Data Science Project 2018

TEAM #06

Projecto de Ciência dos Dados

Bernardo Furet, ist180844
Ana Costa, ist424729
Luis Antunes, ist424802

Contents

1. INTRODUCTION

No projecto foi pedido para aplicar os nossos conhecimentos acerca de técnicas de ciência dos dados na exploração de dados e obtenção de informação em dois problemas distintos.

No primeiro problema são fornecidos dados acerca de sensores relacionados com o Sistema de Pressão a ar (Air Pressure system, APS) dos camiões Scania com o intuito de perceber se este poderá estar avariado necessitando por isso de reparação para evitar maiores problemas posteriormente. Para os algoritmos de classificação, são nos dados também os custos de detetar falsamente uma avaria (10), e não detetar uma avaria no sistema apesar esta existir (500).

No segundo problema temos como tarefa avaliar a qualidade dos dados obtidos em exames de colposcopia, nos seus diferentes métodos (green, hinselmann, schiller), e como estes dados permitem a especialistas identificar problemas no colo do útero, nomeadamente de modo a detetar a presença de cancro.

2. NON-SUPERVISED MINING

2.1 PRE-PROCESSING

Normalizou-se as escalas dos atributos de modo a serem comparáveis para os algoritmos.

2.2 CLUSTERING

O k-means serviu como base para escolher o número de clusters (k-clusters) para os outros algoritmos de *clustering* que têm também este parâmetro. Escolheu-se o k-cluster tendo por base as métricas: silhouette, Rand Index, Completeness, Mutual Information, Homogeneity e V-measure.
Na análise do Colposcopy, os algoritmos testados foram: os clusters aglomerativos com as variantes ward, complete, average e single, e ainda os algoritmos spectral, affinity, DBSCAN, Birch, gaussian mixture, mean shift, mini batch k means.
Os gráficos 1, 2 e 3 representam a performance variando os vários parâmetros para o algoritmo K-Means, Affinity e DBSCAN, respectivamente, na modalidade green. Os gráficos para a modalidade hinselmann e Schiller eram bastante semelhantes.
Para escolher o k-means, a dúvida esteve entre escolher 2 ou 3. Na figura 1, pode-se verificar que o pico do silhouette apresenta-se para o k-cluster 3, mas há um bom trade-off entre as diferentes métricas no 2 e no 3. No entanto, treinando os algoritmos clustering que recebiam este parâmetro, não se obteve um tão bom resultado como com 2. Sendo assim, usou-se 2 k-clusters para todas as modalidades.
Usou-se os gráficos da figura 2 para escolher os parâmetros de *damping* do *Affinity Propagation* e o *epsilon* do *DBSCAN*. Pela análise dos gráficos e treinando os algoritmos concluiu-se que os melhores valores para o damping era 0.7 e o eps a 7.
Na análise do *APS Trucks* os únicos algoritmos testados foram o *Mini Batch K Means* e o *Gaussian Mixture*. Testou-se apenas estes algoritmos porque eram os que tinham menor complexidade. Os outros algoritmos devido ao tamanho do dataset ocupavam demasiada memória.

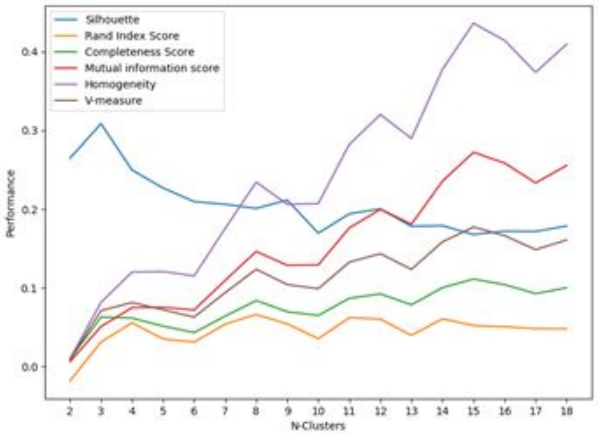


Figura 1 Performance variando o número de N-Cluster (k-clusters) para a modalidade green

