Team Nr. 57

Data Science Project

Nr 81478 Name Joana Godinho
Nr 81811 Name João Crespo
Nr 81888 Name Pedro Caldeira

MEIC-A SAD 2017

INDEX

1 PRE PROCESSING

1.1 Crabs

Em primeiro lugar retirou-se o índice (*index),* dado que este é um falso preditor, ou seja, é um atributo gerado com base na classe (*species).* Para a execução do algoritmo *KNN* foi necessário alterar o atributo *sex* de categórico para numérico. Testou-se ainda **normalização** e ***PCA*** para este algoritmo de modo a minimizar a influência de atributos irrelevantes para a variância da data, sobre a distância entre "vizinhos". No caso do *Naïve Bayes*, pareceu especialmente relevante o uso de ***PCA*** e ***Feature Selection*** para contornar uma provável dependência dos atributos deste *dataset*, já que o *Naïve Bayes* apresenta uma melhor performance em *datasets* cujos atributos são independentes. Para a aplicação das *decision trees,* ao contrário dos anteriores, **não foi necessário fazer pré-processamento** (nomeadamente, *PCA* e *Feature Selection*) pois o próprio algoritmo dá mais importância a atributos que melhor servem para particionar o *dataset*. No entanto, tentou-se discretizar os dados por intervalos de forma a ver se se obtinha uma divisão com melhores resultados que a encontrada pelo algoritmo.

Ao fazer uma primeira análise do *dataset*, verificou-se uma relação entre a distribuição das espécies e o rácio entre os atributos ***FL*** e ***CW***, claramente visível no gráfico que relaciona estas duas variáveis. Assim, começou-se por gerar a equação da recta que melhor relacionava as duas variáveis - FL=0.4291CW - 0.192. De seguida, criou-se uma medida derivada dos dois atributos que toma o valor 1 se o valor de FL da instância estivesse abaixo da recta e 0 se estivesse acima. Esta transformação foi utilizada de forma a melhorar os resultados no *Naïve Bayes*.

1.2 No-shows

À semelhança do primeiro projecto, para os atributos *ScheduledDay* e *AppointmentDay,* separaram-se as horas das datas. Consequentemente, os atributos *ScheduledDay* e *AppointmentDay* passaram a ter apenas informação referente ao dia, informação esta que foi traduzida em valores de 1 a 366, consoante o dia do ano. **Removeram-se os atributos *AppointmentID* e *PatientID*** por apresentarem sempre ou quase sempre, respectivamente, valores diferentes para cada entrada. No caso do *KNN*, também o atributo ***Neighbourhood*** **foi removido** pelo facto de não ser quantificável devido ao fraco conhecimento do domínio. Assim como no *dataset* do crabs, aplicou-se **normalização** e ***PCA*** para o *KNN* e para o *Naïve Bayes*, pelas mesmas razões apresentadas acima. Ademais, para as *Decision Trees*, discretizou-se os intervalos de idades em 5 categorias (Bebé, Criança, Adolescente, Jovem Adulto, Adulto e Idoso) e as datas foram transformadas de dias do ano para semanas do ano.

Experimentou-se ainda a realização de ***feature selection***, em que se removeram de forma aleatória grupos de atributos.

Devido ao facto de o *noshows* ser um *dataset* desequilibrado, adoptaram-se **técnicas de balanceamento** de forma a equilibrar as classes no conjunto de treino. Deste modo, evita-se que o classificador se torne tendencioso para classificar as instâncias como sendo da classe maioritária, dado que este pretende minimizar a taxa de erro.

2 EXPLORATION

2.1 Crabs

2.1.1 Methods and Parametrization

Como estratégia de treino utilizou-se **Cross-Validation** com 10 folds, dado que é a mais aconselhável para *datasets* como o *crabs,* cujo número de instâncias está na ordem das centenas.

