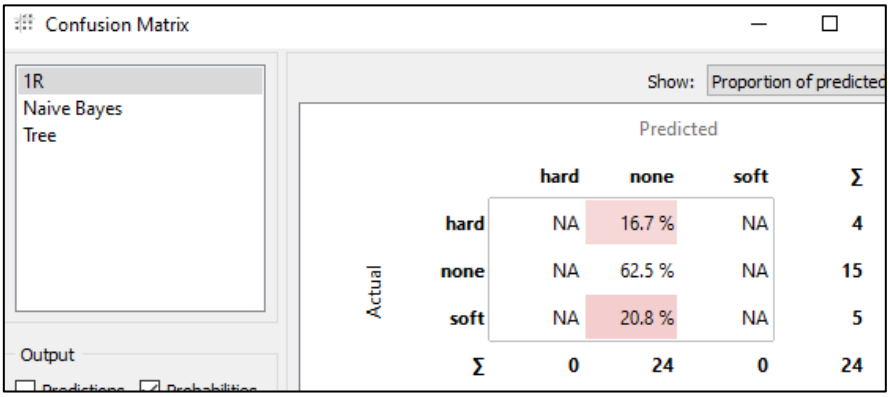


Figura 2 - Matriz de Confusão do método 1R

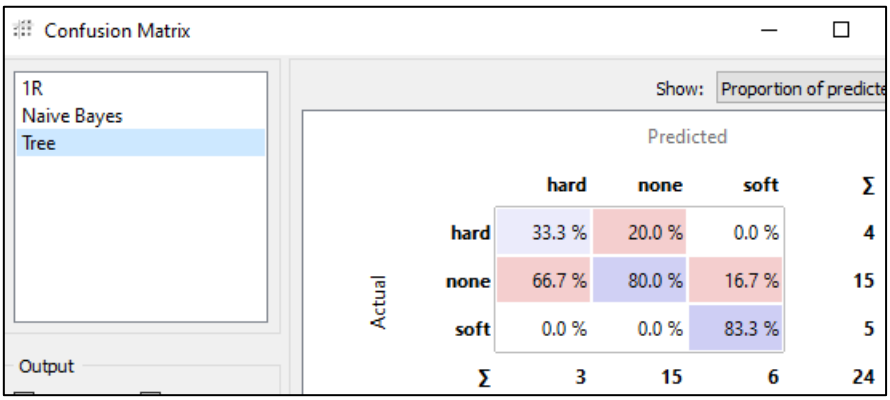


Figura 3 - Matriz de Confusão do método ID3-Árvore de Decisão

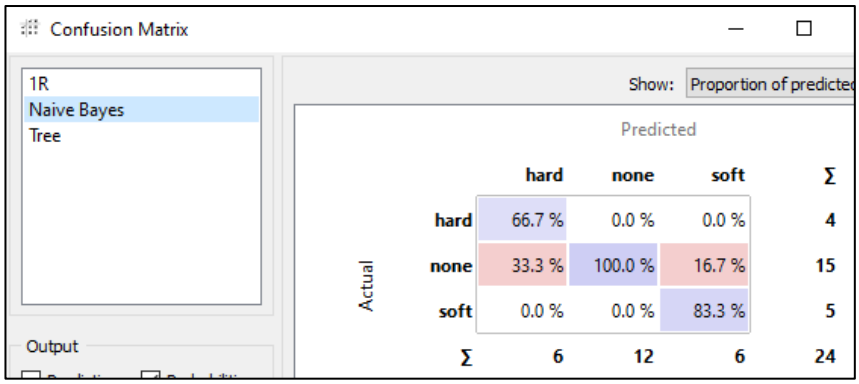


Figura 4 - Matriz de Confusão do método Naive Bayes

Projeto A1

```
ONE R
( attr, attr-value, class value ) : (error, total)
('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 5 - Resultados 1R

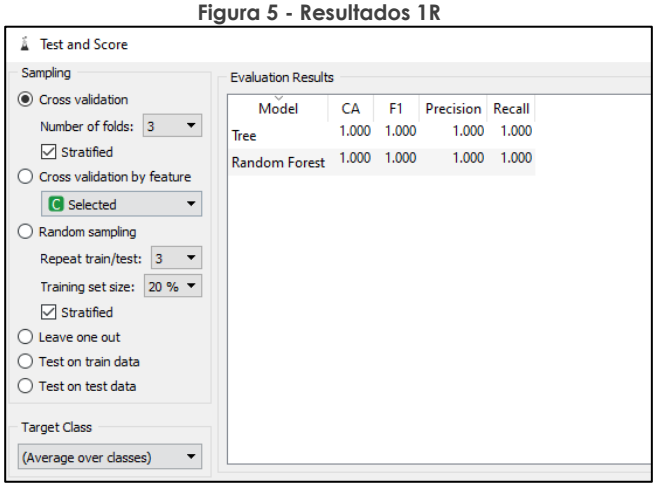


Figura 6 - Avaliação Cross Validation 3-fold stratified

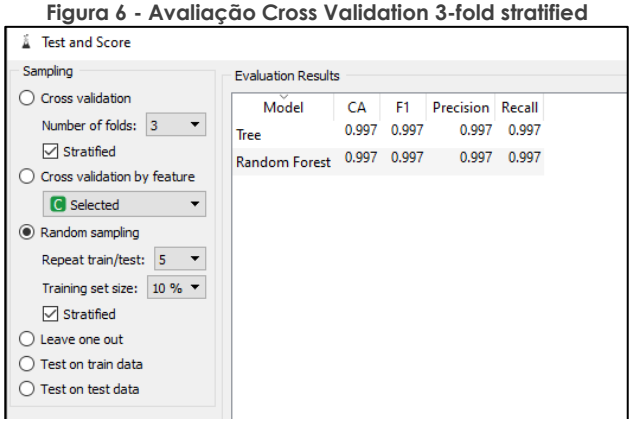


Figura 7 - Avaliação Random Sampling Estratificado com 5 testes, e 10% do dataset como dataset de treino

Análise de Resultados

Projeto A

Esta análise de resultados será feita com base nas métricas, e após isso nas matrizes de confusão.

De todos os métodos de classificação, o 1R apesar de ter tendência apresentar resultados bons tendo em conta a sua simplicidade, foi o pior. A accuracy e o recall do mesmo foram interessantes, no entanto todas as outras métricas foram bastante medíocres, que acaba por se provar ao apresentar apenas 62,5% de acerto total. Este método falha demasiado para ser sequer ponderado como solução a apresentar.

O método ID3, teve métricas bastante boas, e equilibradas, sendo que todas ficaram por volta dos 0.75, o que é interessante e prova a sua consistência como método de classificação. Dá a sensação de ser um método que classifica mais facilmente valores com mais instâncias, e apresenta mais dificuldade em valores com menos instâncias. Apesar de apresentar um bom desempenho não podemos recomendar este método também por não ser o melhor.

Por fim, o melhor método e desde já escolhido para apresentar à Medknow, é o Naive Bayes. Este método apresenta todas as métricas com valor superior a 0.88, sendo que o F1 é bastante alto, o que garante um bom recall e precision. Na figura 4 podemos ver que a maior falha é de 33%, pelo que ainda é aceitável, e, portanto, é o escolhido para apresentar à empresa.

Todo o trabalho realizado para este relatório encontra-se junto no ficheiro all-methods-evaluation.ows.

Projeto A1

Este projeto não apresentou tantos resultados como o anterior, mesmo devido a ser de uma dimensão menor que o outro. Por essa razão vai ser focada cada imagem deste projeto para analizar.

Na figura 5 são apresentados os resultados da aplicação do Modelo 1R nesse dataset. Neste caso consegue-se reparar que o 1R apresenta um óptimo desempenho, sendo que apenas apresenta erros numa das instâncias, e mesmo aí são extremamente poucos, aproximadamente 3% das entradas. Tendo em conta a simplicidade do modelo, isso é um resultado bastante bom. Escolhendo esta regra como solução seria interessante, devido à sua simplicidade.