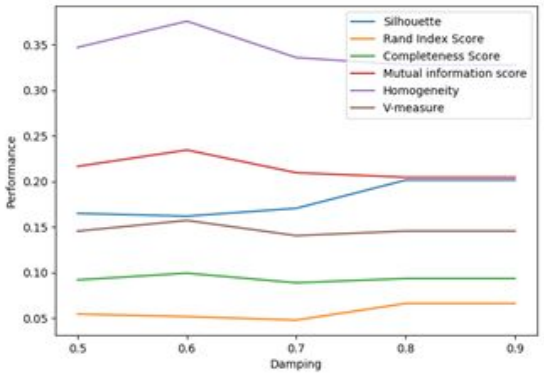


Figura 2 Performance variando o parâmetro Damping para a modalidade green

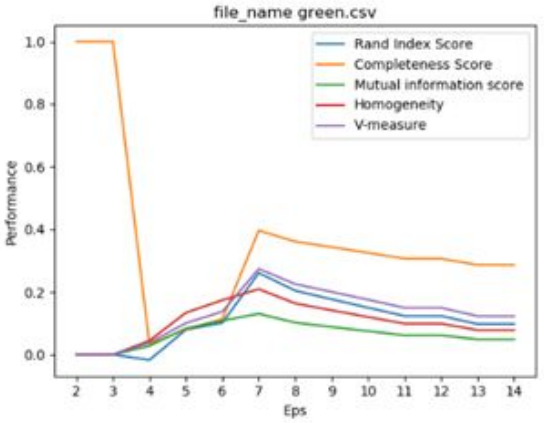


Figura 3 Performance variando o parâmetro Eps para a modalidade green

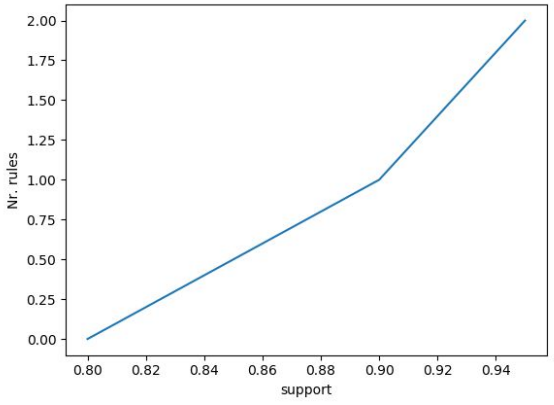
.

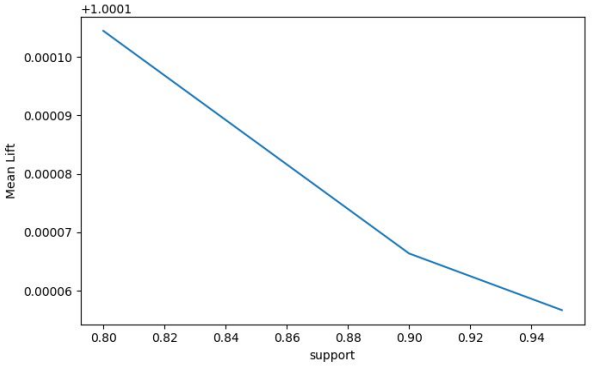
2.3 PATTERNS

Experimentou-se vários valores de suporte mínimo para o algoritmo *apriori*. No colposcopy escolheu-se o valor de suporte mínimo 0.95 porque era o valor máximo para o qual havia regras. O mesmo valor foi escolhido para a análise do APS Trucks. O tamanho máximo de itemsets gerados foi de 2. Para filtrar as regras usou-se o algoritmo de association rules, com um threshold de 0.99.

Para o dataset de colposcopy, as regras encontradas com maior lift têm como valor 1.04. Com menor lift 1.01. A confidence mantinha-se a 1 para todas as regras.

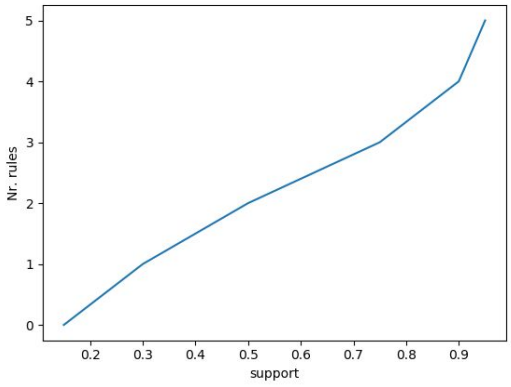
Para o datasets APS Trucks, o lift máximo é de 1.00952 e o confidence também é de 1. O lift minimo foi de 0.99. De confidence 0.990.