Naïve Bayes: Após a aplicação de *PCA* ao *dataset*, testou-se usar os dois primeiros *principal components*, visto cobrirem 98% da variância nos dados. No entanto, dado que no segundo componente, o sexo tinha muita relevância (quádruplo dos restantes atributos), testou-se também com 3 *principal components.*

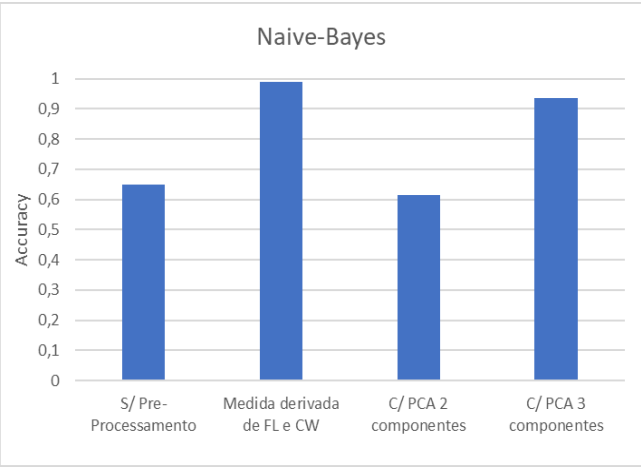
KNN: Testou-se o algoritmo com diferentes valores de *k* (1,3,5,7,9,11,13,15,17), todos ímpares de modo a evitar empates na classificação das instâncias. As distâncias entre vizinhos foram calculadas com a distância euclidiana. Tal como no *Naïve Bayes*, testou-se o uso de 2 e 3 *principal components.*

Decision Trees: Testaram-se os algoritmos ***C4.5*** e ***CART*** (*rpart*). No último variou-se o número de instâncias necessárias para particionar cada nó (*minsplit*) entre 2 e 50. Para as discretizações referidas no pré-processamento, variou-se o número de intervalos (*bins* = 3,5,7,11).

Para comparar as diferentes técnicas usou-se a medida de classificação ***accuracy***. Não se usou as medidas ***Sensibility*** e ***Specificity***, dado que o *crabs* é um *dataset* balanceado em que a preocupação é classificar as duas classes (*sp* = "B" ou "O") ao contrário do que acontece no caso do *noshows* em que a preocupação é classificar bem os casos positivos (*No-show* = "Yes").

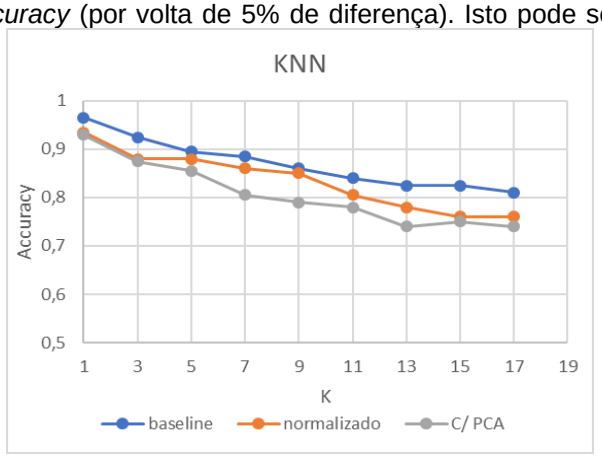
2.1.2 Results

Naïve Bayes: Relativamente ao uso do *PCA*, ao considerar apenas 2 componentes, os resultados não melhoraram, porém, ao considerar 3 componentes verifica-se uma **melhoria significativa** (por volta dos **93%** de accuracy). Esta subida pode ser justificada pelo facto de a segunda componente ser fortemente influenciada pelo atributo *sex*, que poderá estar correlacionado com os restantes atributos. Ao considerar a terceira componente, cuja influência de cada atributo é relativamente semelhante, o efeito da segunda componente nos resultados é atenuado. A medida derivada foi a que obteve **melhores resultados** com **99%** de *accuracy*. É um resultado expectável tendo em conta a análise dos gráficos gerados pelo *Weka*.



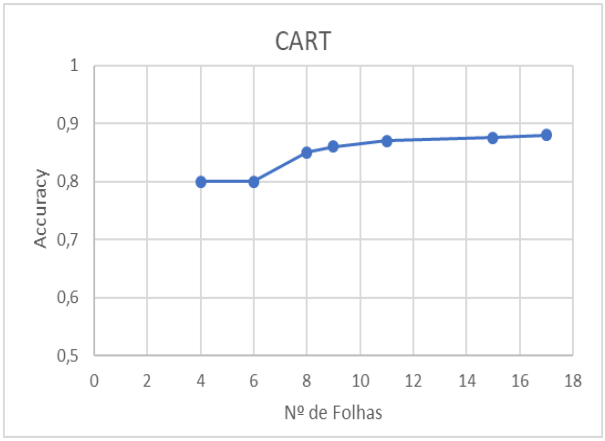
Apesar de não ter sido testada, a medida derivada teria muito provavelmente a mesma qualidade de resultados com os outros algoritmos (KNN e CART)

KNN: Pelo gráfico verifica-se que os resultados sem pré-processamento e com normalização foram muito semelhantes em termos de *accuracy* (por volta de 5% de diferença). Isto pode ser justificado pelo facto de **todos os atributos terem intervalos de valores semelhantes**. Ao se utilizar o *PCA*, os resultados pioraram ligeiramente, possivelmente pelo facto de as componentes não descreverem bem as espécies.



