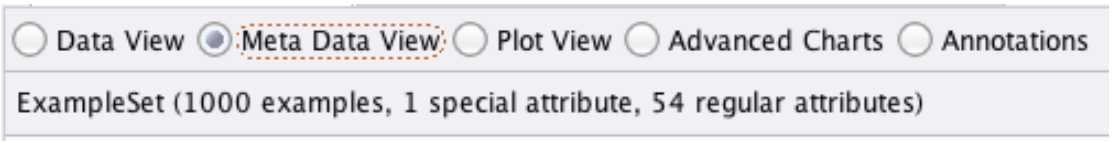
Pedro Miguel Correia Leite
PG25330
Data Mining

Tarefa 1

a)



Tal como se pode analisar pela imagem anterior, o data set pFC tem um total de 1000 instâncias, possuindo 54 atributos regulares e 1 atributo especial, sendo que este atributo especial corresponde ao atributo "cover type", que foi considerado como sendo a label (classe objetivo). Este foi definido como polinomial. Além disso, todos os Soil_Type* e Wilderness_Area* foram definidos como binomiais. Os restantes são numéricos.
Este data set não possui missing values.

b)

Relativamente aos numéricos, possuem as médias e respetivos desvios padrão:

Role	Name	Type	Statistics	Range	Missings
regular	Slope	integer	avg = 11.262 +/- 6.023	[1.000 ; 40.000]	0
regular	Hillshade_3pm	integer	avg = 139.768 +/- 29.329	[28.000 ; 240.000]	0
regular	Aspect	integer	avg = 141.844 +/- 108.431	[0.000 ; 359.000]	0
regular	Hillshade_9am	integer	avg = 218.274 +/- 20.804	[115.000 ; 254.000]	0
regular	Hillshade_Noon	integer	avg = 225.413 +/- 14.616	[141.000 ; 254.000]	0
regular	Horizontal_Distance_To_Hydrol	integer	avg = 236.588 +/- 189.965	[0.000 ; 997.000]	0
regular	Elevation	integer	avg = 2867.562 +/- 173.698	[2486.000 ; 3267.000]	0
regular	Vertical_Distance_To_Hydrolog	integer	avg = 30.960 +/- 37.224	[-45.000 ; 245.000]	0
regular	Horizontal_Distance_To_Fire_Pc	integer	avg = 3184.821 +/- 1746.28	[120.000 ; 6853.000]	0
regular	Horizontal_Distance_To_Roadw	integer	avg = 3351.210 +/- 1678.14	[67.000 ; 6890.000]	0

Relativamente aos não numéricos, possuem as seguintes modas:

Role	Name	Type	Statistics	Range	Missings
regular	Wilderness_Area2	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Wilderness_Area3	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Wilderness_Area4	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type1	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type2	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type3	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type4	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type5	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type6	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type7	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type10	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type11	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type13	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type14	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type15	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type17	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type21	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type25	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type26	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type27	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type28	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type31	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type32	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type33	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type34	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type35	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type36	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type37	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type38	binominal	mode = false (1000), least = t	false (1000), true (0)	0
regular	Soil_Type29	binominal	mode = false (567), least = tru	false (567), true (433)	0
regular	Soil_Type12	binominal	mode = false (827), least = tru	false (827), true (173)	0

Uma vez que o data set é bastante extenso, a imagem seguinte apresenta os restantes atributos não numéricos que não couberam na imagem anterior.

regular	Soil_Type12	binominal	mode = false (827), least = tru	false (827), true (173)	0
regular	Soil_Type30	binominal	mode = false (865), least = tru	false (865), true (135)	0
regular	Soil_Type23	binominal	mode = false (898), least = tru	false (898), true (102)	0
regular	Soil_Type20	binominal	mode = false (951), least = tru	false (951), true (49)	0
regular	Soil_Type18	binominal	mode = false (954), least = tru	false (954), true (46)	0
regular	Soil_Type16	binominal	mode = false (975), least = tru	false (975), true (25)	0
regular	Soil_Type24	binominal	mode = false (984), least = tru	false (984), true (16)	0
regular	Soil_Type19	binominal	mode = false (993), least = tru	false (993), true (7)	0
regular	Soil_Type9	binominal	mode = false (995), least = tru	false (995), true (5)	0
regular	Soil_Type22	binominal	mode = false (997), least = tru	false (997), true (3)	0
regular	Soil_Type40	binominal	mode = false (997), least = tru	false (997), true (3)	0
regular	Soil_Type39	binominal	mode = false (998), least = tru	false (998), true (2)	0
regular	Soil_Type8	binominal	mode = false (999), least = tru	false (999), true (1)	0
regular	Wilderness_Area1	binominal	mode = true (1000), least = fa	false (0), true (1000)	0