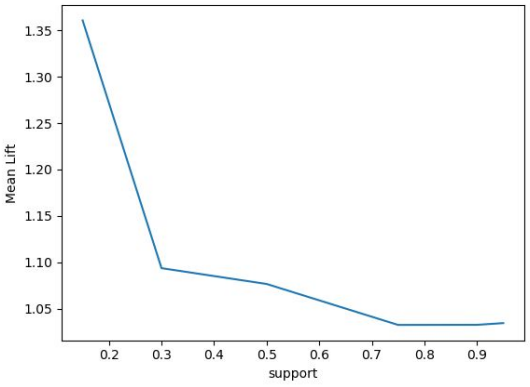




Foram encontrados muitos casos em que o antecedente e o consequente tinham uma relação bidireccional.

As seguintes imagens mostram que há um maior número de regras quanto maior for o Lift e o Lift médio vai diminuindo à medida que o suporte aumenta. Isto é o comportamento esperado[2].





2.4 RESULTS

Na *análise da Colposcopy,* nos algoritmos de clustering, aquele que teve melhor performance foi o dbscan, tendo melhores valores nos *scores* de *mutual information* (0.12), *homogeneity* (0.2) ,*V-measure* (0.25) e *adjusted index rand score*(0.25). Apesar de ser melhor entre todos os testados, são valores maus. Provou-se que com as features consideradas, os algoritmos de clustering não são bons.
Já o *Mean shift* teve melhor performance no *completeness score* (1).
Na análise do APS Trucks, o *Mini Batch K Means* possui melhores valores nas métricas testadas (adjusted rand index 0.47, completeness score 0.25, mutual information score 0.03, homogeneity 0.43, v-measure 0.31) do que o *Gaussian Mixture*.

No pattern mining, os lifts das regras estavam muito perto de 1, motivo pelo qual se considerou estas padrões inúteis, visto que os seus acontecimentos seriam independentes.

3. CLASSIFICATION

3.1 PRE-PROCESSING

O dataset APS Failure at Scania Trucks não estava balanceado. Usou-se a técnica SMOTE para balancear a data. Para este dataset a target variables é o atributo ‘class’.

Para o dataset Quality Assessment of Digital Colposcopies, as target variables são consensus, experts::0: experts::1 experts::2, experts::3, experts::4 e experts::5. Retirou-se do X do train, os experts de forma a não considerar a moda, já que esta já é implicitamente calculada na variável consensus. Dividiu-se o data em 80% train e 20% test. Mais uma vez, o dataset não estava balanceado. Usou-se a técnica SMOTE. É de salientar que na classificação com random forests não foi usado o balanceamento de dados, em vez disso recorremos ao parâmetro inerente do classificador de random forests do sklearn chamado class_weight com o valor "balanced" que atribui um peso inversamente proporcional à presença das classes nos dados de input.

Usando o Naive Bayes como base conclui-se que a melhor forma de preencher os NA em ambos os datasets seria preencher no train os valores com a média do atributo dentro da classe, e no teste com a média de todo o atributo (não considerando as classes).

Não se fez tratamento de outliers. Os outliers correspondiam à mudança da target variable. Ao retirar o outlier, ficava-se só com uma classe ou muito poucas, sendo crucial para a mudança de classe.

3.2 NAIVE BAYES

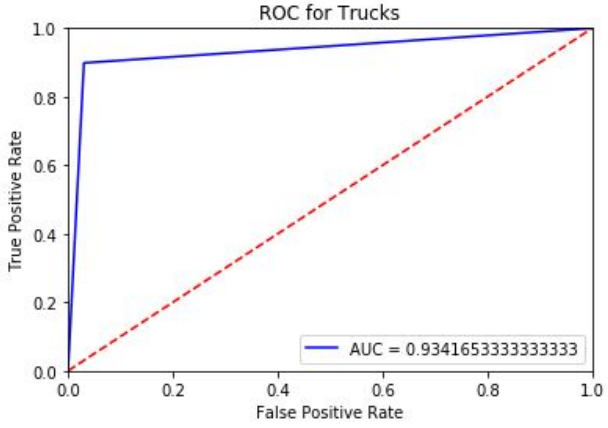
O Naive Bayes é usado como baseline para os outros classificadores. Assim, aplicámos algumas técnicas usando Naive Bayes, para ter uma estimativa/guia quanto aos outros classificadores.