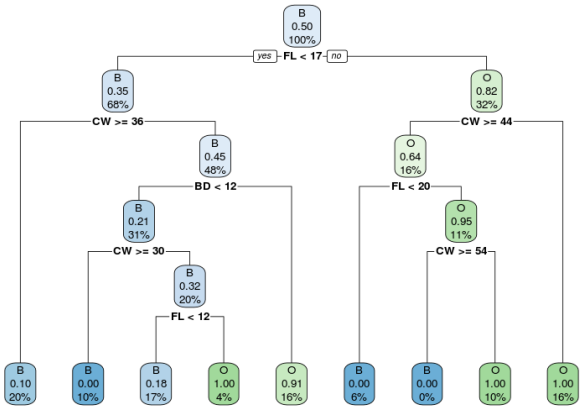
Pelo gráfico é possível ainda concluir que o número de vizinhos (*k*) igual a **1** conduz a melhores resultados. Tal pode dever-se ao facto das **características físicas**, entre instâncias de classes diferentes **não variarem substancialmente**. Logo, ao classificar uma dada instância, quando se aumentar o valor de *k*, a probabilidade de caranguejos de classe oposta serem considerados vizinhos aumenta.

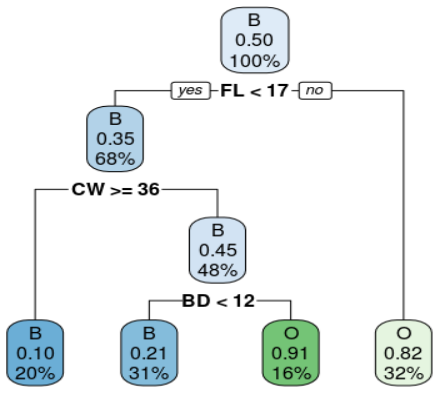
Decision Trees: As discretizações em intervalos por *bins* não geraram modelos melhores que os gerados pelos algoritmos *C4.5* e *CART* com os dados originais, o que era expectável, visto que ambos os algoritmos procuram encontrar o valor do atributo que melhor divide as instâncias pelas classe (caso seja um atributo numérico).



Pelo gráfico, é possível constatar que a ***accuracy* aumeta** com o número de folhas na árvore, já que um elevado número de folhas implica um aumento no número de testes aos atributos (nós). Esta correlação era expectável uma vez que a diminuição de nós pode levar à perda de informação. Isto implica que num *dataset* como o *crabs*, não tendo em conta relações entre atributos (nomeadamente *FL* e *CW*), todos estes influenciam positivamente a classificação. Assim, a melhor árvore é a **não *pruned***.

Ao analisar as diferentes árvores geradas é possível verificar que os atributos primeiramente selecionados para teste são o *FL* e o *CW*. Tal significa que estes dois atributos conduzem a um **maior ganho de informação** durante a classificação, algo expectável pelos resultados da medida derivada do *Naïve Bayes*.





2.2 No-shows

2.2.1 Methods and Parametrization

Como estratégia de treino utilizou-se a divisão em ***sets* de treino e teste**, visto ser o método mais adequado para um *dataset* tão grande como o *noshows* (milhares de instâncias).

Para balancear o conjunto de treino, de modo a que este tivesse o mesmo número de instâncias classificadas como "*No*" e como "*Yes*", aplicou-se ***undersampling*** aleatório da classe maioritária (No-show = "No") e ***SMOTE***. Experimentou-se ainda balancear o conjunto de treino de forma a ter 70% de instâncias "*No*" e 30% de instâncias "*Yes*", também através de *undersampling*, de forma a obter melhores resultados na medida de *accuracy*.