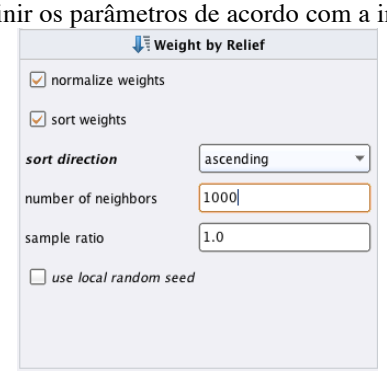
c)

Role	Name	Type	Statistics	Range	Missings
regular	Slope	integer	avg = 11.262 +/- 6.023	[1.000 ; 40.000]	0
regular	Hillshade_3pm	integer	avg = 139.768 +/- 29.329	[28.000 ; 240.000]	0
regular	Aspect	integer	avg = 141.844 +/- 108.431	[0.000 ; 359.000]	0
regular	Hillshade_9am	integer	avg = 218.274 +/- 20.804	[115.000 ; 254.000]	0
regular	Hillshade_Noon	integer	avg = 225.413 +/- 14.616	[141.000 ; 254.000]	0
regular	Horizontal_Distance_To_Hydrol	integer	avg = 236.588 +/- 189.965	[0.000 ; 997.000]	0
regular	Elevation	integer	avg = 2867.562 +/- 173.698	[2486.000 ; 3267.000]	0
regular	Vertical_Distance_To_Hydrology	integer	avg = 30.960 +/- 37.224	[-45.000 ; 245.000]	0
regular	Horizontal_Distance_To_Fire_Pc	integer	avg = 3184.821 +/- 1746.28	[120.000 ; 6853.000]	0
regular	Horizontal_Distance_To_Roadw	integer	avg = 3351.210 +/- 1678.14	[67.000 ; 6890.000]	0

Tal como se pode analisar na imagem anterior, há dois atributos que se aproximam bastante em termos de amplitude, sendo eles o "Horizontal_Distance_To_Fire_Points" e o "Horizontal_Distance_To_Roadways", contudo, a amplitude do atributo destacado na imagem "Horizontal_Distance_To_Roadways", apresenta uma amplitude superior, relativamente a todas as restantes.

d)

Para responder a esta questão, usei um dos operadores presentes em "Modeling" -> "Attribute Weighting" -> "Weight by Relief". Optei pela utilização deste, porque é o sugerido pelo administrador do blog "rapid-i", e pela documentação verifiquei que é o mais utilizado devido à sua baixa complexidade e alta fiabilidade.
Para este componente, decidi definir os parâmetros de acordo com a imagem seguinte:



O motivo pelo qual o parâmetro "number of neighbors" está aumentado, deve-se ao facto de que utilizando este valor, serão necessários mais nós "vizinhos" para se definir o resultado final, o que oferece um acréscimo de fiabilidade, uma vez que pode haver algum valor que possa ter sofrido uma ligeira entropia.
Deste modo, verificou-se que os atributos úteis, para o data set pFC são os seguintes:

Elevation	1
Hillshade_Noon	0.348
Hillshade_9am	0.208
Soil_Type12	0.160
Soil_Type29	0.154
Aspect	0.107
Vertical_Distance_To_Hydrology	0.096
Horizontal_Distance_To_Roadways	0.087
Horizontal_Distance_To_Fire_Points	0.083
Soil_Type30	0.059
Horizontal_Distance_To_Hydrology	0.052
Slope	0.048

Tal como se pode analisar na imagem anterior, estão presentes mais de 5 atributos úteis, sendo que apenas vou analisar os atributos de "Elevation", "Hillshade_Noon", "Hillshade_9am", "Aspect" e "Horizontal_Distance_To_Roadways".
Se o objetivo é identificar o "cover type", os atributos relativos à elevação (Elevation) e à sombra presente na colina (Hillshade_Noon e Hillshade_9am) são fatores essenciais para identificar, o "cover type", pois estes influenciam, e muito, por exemplo, a qualidade do terreno, e por si só identificam o seu tipo. Apesar do tipo de solo ser bastante importante, considero o aspeto e a distância vertical para as estradas fatores mais decisivos, e tal como se pode observar, a importância entre eles possui uma diferença bastante reduzida.