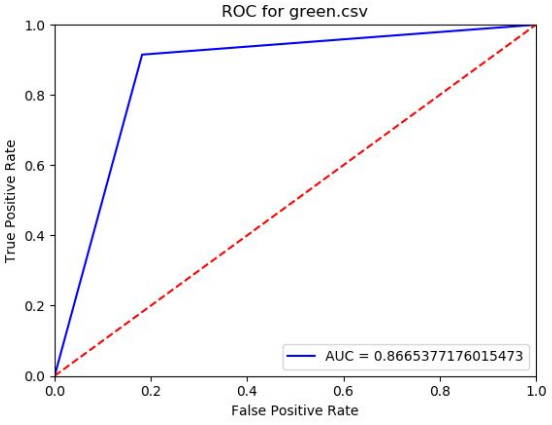
Ao aplicar o Naive Bayes, para cada um dos procedimentos de tratamento de valores desconhecidos, calculamos o *accuracy score* e a *confusion matrix* (extraindo a sensitivity e a specificity).

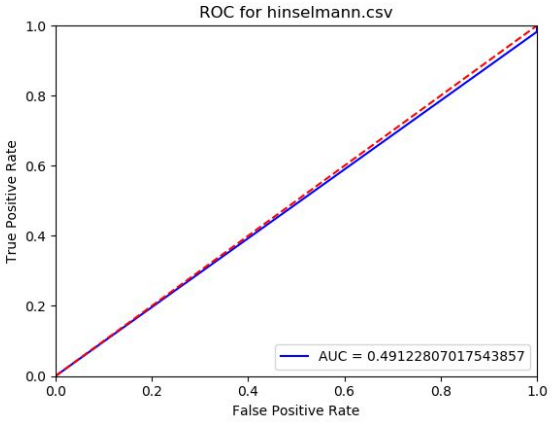
O *accuracy score* computado pelo método sklearn.metrics.accuracy_score é diferente da calculada através dos dados da confusion matrix. Isto é porque a primeira foca-se num único cutpoint.[1]

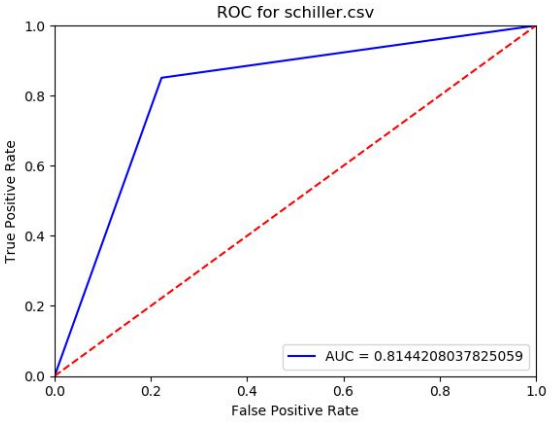
Para o caso dos APS, em relação às instâncias não numéricas, como as do atributo “class”, que era o atributo que queríamos estudar, convertemos os “pos” para 1 e os “neg” para 0. Após aplicar o procedimento do Naive Bayes, chegámos à conclusão que, como os valores calculados eram todos iguais, não havia valores desconhecidos para a classe “class”, pelo que basta analisar um dos casos.

Para o caso da Colposcopia, não existe nenhum valor n/a, portanto, para cada procedimento de tratamento dos n/a, obtêm-se os mesmo resultados. O hinselmann é o que tem pior resultados.



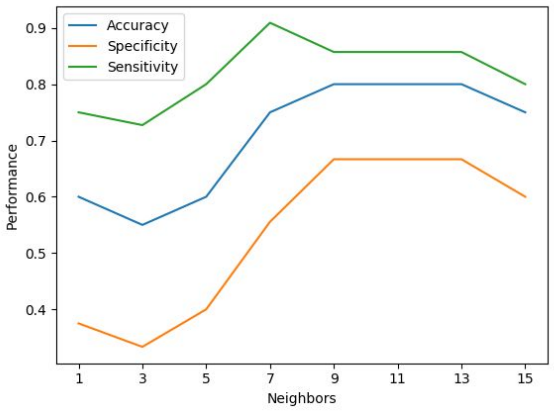






3.3 KNN

Na análise do *colposcopy*, na modalidade *Green* usou-se 9 neighbors, porque são os picos dos gráficos na figura abaixo. No gráfico ROC verifica-se que tem bons resultados, tendo um AUC de 90%.