As figuras 6 e 7 são duas situações de avaliação dos modelos Tree e Random Forest. Foram feitas duas avaliações pois, achava-se que um deles poderia ser uma indicação de overfitting. Foi escolhido o método de cross-validation estratificado 3-fold por não ser tão pesado computacionalmente

como o leave-one-out e ainda assim apresentar bons resultados. Foi escolhido o 3-fold, de forma a evitar o overfitting, mas ainda assim os resultados foram extremamente bons. Os resultados apresentados por este modelo foram tão bons quanto poderiam ser, pelo que este seria o modelo a escolher. Aqui a escolha da árvore ou da floresta acaba por ser uma escolha pessoal, no entanto o grupo pensa que a árvore será uma melhor escolha devido ao menor carácter estocástico do método.

Por fim foi feita a avaliação com Random Sampling estratificado, de forma a utilizar um método que fosse extremamente estocástico, de forma a evitar overfitting. Por essas razões também foi escolhido um tamanho de dataset de treino extremamente pequeno, mas ainda assim os resultados foram muito bons, pelo que com este teste o grupo chegou à conclusão que a razão dos resultados se deverá prender com o dataset em si. Em relação a este método, teve um bom comportamento, no entanto os dados para uma avaliação noutra situação não são ideais.

Por fim, pensamos que nestas avaliações não houve overfitting, o que se passou foi que o dataset era extremamente detalhado e preciso. Quiçá também os cogumelos tenham características que se correlacionam quase em 100% a sua comestibilidade. Para além disso grande parte dos atributos tinham bastantes valores, e a classe apenas tinha dois, assim sendo é possível fazer uma óptima junção entre atributos e a classe que acaba por ser expressa.

Conclusão

Com este trabalho foi possível compreender todo o trabalho desde ter os dados, até os conseguir inserir nalguma plataforma que consiga extrair conhecimento dos mesmos.

Também foi aprendido como implementar e utilizar métodos de classificação de dados, bem como ler e avaliar os mesmos. Tendo em conta que esta parte do trabalho foi maioritariamente implementa com base no uso de Orange3, também foram ganhas competências nesse software. Deu para compreender a utilidade e importância de utilizar Orange3, e de algumas das funcionalidades que têm, nomeadamente as relacionadas com classificação de datasets. Compreendemos que é uma ferramenta bastante poderosa e que consegue poupar bastante tempo em partes técnicas, que pode ser utilizado para realmente analisar os dados.

A partir dos deste trabalho é possível ganhar e investigar dados de outro campo, pois foi apresentado o método de trabalho para tal. Isto apenas aplicado à classificação dos dados.