Para avaliar os diferentes classificadores usaram-se as medidas ***accuracy* e *sensibility*,** (rácio de *No-shows = "Yes"* que o classificador acerta) dado que a dificuldade neste *dataset* é classificar correctamente a classe minoritária, isto é, os casos positivos (*No-show* = "*Yes*"). Torna-se relevante classificar corretamente estes casos tendo em conta o domínio do problema, já que ao prever que um paciente não vai aparecer na consulta, permite poupar recursos humanos nos hospitais.

Para a *feature selection* aleatória, retiraram-se iterativamente grupos de **uma ou duas *features*** do *dataset* antes de realizar o treino.

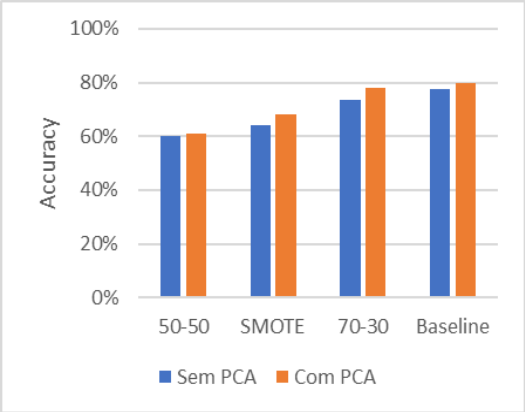
Naïve Bayes: Após a aplicação de *PCA* ao *dataset*, testou-se usar os **nove primeiros *principal components***, visto cobrirem **92%** da variância nos dados.

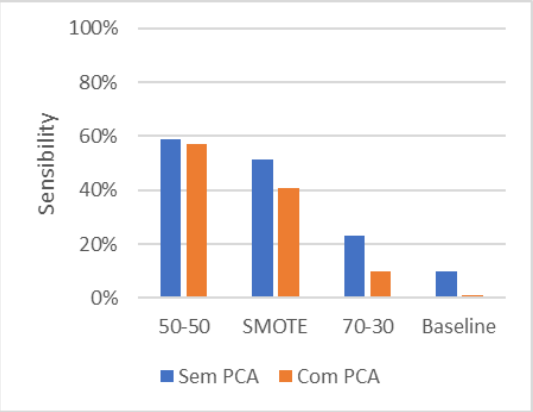
KNN: Testou-se o algoritmo com diferentes valores de *k* (1, 3, 5, 7, 9, 11, 13, 15, 17, 21, 27, 33, 41, 55, 67 e 101). Tal como no *Naïve Bayes*, testou-se o uso de 9 *principal components.*

Decision Trees: Testou-se o algoritmo ***CART***, com e sem a aplicação de ***pre-prunning***, onde se variou o número de instâncias necessárias para particionar cada nó (*minsplit*) entre 500 e 20000, o que refletiu uma variação do número de folhas entre 2 e 1500. Esta abordagem foi aplicada tanto para as discretizações referidas no pré-processamento, como para o *baseline* (*dataset* sem balanceamento).

2.2.2 Results

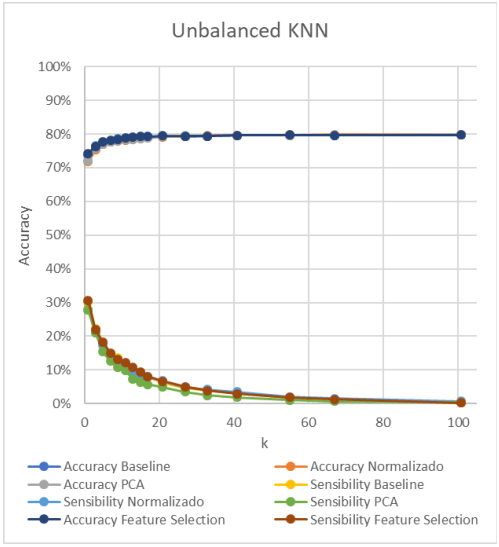
Com base nos gráficos, é possível concluir que o uso do *PCA* **não melhora as medidas**, pelo contrário piora, algo que não era expectável. Adicionalmente, verifica-se que para balanceamentos em que a ***accuracy* é alta, a *sensibility* é baixa**, logo, apesar de com esses métodos se classificarem corretamente mais instâncias, o número de instâncias *No-show = "Yes"* classificadas acertadamente **diminui substancialmente**.