Relativamente aos atributos inúteis, importa referir que existem bastantes, uma vez que existem mais de 10 atributos cujo "weight" é 0, contudo, para uma melhor análise, recorri às potencialidades da ferramenta, e ordenei por ordem crescente de pesos, obtendo assim os seguintes resultados:

attribute	weight
Wilderness_Area1	0
Wilderness_Area2	0
Wilderness_Area3	0
Wilderness_Area4	0
Soil_Type1	0

Tal como referi anteriormente e como só são pedidos 5 atributos, decidi inserir os 5 menos importantes, porque pela minha análise, faz todo o sentido os valores que foram obtidos. Relativamente à "Wilderness_Area*", não são considerados como fatores de decisão, porque estes, tal como diz na descrição do data set em http://archive.ics.uci.edu/ml/datasets/Covertype, estas áreas estão todas localizadas no "Roosevelt National Forest", portanto, dá a entender que os valores serão bastante aproximados independentemente da área escolhida, além disso, o tipo de solo 1 que está representado terá a mesma análise. Na minha opinião, fatores como o tipo de solo, por exemplo, serão bastante aproximados. Além de que após um "browse" sobre o data set, verifiquei que a maioria dos valores binomiais estão todos muito semelhantes, o que faz a minha análise ter mais relevância.

e)

Fazendo a mesma análise da alínea anterior, mas agora com o data set rFC, verifica-se que os atributos não se mantêm, o que eu já estava a prever, porque uma coisa é fazer uma análise com 1000 registos ordenados de acordo com a sua inserção e outra é fazer a análise com 1000 registos retirados do data set original por ordem, quase que, aleatória. No conjunto dos dados pFC, como os dados são retirados por ordem sequencial (primeiros mil registos), após um "browse" no data set, dá me a sensação que estes são bastante similares porque a análise é relativa à mesma área, isto é, foi realizada a análise por áreas, e os primeiros 1000 registos foram os primeiros 1000 dados retirados desta análise. Relativamente ao rFC já não se pode dizer o mesmo, porque como os dados são retirados de 100 em 100, é fácil de concluir que os dados obtidos são de áreas diferentes, e por isso têm uma maior dispersão de informação.

Relativamente aos atributos úteis:

attribute	weight
Hillshade_Noon	1
Elevation	0.845
Hillshade_9am	0.447
Soil_Type12	0.339
Soil_Type29	0.318
Vertical_Distance_To_Hydrology	0.253
Aspect	0.222
Soil_Type30	0.197
Horizontal_Distance_To_Roadways	0.191
Horizontal_Distance_To_Fire_Points	0.153
Wilderness_Area1	0.149
Horizontal_Distance_To_Hydrology	0.105
Slope	0.096

De acordo com a imagem anterior, verifico que o "Hillshade_Noon" e "Elevation" encontram-se no topo da importância, apesar de estarem agora em ordem inversa. O aspeto também está, apesar de ter passado de 6º para 7º em termos de importância, contudo, tem um peso superior. Relativamente ao "Slope" também tem um peso superior mas já não pertence aos mais importantes. Aqui considero como mais importantes, "Hillshade_Noon", "Elevation", "Hillshade_9am", "Vertical_Distance_To_Hidrology" e "Aspect".

Relativamente aos menos úteis:

attribute	weight
Soil_Type15	0
Soil_Type21	0
Soil_Type25	0
Soil_Type27	0
Soil_Type28	0

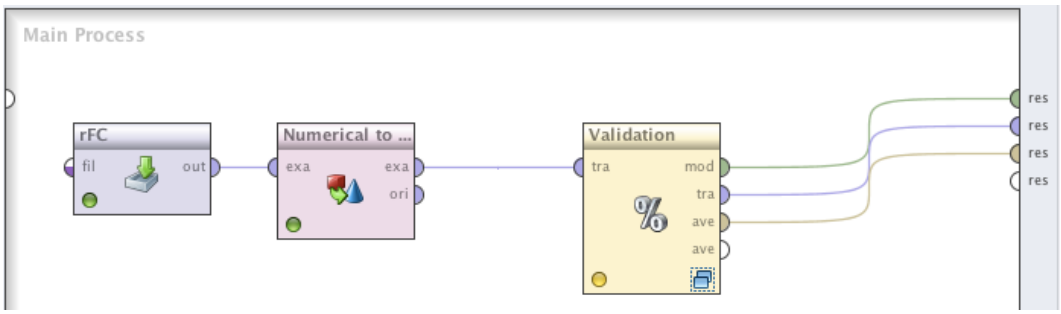
Aqui considero os que foram sugeridos pelo RapidMiner, e que estão presentes na imagem anterior. É fácil de verificar que os atributos de "Wilderness_Area*" já não estão presentes nesta imagem, o que dá mais ênfase à conclusão que retirei anteriormente. Como o pFC tinha os registos em abundância relativos a um tipo de área, então a escolha da mesma não era importante, mas neste caso (data set rFC), como os registos são mais aleatórios, a escolha da área já é um fator pertinente para a obtenção de resultados. Verifica-se que o Wilderness_Area1 passou de último para um dos primeiros, ordenados por ordem decrescente de importância.