Performance para a modalidade Green variando os neighbors

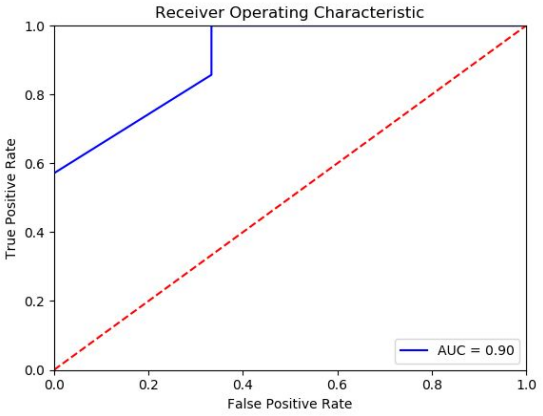
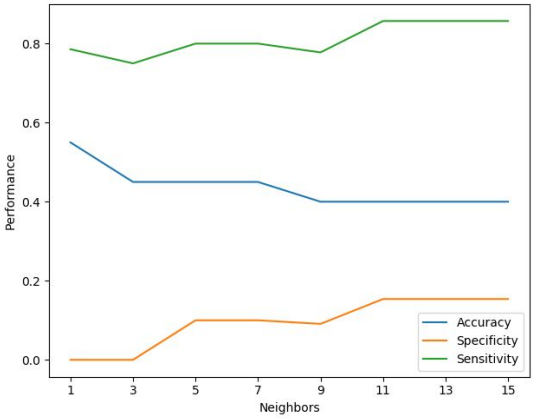


Gráfico ROC para a modalidade Green

Na modalidade hinselmann usou-se 9 neighbors, pois era o que representava um melhor trade-off entre as diferentes métricas usadas. Não possui um bom resultado pois tem um AUC de 0.40.



Performance para a modalidade *Hinselmann* variando os neighbors

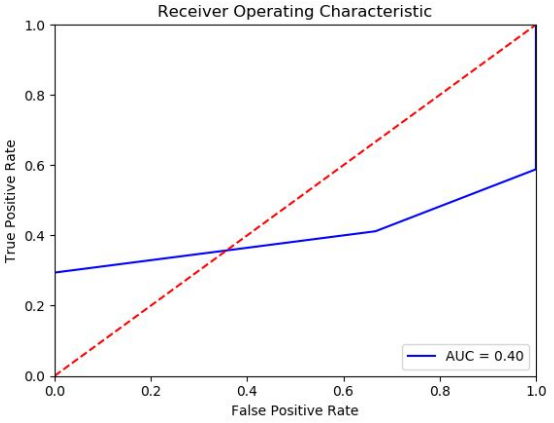
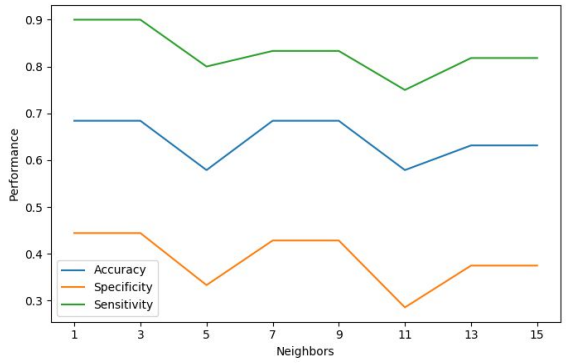
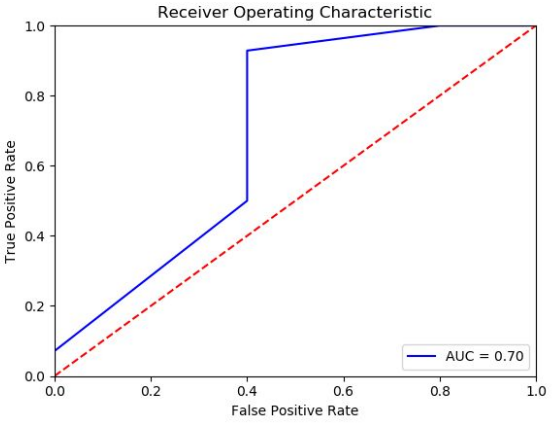


Gráfico ROC para a modalidade *Hinselmann*

Na modalidade schiller usou-se 7 neighbors, pois representava um pico entre as várias métricas. O resultado AUC foi bom, sendo de 0.70.