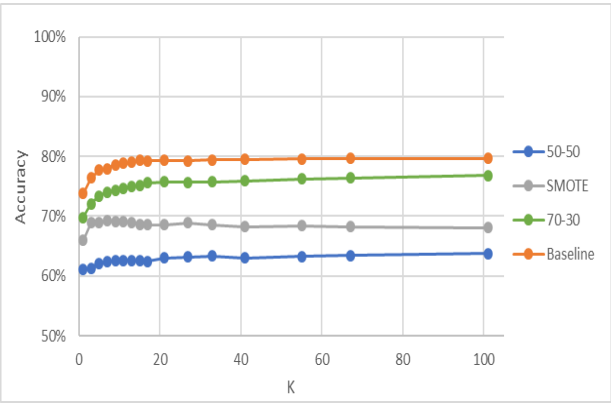
A elevada *accuracy* observada para o *baseline* e para o balanceamento *70-30*, é provocada pelo comportamento tendencioso dos classificadores para com a classe maioritária (*No-show = "No"*), explicado no pré-processamento.

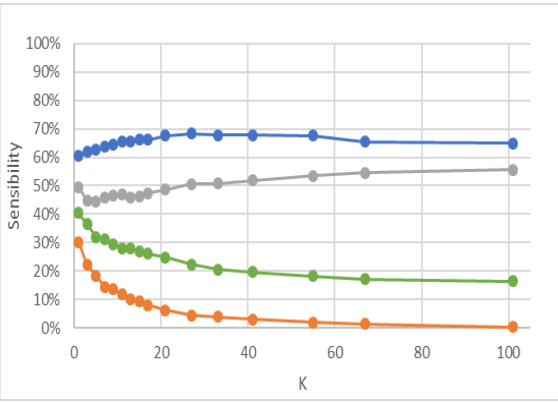
KNN: Pela análise do gráfico podemos verificar que todas as técnicas de pré-processamento (Normalização, *PCA e Feature Selection*) tiveram uma **performance equivalente** tanto em termos de *Accuracy* como de *Sensibility*. Ainda assim, testou-se o algoritmo KNN com estas técnicas, para os diferentes balanceamentos, sem grande diferença.



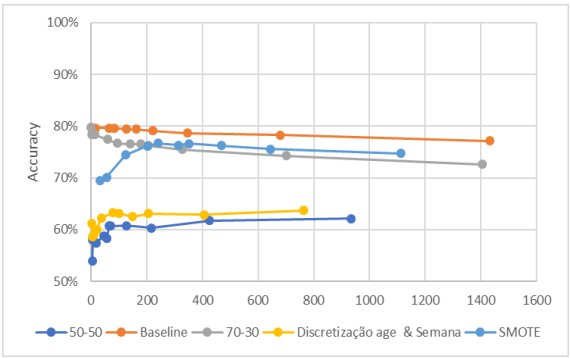
Por outro lado, o que realmente influenciou os resultados foi o **balanceamento do *dataset***. Tal como é possível observar pelos gráficos abaixo, os dois balanceamentos mais próximos da realidade, *baseline* (proporção de *Yes* e *No* igual à do *dataset*) e *70-30*, foram os que obtiveram melhores resultados em termos de *accuracy*. No entanto, pela mesma razão que no caso do *Naïve Bayes*, a *sensibility* é **demasiado baixa**.

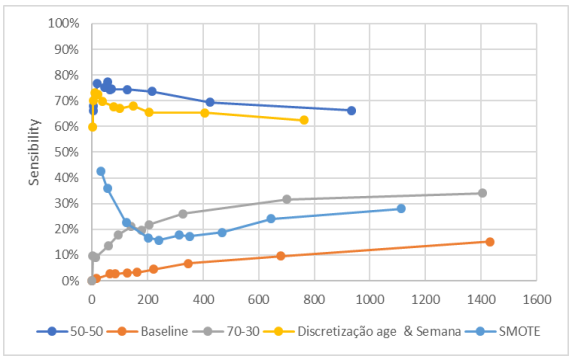
Assim, os **melhores resultados encontrados** foram aqueles em que o *dataset* foi equilibrado com ***undersampling***, de forma a ter o mesmo número de instâncias "*Show"* e *"No-Show"* (50-50), e aqueles em que se usou ***SMOTE***, pois apresentam as **melhores relações *accuracy-sensibility***. Segundo a tendência visível nos gráficos, para valores de *K* ainda mais elevados, o algoritmo *SMOTE* encontraria possivelmente uma melhor relação entre *Accuracy* e *Sensibility*, no entanto, tal não foi possível verificar devido a limitações de *hardware*.