Tarefa 2

a)

Para esta tarefa optei pela utilização das árvores de decisão, uma vez que são as mais adequadas à classificação, permitindo obter resultados de forma mais rápida que outras técnicas. É óbvio que estas podem em alguns casos apresentar resultados inferiores a outras técnicas, mas os objetivos destas podem passar por uma melhor compreensão do modelo, uma vez que as regras que induzem, possuem a vantagem de uma fácil tradução para a linguagem humana. Deste modo, são mais compreensivas quando comparadas com outras técnicas.

Além disto, decidi colocar este modelo aninhado dentro de um split validation, com a finalidade de conseguir definir a percentagem de resultados utilizados para teste. Neste caso utilizarei 70% para treino e 30% para teste, utilizando ainda um particionamento relativo com "shuffled sampling". Importa ainda salientar que para o caso das decision tree, utilizarei o critério de information gain, colocando uma profundidade máxima de 5, com o número de preprunning a 1. Não utilizei o gain ratio, uma vez que esta faz o rácio da information gain utilizando para isso a sua informação intrínseca, e não é isso que pretendo. Apenas quero analisar o ganho de informação com este particionamento. O modelo utilizado está presente na imagem seguinte:



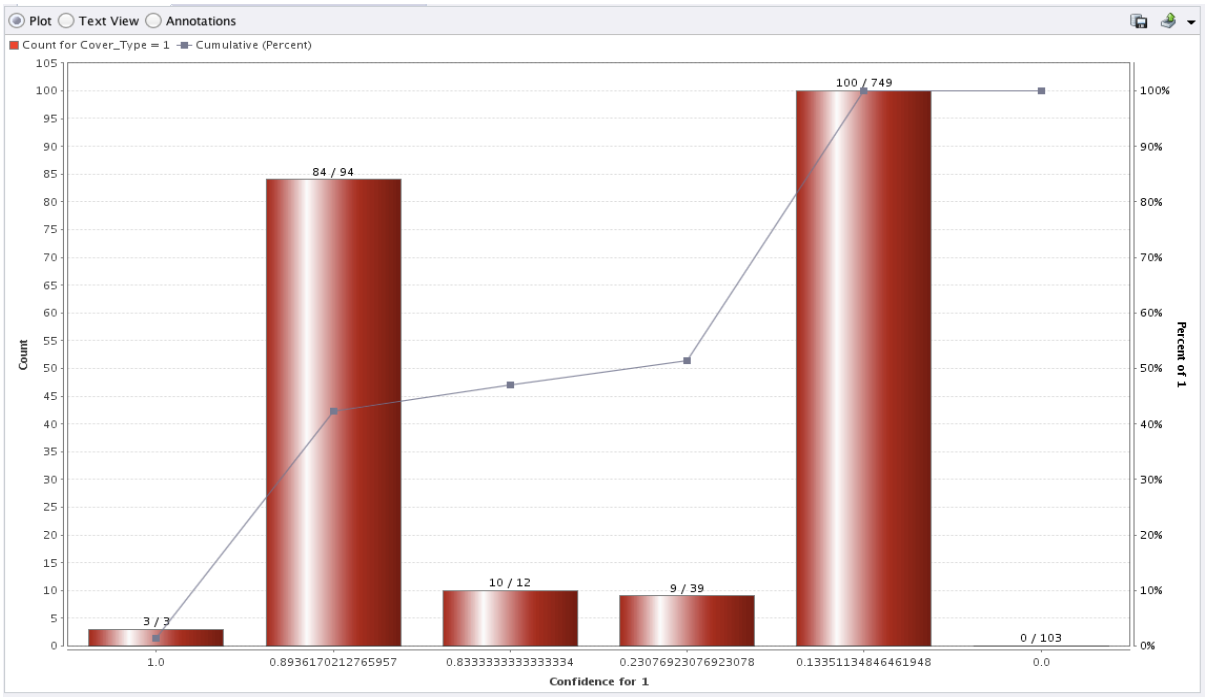
Com estes parâmetros consegui obter resultados bastante bons, possuindo uma acurácia de 83,67%. Importa salientar que estes resultados foram obtidos ainda sem otimizações, sendo que se for otimizado é garantido que esta proporção do número total de previsões será melhorado, diminuindo assim o erro.