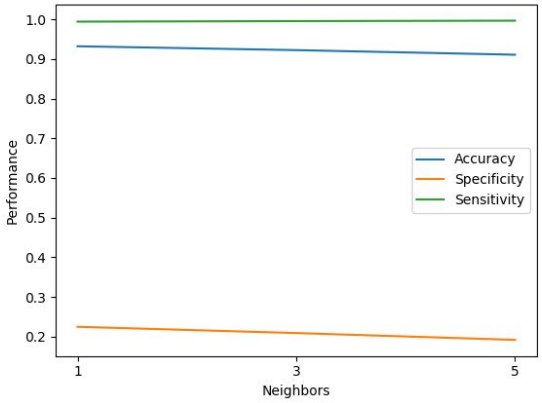


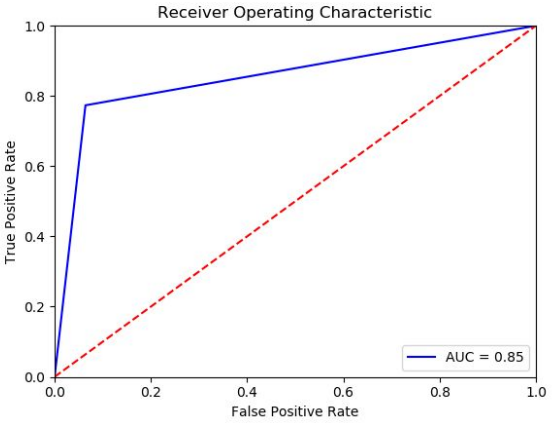


Performance para a modalidade *Schiller*

Gráfico ROC para a modalide *Schiller*

Na análise dos *APS Trucks*, o vizinho foi de 1. Aumentando a quantidade de vizinhos, a performance mantinha-se quase inalterada.. Tem uma bom resultado, tendo uma AUC de 0.85.





Performance no APS Trucks variando neighbors

Figura Gráfico ROC para o APS Trucks

3.4 DECISION TREES

O projecto foi desenvolvido em Python, com o auxílio da biblioteca scikit-learning, que não implementa explicitamente os algoritmos ID3, C4.5 e CART. Permite apenas distinguir o critério (criterion), que corresponde ao Information Gain (“Entropy”) ou a Impurity (“Gini”). Ambos foram aplicados.

Para o caso do APS, o tratamento dos valores desconhecidos não influencia o classificador. No entanto, há diferença entre os dois critérios, mas, como se pode ver abaixo, são mínimas. O classificador consegue prever os casos correctos, com regularidade.

Entropy:

Accuracy score: 0.989

Sensitivity: 0.993042701219123

Specificity: 0.7987987987987988

Gini:

Accuracy score: 0.9885625

Sensitivity: 0.9931025673776983

Specificity: 0.7807017543859649

Para o caso da Colposcopia, não há diferença entre os dois critérios e como os dados não têm valores desconhecidos, não há diferença entre os procedimentos que lidam com valores desconhecidos. Assim, basta analisar um critério para cada ficheiro de dados.

Como o objectivo é tentar minimizar o facto de uma pessoa ter cancro e não ser acusado esse facto ou uma pessoa não ter cancro e ser acusado esse facto, vamos querer maximizar os casos correctos. Ou seja, maximizar a sensitivity e a specificity. O conjunto de dados que apresenta maior sensitivity é o green e o que apresenta maior specificity é o hinselmann. No entanto, o hinselmann apresenta uma sensitivity muito baixa, tal como a menor accuracy dos três, pelo que deve ser descredibilizado (demasiados casos false negatives, que se traduzem em enviar pacientes para casa, como se estivessem saudáveis, não estando).

green.csv:

Accuracy score: 0.7101449275362319

Sensitivity: 0.5555555555555556

Specificity: 0.7647058823529411

hinselmann.csv:

Accuracy score: 0.6323529411764706

Sensitivity: 0.15

Specificity: 0.8333333333333334

schiller.csv:

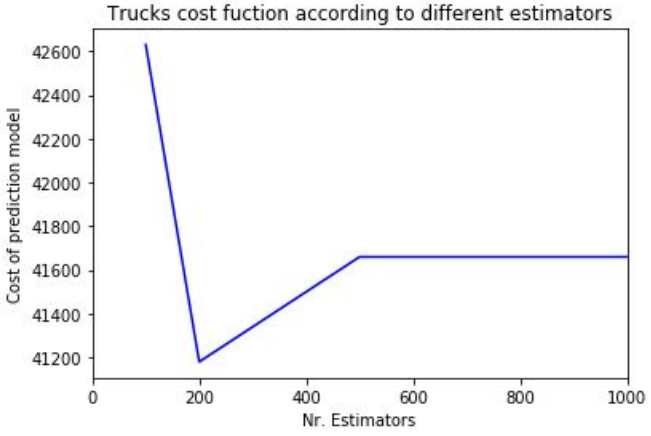
Accuracy score: 0.7384615384615385

Sensitivity: 0.5294117647058824

Specificity: 0.8125

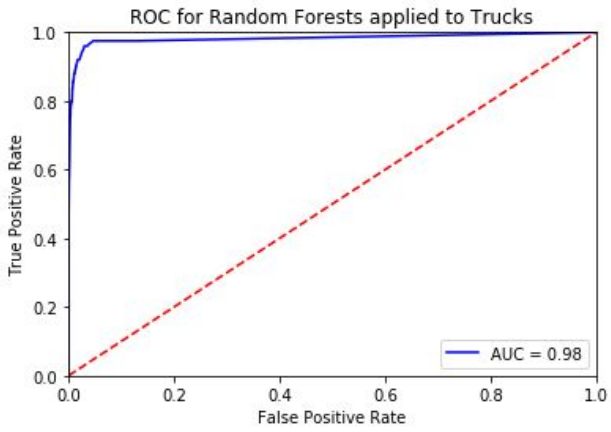
3.5 RANDOM FORESTS

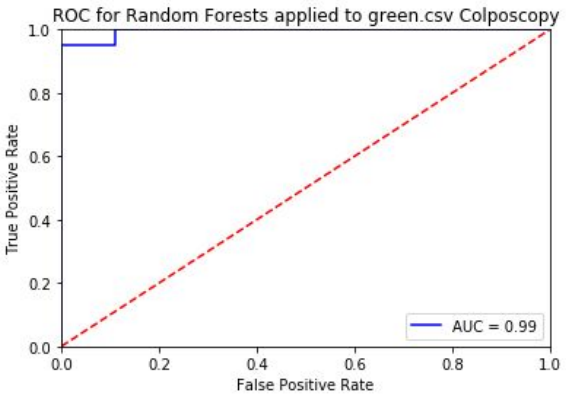
Em relação à classificação com random forests no problema do APS Failure, começamos por encontrar o número ideal de estimadores(decision trees). Corremos um ciclo de operações para 100, 200, 500 e 1000 estimadores. As diferenças em valores de accuracies não pareciam variar muito por isso recorremos à função custo do modelo dada pelo problema. Aí já foi possível observar um melhor desempenho dos 200 estimadores para o set de validação tendo sido então esse valor usado nos futuros resultados.

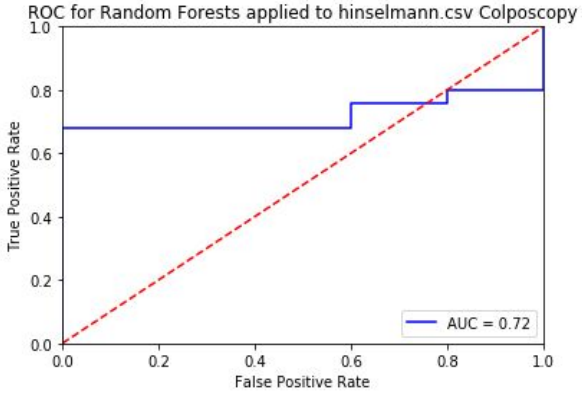


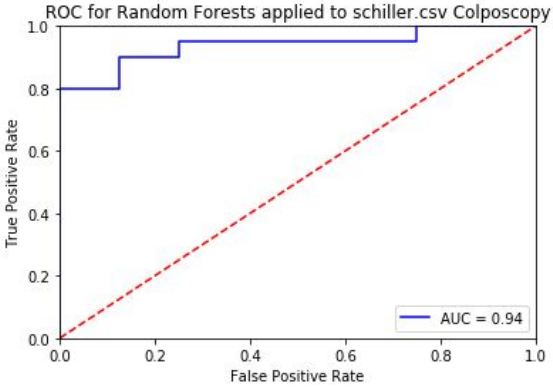
Com os thresholds devolvidos do roc_curve iteramos sobre estes tendo sempre em vista o menor valor possível para a função custo tendo acabado com um custo mínimo de 7560 para o threshold de 0.0450. Este valor é bem menor que o obtido inicialmente que rondava os 40000.

Na exploração do segundo problema não havendo uma função custo/objetivo inerente a este, recorremos simplesmente aos valores de accuracy para a escolha do número ideal de estimadores. Porém sendo os valores obtidos todos próximos ou iguais acabamos por optar por usar 200 estimadores simplesmente por consistência ao longo do projeto.