Decision Trees: Tal como nos métodos anteriores, o *baseline* e o balanceamento *70-30* originaram valores de *accuracy* superiores, mas uma *sensibility* pior, do que os balanceamentos com ***undersampling*** *50-50* e com *SMOTE*.



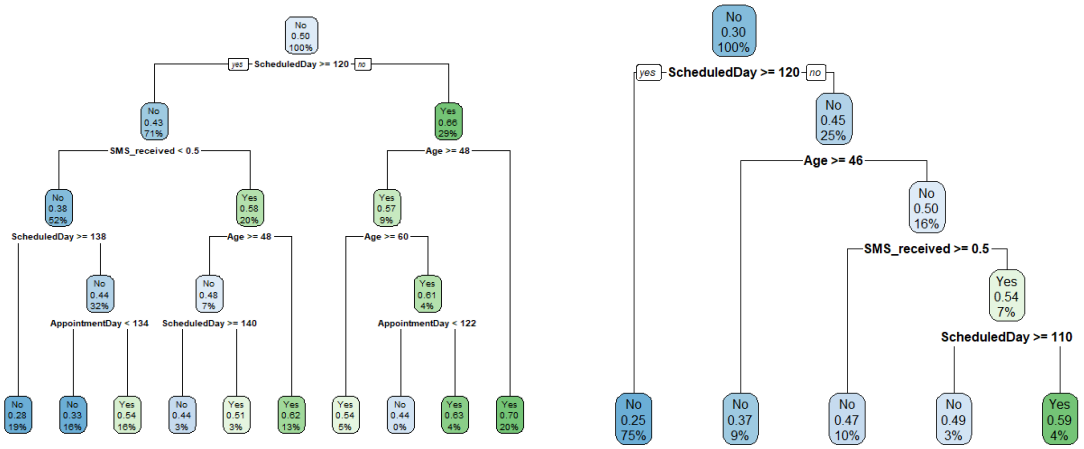


Pelos gráficos pode-se constatar, no caso dos balanceamentos *50-50* e *SMOTE* que, à medida que se aumenta o número de folhas, a ***accuracy* aumenta e a *sensibility* diminui**. O mesmo acontece para o balanceamento *50-50* com discretização sobre a idade e as datas. Em qualquer uma das três alternativas, verificou-se uma maior variação das métricas entre 2 a 100 folhas (aproximadamente), dado que ao variar a quantidade destas dentro desse intervalo, dá-se um maior aumento do número de testes aos atributos, o que implica mais informação para a classificação.

No caso do *baseline* e do balanceamento por *undersampling 70-30*, a *accuracy* diminui e a sensibility aumenta, com o número de folhas. Ou seja, com o aumento do número de folhas na árvore, o classificador **tende a catalogar as instâncias negativas** (*No-shows = "No"*), **progressivamente pior, mas as positivas** (*No-shows = "Yes"*), **melhor**. Tal como acontece nos dois algoritmos anteriores, estes apresentam um **enviesamento dos dados**, especialmente visível quando há poucas folhas. No entanto, a variação da *sensibility* é bastante superior à da *accuracy*, logo ao se aumentar significativamente o número de folhas, o classificador pode tornar-se melhor.

Em todas as alternativas, a certa altura, as **métricas tendem a convergir**. Isto significa que, a partir de um dado momento, o aumento do número de testes leva a computação desnecessária pois o **ganho de informação não aumenta** e, portanto, não se dá um melhoramento da classificação.

Por fim, ao analisar algumas árvores geradas por diferentes balanceamentos e discretizações, verificou-se que alguns dos primeiros atributos testados eram quase sempre os mesmos, caso do *ScheduledDay* e *Age*. Isto significa que **esses atributos acarretam mais informação** relevante para verificar se um paciente vai à consulta ou não, ao contrário das doenças



3 CRITICAL ANALYSIS

No caso do *dataset* do *crabs*, o *KNN* apresentou os melhores resultados sem ser necessário um pré-processamento complexo. Já ao aplicar o *Naïve bayes*, a utilização de uma medida derivada revelou ser importante, sendo que melhorou substancialmente a classificação, acabando por se obter um classificador quase perfeito. Contudo, se a relação entre *FL* e *CW*, presente no *dataset*, utilizada para gerar o novo atributo, não se verificar na realidade, isto é, não apresentar diferenças tão aparentes entre instâncias de classes diferentes, o modelo encontra-se em *overfitting* e pode mostrar resultados bastante piores que as restantes alternativas.