accuracy: 83.67%

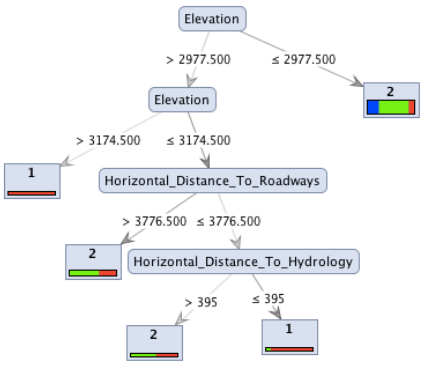
	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	0	2	0	1	0	0	0	0.00%
pred. 2	4	210	29	0	0	0	0	86.42%
pred. 1	1	2	26	0	2	0	0	83.87%
pred. 6	1	1	0	1	0	0	0	33.33%
pred. 7	0	0	0	0	5	0	0	100.00%
pred. 4	0	0	0	0	0	8	3	72.73%
pred. 3	0	0	0	1	0	2	1	25.00%
class recall	0.00%	97.67%	47.27%	33.33%	71.43%	80.00%	25.00%	

Analisando a imagem anterior, verifica-se que que a proporção de casos positivos corretamente identificados (class precision) não está muito boa para a previsão das classes 5 (não acertou em nenhum), 6 (acertou em 33,33%) e 3 (acertou em 25%). Além desses, as restantes classes possuem

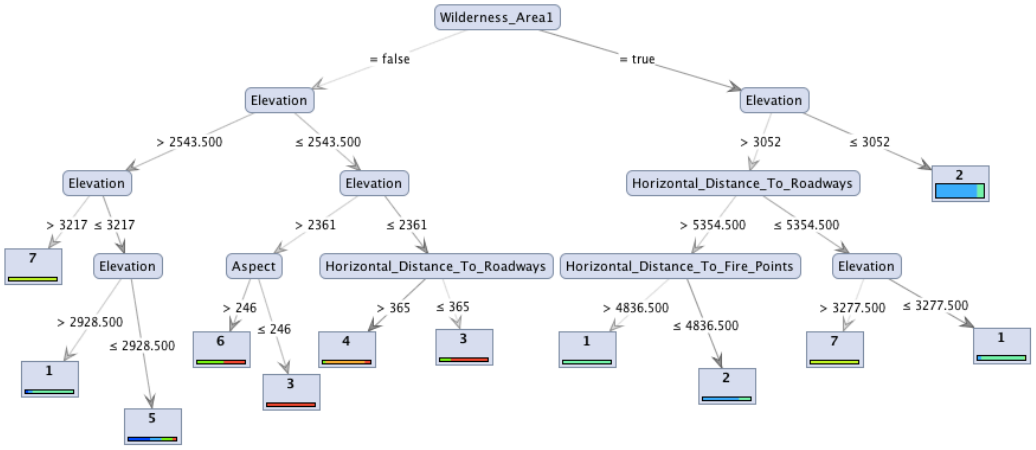
valores razoáveis, tendo a classe 7, o melhor valor, possuindo 100% dos casos acertados. Relativamente à proporção dos casos reais que foram corretamente identificados (class recall), pode verificar-se que analogamente à "class prediction", estes também ainda não estão muito precisos, verificando que no caso da classe 1, a percentagem continua a 0%. Apesar de tudo, concluo que este modelo revela alguma confiança, uma vez que tem um erro de sensivelmente 16%, o que é bastante bom. Um erro de 16% neste volume de dados, é algo que obviamente terá de ser otimizado mas que revela uma boa que terá uma boa performance na previsão do tipo de cobertura florestal.
O problema deste modelo é que para conseguir que a classe objetivo obtenha os 100%, ainda é bastante custoso, tal como se pode verificar na imagem apresentada de seguida.



b)
Analisando os dois data sets com os mesmos parâmetros, posso dizer que fiquei um pouco surpreendido, com os resultados obtidos. Já sabia que o pFC daria um valor inferior de acurácia, mas não estava à espera que fosse tão reduzido. Aqui, a acurácia do pFC é de cerca de 60% enquanto que o do rFC é de cerca de 83%. Também a sua performance é similar nos dois casos, tendo uma maior percentagem de acertos no caso do data set do rFC.
Árvore de decisão do pFC:



Árvore de Decisão do rFC:



Analisando as duas árvores importa salientar que ambas foram geradas com o mesmo critério e com a mesma definição, contudo as duas árvores são bastante diferentes. Isto é normal, e vai de encontro ao que introduzi anteriormente, uma vez que o pFC é relativo a uma área mais localizada, daí não ter na sua árvore (que foi definida com uma profundidade máxima de 5), o atributo relativo à área. Também se verifica que esta árvore só possui as classes 1 e 2. Com base nisto, e com a baixa acurácia (que implica um erro superior), concluo que este não é o melhor modelo para se obter conclusões.