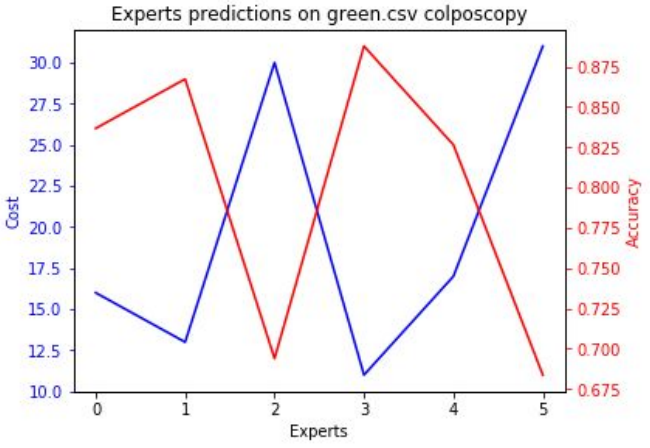








3.6 MÉTODOS DE EXPLORAÇÃO DE DADOS ADICIONAIS



Além dos tradicionais algoritmos de exploração de dados, devido à inexistência de uma função objetivo/custo para o problema da colposcopia, decidimos tomar os valores dos consensus como um resultado final certo e correto de diagnóstico dos vários samples fornecidos e comparamos estes valores com os diagnósticos dos vários experts de forma a avaliar o quão certos os experts tendem a estar. Apesar do expert mais correto variar de acordo com diferentes métodos de colposcopia foi possível obter uma accuracy média dos experts ao longo dos vários métodos.

Foi criado um custo simplificado ([0:1]) para este problema tendo os diagnósticos corretos (verdadeiros positivos e verdadeiros negativos) custo 0 e os diagnósticos errados (falsos positivos e falsos negativos) custo 1. Assim obtivemos também os valores médios deste custos para os experts em todos os métodos de colposcopia. Estes valores obtidos serviram então de comparação com as accuracies obtidas nos classificadores treinados referidos ao longo do relatório.

Green:	Hinselmann:	Schiller:
Mean costs: 19.66	Mean costs: 16.83	Mean costs: 20.16
Mean accuracies: 0.7993	Mean accuracies: 0.8264	Mean accuracies: 0.7807

3.7 RESULTS

Analisando os resultados apresentados nos gráficos dos diversos classificadores usados ao longo do projeto podemos concluir que os melhores modelos para o problema 1 foram obtidos para as decision trees e random forests tendo estes classificadores valores de AUC a 0.98. Os restantes classificadores apesar de não apresentarem resultados de AUC tão elevados, também obtiveram modelos com um bom comportamento, bem melhor que a predição randomizada.

Dado que os dois melhores classificadores deste problema têm características comuns poderia-se pensar que terá havido overfitting das features no entanto, recorrendo ao dataset de teste não usado no treino e ajuste de parâmetros do modelo de random forests continuamos a obter uma accuracy de 0.98 para este modelo.

Em relação ao segundo problema os resultados obtidos para o AUC foram geralmente melhores para o método green da colposcopia sendo que as random forests continuaram a ser o melhor modelo a prever os resultados corretos porém o mesmo não se pode dizer dos restantes métodos de colposcopia. No caso do método de schiller o KNN e o Naive Bayes ainda são capazes de fazer predições corretas mas com valores de AUC a raramente passarem os 0.70. No método de Hinselmann os modelos obtidos foram geralmente inúteis, tendo comportamentos semelhantes ou por vezes piores à escolha aleatória.

4. CONCLUSIONS

Conclui-se que para os resultados obtidos, ambos os *datasets* não são bem modelados com a aprendizagem não supervisionada. Assumimos por isso que este tipo de aprendizagem não têm grande vantagem de aplicação nas situações reais estudadas.

Já para aprendizagem supervisionada as conclusões variam para os dois problemas.

No caso do problema do APS dos camiões da Scania os modelos foram geralmente bons e a maioria deles poderia ser aplicado em contexto real permitindo à empresa poupar dinheiro nas reparações de camiões e evitar acidentes de prejuízos avultados. Porém no caso do problema da qualidade dos datasets de colposcopia, a maioria dos modelos criados não conseguiu obter resultados razoáveis sendo que ao compararmos com as accuracies de predições do experts, estes tendem sempre a ter um melhor desempenho que os modelos criados, fazendo com que não seja recomendável o uso destes modelos num sistema automático de diagnóstico em preferência da opinião de um especialista.

5. REFERENCES

[1] https://stats.stackexchange.com/questions/68893/area-under-curve-of-roc-vs-overall-accuracy#answer-68921

[2] https://stats.stackexchange.com/questions/229523/association-rules-support-confidence-and-lift