1. Considere o dataset ***labor***. Usando os algoritmos *J48*e *simpleCart* (Weka) apresente resultados (as diferentes árvores) que ilustram o uso das várias opções de pruning estudadas e.g. sub-tree raising, replacement, laplace, minimal complexity pruning, etc.

Algoritmo J48:

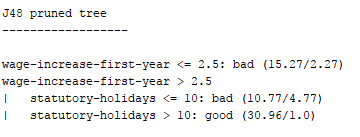


Figura 1 - J48 com pruning: sub-tree replacement **true**, sub-tree raising **true**.   
Laplace **false**

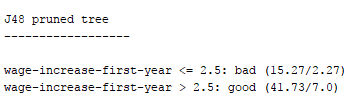


Figura 2 - J48 com pruning: sub-tree replacement **true**, sub-tree raising **false**. Laplace **false**

Uma imagem com texto

Descrição gerada com confiança alta

Relativamente ás árvores criadas para os casos analisados, é possível constatar que sempre que a função de pruning se encontra desativada, as árvores resultantes apresentam uma maior complexidade a nível de estrutura e profundidade. Isto acontece porque os métodos de sub-tree raising ou sub-tree replacement não são realizados, e, portanto, o modelo cresce conforme os dados de treino que lhe são fornecidos.

O uso do método de pruning torna-se assim vital para reduzir o risco de overffiting perante os dados de treino, levando à criação de uma árvore mais simples de compreender e iterar.

Dentro dos métodos de pruning, considerando que o método de sub-tree raising é mais exigente a nível de tempo de execução e complexidade, pode até afirmar-se que utilizar apenas o método de sub-tree replacement já é suficiente para reduzir significativamente a complexidade da árvore.

Perante a opção que permite a utilização do método de Laplace, pode constatar-se que o método não altera a estrutura da árvore, mas leva a melhores resultados, reduzindo o parâmetro associado ao erro *RMSE.*

Figura 3 - J48 sem pruning: sub-tree replacement **false**, sub-tree raising **false**. Laplace **false**

Algoritmo SimpleCart:

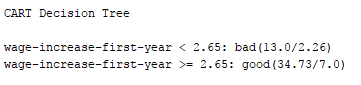


Figura 4 - SimpleCart com pruning.

Uma imagem com texto

Descrição gerada com confiança muito alta

Figura 5 - Figura 6 - SimpleCart sem pruning.

1. Apresente a justificação para os resultados obtidos com as avaliações por 10-xval e por training data.



Independente do algoritmo escolhido, é possível analisar que com o uso de dados de treino para teste, o rácio de erro diminui de forma acentuada.

Apesar dos parâmetros apresentados melhorarem quando se usa a opção “*Training set”*, estes dados de avaliação podem ser falaciosos para tirar conclusões, isto porque estamos a usar como dados de teste registos que já foram utilizados para treinar o modelo, sendo assim já conhecidos pelo mesmo. Desta forma, a árvore gerada vai modelar muito bem os dados de treino, mas será pouco apropriada para analisar novos casos em utilização futura.

1. Considere o dataset **soybean**. Apresente um estudo comparativo usando validação cruzada e os algoritmos NaiveBayes, SimpleCART e J48. Apresente conclusões sobre os melhores desempenhos para cada classe (das 19 existentes).



Figura 7 - Valores da área ROC por classe, usando três diferentes classifeirs

Como podemos analisar pelos valores obtidos na tabela apresentada, conforme a importância que seja necessária de atribuir a uma determinada classe, pode ser relevante o uso de um diferente algoritmo, que favoreça a precisão para os casos dessa classe.

Ou seja, em determinados casos um algoritmo pode ser vantajoso mas, para o estudo de outras classes ou de um outro contexto, essa algoritmo pode já não ser o mais indicado.

De uma forma geral, para o dataset SoyBean, o classifer *Naive Beyes*foi em média o que obteve melhores resultados para qualquer classe. Isto deve-se ao facto do algoritmo olhar para cada classe de forma independente, ao contrário do que acontece com os dois algoritmos de árvores também em análise.

1. Escolha uma classe e 2 classificadores onde o valor de AUC contraria o valor do erro nessa classe. Apresente todos os resultados.

Considerando como classifiers os algoritmos Naive Bayes e J48 (com uso de LaPlace), foi possível analisar que para a classe j = bacterial-pustule, o valor da área da curva contradiz o valor do erro nessa classe.

Dado que em Naive Bayes com 2 previsões erradas a área da curva ROC = 1, seria de esperar que no J48, dado que tem menos previsões erradas, a área da curva fosse superior. Contudo, o seu valor é inferior a 1.



Figura 8 - Comparação dos valores ROC Area e rácio de erros para a classe bacterial-pustule

Isto indica que a medida da área da curva ROC é mais robusta para avaliar um modelo, dado que o número de previsões erradas pode variar conforme o método de teste ou os novos dados que sejam avaliados no futuro com o modelo