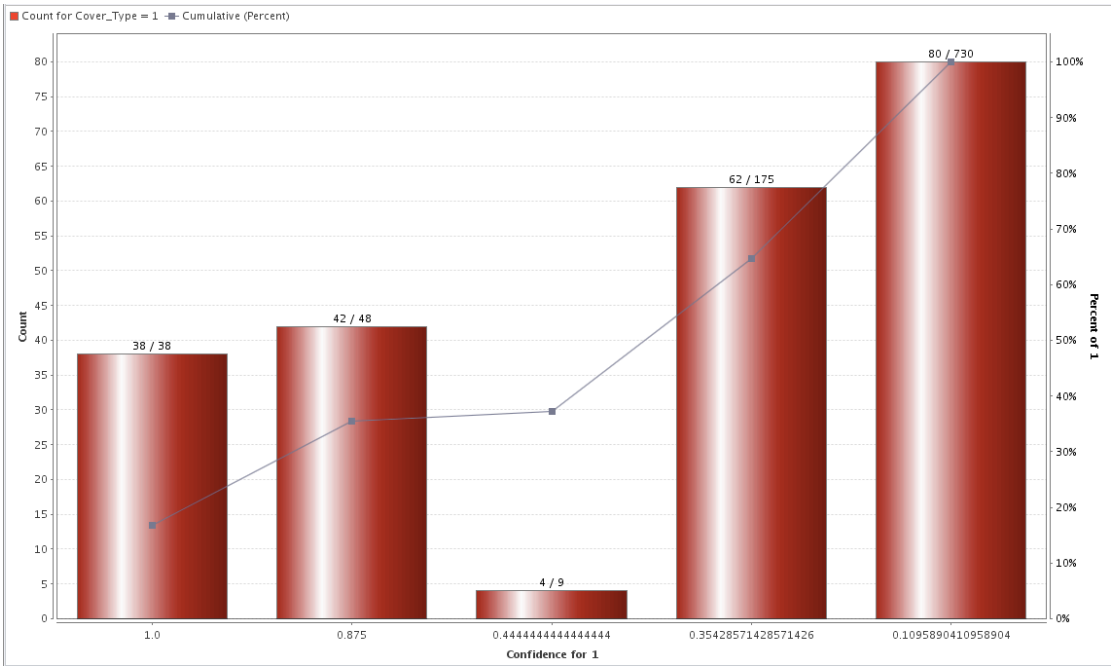
Relativamente à segunda árvore (relativa ao data set rFC), verifica-se que esta tem um conjunto de dados mais abrangente, englobando mais classes (1,2,3,4,5,6,7), melhor dizendo, englobando todas as classes. Nesta árvore é possível verificar ainda que a análise tem como primeiro fator decisivo a área em que se pretende influenciar o processo de decisão. Por exemplo, se for objetivo de análise a área 1, então a elevação deve ser centrado entre >3052 e <=3052, no caso de não se pretender a área 1, então a elevação deve estar entre >2543,5 e <=2543,5. Com base neste parágrafo e com a acurácia que foi analisada no início desta alínea (aproximadamente 83%), então verifica-se que o data set que deve ser utilizado deve ser o rFC.

Em suma, o pFC é relativo a uma área mais localizada, e o rFC é relativo a uma área mais genérica.