Aquando a realização de *association rules* no primeiro projecto, observaram-se dependências consideráveis entre as características físicas dos caranguejos (por exemplo: FL-> CW e CL->BD). Assim, os piores resultados no *Naïve Bayes*, não considerando a medida derivada, podem ser justificados por estas dependências.

No caso do *noshows*, a utilização de *feature selection* no *KNN* não apresentou resultados **promissores**. A **falta de conhecimento de domínio e poder computacional** não permitiu a tentativa de todas as *selections* possíveis. Consequentemente, existe a possibilidade do método se comportar melhor com *selections* que não foram passíveis de teste.

Ainda assim, através de algumas conclusões da análise de domínio era **esperado** que o atributo ***SMS_received*** tivesse uma **importância considerável** na classificação, já que um paciente que recebe uma mensagem a relembrar a consulta devia ser mais susceptível a aparecer na consulta do que um que não recebe. Previa-se, portanto, que ao retirar este atributo (*feature selection*) os resultados tendessem a piorar, mas tal não se verificou. Esperava-se também que o **ganho de informação** deste atributo fosse maior em relação aos outros. Porém, quando foram geradas as árvores de decisão, este não era escolhido como **raiz** da árvore, corroborando, assim, os resultados obtidos no KNN com *feature selection*. Assim, com base neste resultado, o hospital teria que repensar as medidas adoptadas para relembrar os seus pacientes.

Ademais, dada a **relevância do *ScheduledDay***, poderia ter sido interessante analisar o impacto da **diferença entre o *ScheduledDay* e o *AppointmentDay***, dado que, se o tempo de espera entre a marcação da consulta e a consulta em si for muito elevado, um paciente pode se esquecer da consulta, levando a que depois não apareça. Para além disso, tendo em conta os resultados das árvores de decisão, poderia ter sido curioso analisar o KNN com uma ***feature selection*** em que apenas se utilizariam o ***ScheduledDay* e a *Age***.

Apesar de se terem utilizado diversos métodos de **balanceamento**, estes não obtiveram grande sucesso com o *dataset no-shows*. Isto pode dever-se ao facto de os atributos **não refletirem** a ida ou não dos pacientes às consultas, ou seja, possivelmente deviam ser **necessários mais ou outros atributos** para a predição ser mais correta. No caso das árvores de decisão, tal pode ser verdade pois, ao se usarem mais atributos do que os presentes no *dataset*, o tamanho da árvore aumenta, o que pode levar a um **melhor classificador**, como explicado anteriormente.

Inversamente, pode-se concluir que os atributos do *crabs* são **suficientemente descritivos** sobre a espécie de cada caranguejo, não sendo necessário recorrer a mais informação.

Comparando os resultados, é possível verificar que aplicando diferentes classificadores, a **performance** destes **não diverge** significativamente. Contudo, o método do *Naïve bayes* foi o algoritmo com piores resultados para ambos os *datasets*, muito provavelmente devido à sua **sensibilidade** para com a **dependência das variáveis** e o balanceamento do *dataset*. Tanto o KNN como as árvores de decisão demostraram-se ser mais robustos quando confrontados com um *dataset* desequilibrado, tendo obtido resultados melhores.

4 CONCLUSIONS

Pode concluir-se que os algoritmos de aprendizagem conseguem fazer previsões bastante **boas** para **datasets equilibrados**. Com efeito, o balanceamento do dataset original é imperativo para o treino dos classificadores. No caso do *noshows,* o desequilíbrio dos dados **não é** suficientemente **compensado** por nenhum método de balanceamento.
Conclui-se ainda que, apesar dos vários métodos de classificação utilizarem abordagens e teorias subjacentes diferentes, todas estas acabam por convergir em métricas iguais, pelo que, dentro da pequena variação que existe nos resultados, os métodos categorizam mais ou menos da mesma forma. Conclui-se também que o que mais faz variar os resultados é o pré-processamento e tratamento que é aplicado aos dados, sendo que para diferentes tratamentos os resultados variam substancialmente.