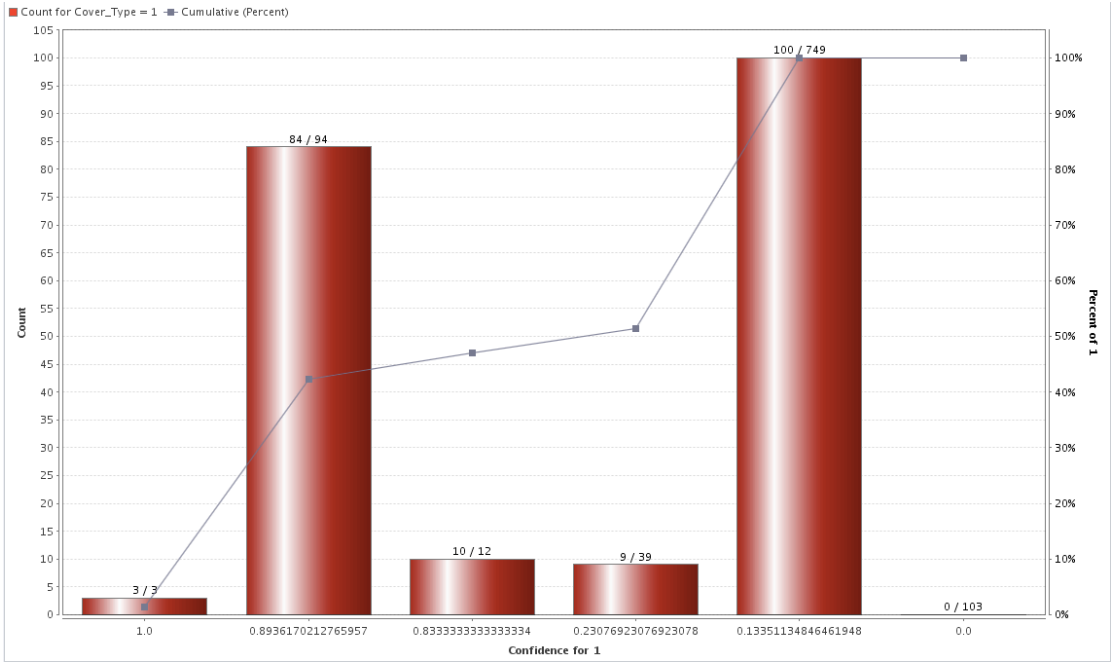
c)

Para esta alínea optei pela utilização do lift chart que é de fácil interpretação, e não é complexo. Importa referir que para esta alínea, defini como target a classe 1 para a geração do lift chart.

Lift Chart pFC:



Lift Chart rFC:



Ao gerar estes lift chart, verifico que estes resultados vão de encontro a tudo o que tenho dito anteriormente, uma vez que o rFC apesar de apresentar uma confiança um pouco baixa, consegue ter uma confiança superior ao pFC. Contudo, em ambos os casos, só se consegue atingir a percentagem de 100% bastante tarde, e tal como se sabe, o que se pretende é atingir esta percentagem o mais rapidamente possível.

Tarefa 3

Ao começar este processo, utilizou-se um split validation e uma decision tree, obtendo-se assim uma acurácia de 83,67%, tal como se pode verificar na imagem seguinte.

accuracy: 83.67%

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	0	2	0	1	0	0	0	0.00%
pred. 2	4	210	29	0	0	0	0	86.42%
pred. 1	1	2	26	0	2	0	0	83.87%
pred. 6	1	1	0	1	0	0	0	33.33%
pred. 7	0	0	0	0	5	0	0	100.00%
pred. 4	0	0	0	0	0	8	3	72.73%
pred. 3	0	0	0	1	0	2	1	25.00%
class recall	0.00%	97.67%	47.27%	33.33%	71.43%	80.00%	25.00%	

Começando por fazer o "discretize by binning", verifiquei que a acurácia diminuiu para 75%, de modo a que optei por não utilizar este componente, contudo, no final da otimização será tentado novamente para verificar se este oferece alguma melhoria. O resultado pode ser analisado na seguinte imagem:

accuracy: 75.33%

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	1	0	0	0	0	0	0	100.00%
pred. 2	4	212	55	0	2	0	0	77.66%
pred. 1	1	2	0	0	3	0	0	0.00%
pred. 6	0	0	0	1	0	0	0	100.00%
pred. 7	0	0	0	0	2	0	0	100.00%
pred. 4	0	0	0	0	0	8	2	80.00%
pred. 3	0	1	0	2	0	2	2	28.57%
class recall	16.67%	98.60%	0.00%	33.33%	28.57%	80.00%	50.00%	

Posteriormente tentei trocar o split validation pelo x-validation, uma vez que o x-validation é ideal no caso de haverem poucos registos, criando k grupos de valores, usando k-1 para treino e o outro para teste. Contudo, com esta alteração consegui uma acurácia de 82,3%, enquanto que com o split validation consegui obter uma acurácia de 83,67.

accuracy: 82.30% +/- 38.17% (mikro: 82.30%)

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	12	5	0	7	0	0	2	46.15%
pred. 2	11	666	112	0	0	0	0	84.41%
pred. 1	1	12	94	0	1	0	0	87.04%
pred. 6	2	2	0	2	0	1	5	16.67%
pred. 7	0	0	0	0	20	0	0	100.00%
pred. 4	0	0	0	4	0	24	6	70.59%
pred. 3	0	0	0	2	0	4	5	45.45%
class recall	46.15%	97.23%	45.63%	13.33%	95.24%	82.76%	27.78%	

Com isto, decidi utilizar um componente para otimizar a seleção, e com este consegui uma melhoria de resultados. Utilizei o "Optimize Selection". Aqui defini o número mínimo de atributos usados para combinações com o valor 1. O tamanho da população passou para 11, uma vez que tentei com vários e foi com este que obtive o melhor resultado. Pus ainda os tamanhos a serem normalizados e defini o esquema de seleção como "roulette whell". Para finalizar coloquei o tipo do crossover a "shuffle" e o p crossover a 0.5. Relativamente ao modelo de decision tree que estava a utilizar coloquei a confiança a 0,05 com o critério de information gain, com o ganho mínimo de 0,1. Com estes parâmetros consegui passar de 83,67% para 87%, tal como se pode observar na imagem seguinte:

accuracy: 87.00%

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	3	0	0	1	0	0	1	60.00%
pred. 2	5	219	22	3	0	0	0	87.95%
pred. 1	0	3	27	0	0	0	0	90.00%
pred. 6	0	0	0	1	0	0	0	100.00%
pred. 7	0	0	0	0	5	0	0	100.00%
pred. 4	0	0	0	0	0	5	3	62.50%
pred. 3	0	0	0	0	0	1	1	50.00%
class recall	37.50%	98.65%	55.10%	20.00%	100.00%	83.33%	20.00%	

Para finalizar, decidi utilizar novamente a sugestão dado no enunciado, contudo, obtive novamente um decréscimo de acurácia, assim como a percentagem de previsão piorou na maioria das classes, tal como se pode ver pela imagem seguinte:

Table View Plot View

accuracy: 78.33%

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	3	1	2	0	2	0	0	37.50%
pred. 2	2	218	49	0	1	0	0	80.74%
pred. 1	0	0	0	0	0	0	0	0.00%
pred. 6	0	1	0	2	0	1	1	40.00%
pred. 7	1	0	0	0	3	0	0	75.00%
pred. 4	0	0	0	1	0	8	2	72.73%
pred. 3	0	0	0	1	0	0	1	50.00%
class recall	50.00%	99.09%	0.00%	50.00%	50.00%	88.89%	25.00%	

Dados estes resultados, consegui otimizar o modelo passando de 83,67% para 87% em termos de acurácia. Também o erro diminuiu de 16,33% para 13%.
Relativamente à classe 5, antes não se conseguia prever nada, pois estava a 0%, e neste momento já se conseguem prever corretamente 60% dos casos que foram positivamente identificados e também se preveem 37,5% dos casos reais que foram corretamente identificados. Este foi o aumento mais significativo que consegui, no entanto, também importa salientar que todas as previsões têm uma maior confiabilidade, sendo que no caso da classe 6, a previsão passou de 33,33% de assertividade para 100% de assertividade.

accuracy: 83.67%

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	0	2	0	1	0	0	0	0.00%
pred. 2	4	210	29	0	0	0	0	86.42%
pred. 1	1	2	26	0	2	0	0	83.87%
pred. 6	1	1	0	1	0	0	0	33.33%
pred. 7	0	0	0	0	5	0	0	100.00%
pred. 4	0	0	0	0	0	8	3	72.73%
pred. 3	0	0	0	1	0	2	1	25.00%
class recall	0.00%	97.67%	47.27%	33.33%	71.43%	80.00%	25.00%	

Figure 1: Resultados antes da otimização

accuracy: 87.00%

	true 5	true 2	true 1	true 6	true 7	true 4	true 3	class precision
pred. 5	3	0	0	1	0	0	1	60.00%
pred. 2	5	219	22	3	0	0	0	87.95%
pred. 1	0	3	27	0	0	0	0	90.00%
pred. 6	0	0	0	1	0	0	0	100.00%
pred. 7	0	0	0	0	5	0	0	100.00%
pred. 4	0	0	0	0	0	5	3	62.50%
pred. 3	0	0	0	0	0	1	1	50.00%
class recall	37.50%	98.65%	55.10%	20.00%	100.00%	83.33%	20.00%	

Figure 2: Resultados após otimização