1) Considere a base de dados *FL_P2* no diretório

P: EAESP Abraham_Laredo MBA BIG DATA PDF CAPITULOS

(SENHA: alfa&beta11).

Variável alvo: STATUS (Note que você não precisa transformar / discretizar nenhuma variável. Utilize-as como estão.)

a) Sem dividir a amostra em duas partes, utilizar o classificador CART (*rpart*) do R para obter a árvore de decisão. Pode a árvore se conveniente. Com a árvore de decisão resultante estime a probabilidade de que o cliente abaixo ("Juquinha") seja "MAU".

Resp: 9,8691% (4 casas decimais)

CLIENTE	IDADE	UF	RESTR	QUANTI	NET
Juquinha	31	SP	SIM	NAO	NAO

b) Calcule a área AUROC para a árvore obtida no item anterior.

Resp: 0,745 (3 casas decimais)

c) qual a taxa de erro estimada pelo método cross validation aplicado a essa amostra?

Resp: 15,173% (3 casas decimais)

d) Obtenha a árvore (sem podar) fixando 100 como número mínimo de indivíduos em um nó terminal. Qual a probabilidade estimada que Cuevas seja "MAU"?

CLIENTE	IDADE	UF	RESTR	QUANTI	NET
Cuevas	31	SP	SIM	SIM	NAO

Resp: 29,386% (3 casas decimais)

[VALOR: 4.00 ponto(s)]

2) Discretize a variável IDADE em 6 classes de mesmas frequências. Gere a variáve kidade.

a) Apresente a distribuição bivariada parcial das frequências absolutas (kidade x STATUS), especificando os limites das duas primeiras classes e as frequências absolutas de BOM e MAU em cada uma destas duas primeiras classes.

	kidade	BOM	MAU
1	[18, 31)	599	258
2	[31, 37)	650	188

b) se Você fosse obrigado a fundir duas ou mais classes adjacentes, quais fundiria? Em caso positivo, dê os limites das novas classes e justifique quantitativamente sua decisão.

Dadas as proporções entre bons e maus entre as classes, eu não faria nenhuma fusão. O que pode ser feito é ajustar os limites inferiores e superiores das classes, considerando que a primeira classe passaria a ser [0, 31) e a última [54, ∞). Essas alterações não tem impacto nas frequências de bons e maus.

(VALOR: 2.00 pontos)

	BOM	MAU	BOM/MAU
[18, 31)	599	258	2,32
[31, 37)	650	188	3,46
[37, 41)	637	113	5,64
[41, 47)	691	183	3,78
[47, 54)	671	101	6,64
[54, 88]	597	110	5,43

3) **Considere a mesma planilha de dados da questão 1 (FL_P2) . Considere a amostra como um todo sem dividi-la em duas partes. Não discretize variáveis, não faça fusões nem elimine observações. Não transforme as variáveis previsoras**

a) Obtenha a regressão logística utilizando todas as cinco variáveis previsoras. Qual a probabilidade estimada de que o cliente 1019 seja "**MAU**" .

Resp: 67,097% (3 casas decimais)

b) Calcule a probabilidade estimada de que o indivíduo seguinte, Pepe seja "**MAU**"

Resp: 14,366 % (3 casas decimais)

CLIENTE	IDADE	UF	RESTR	QUANTI	NET
Pepe	34	SP	NAO	NAO	NAO

c) Calcule a estatística KS para o modelo obtido. Resp: 0,537 (3 casas decimais)

d) Agora selecione as variáveis utilizando o comando "step". Alguma variável foi removida? NÃO . Qual o intercepto do novo modelo? Resp: O mesmo (-0,574) (3 casas decimais)

(VALOR: 4,00 ponto(s))

1) Considere os seguintes resultados ao determinar uma árvore de decisão:

```
1) root 3368 1396 mau (0.4144893 0.5855107)
  2) EDUC=secundária 2094  978 bom (0.5329513 0.4670487)
    4) PRIM_EMP=não 921  274 bom (0.7024573 0.2975027) *
    5) PRIM_EMP=sim 1173  469 mau (0.3998295 0.6001705)
     10) TESTE>=79.5 386  160 bom (0.5854922 0.4145078) *
     11) TESTE< 79.5 787  243 mau (0.3087675 0.6912325) *
  3) EDUC=superior 1274  280 mau (0.2197802 0.7802198) *

Root node error: 1396/3368 = 0.41449

n= 3368

        CP     nsplit   rel error  xerror     xstd
1 0.133596      0      1.00000 1.00000 0.020480
2 0.047278      2      0.73281 0.75215 0.019257
3 0.010000      3      0.68553 0.70344 0.018894
```

a) Qual a probabilidade de que os indivíduo abaixo seja "mau"

a) Qual a probabilidade de que os indivíduo abaixo seja "mau"

IDADE	ECIV	DIST_EMP	TIPORESID	PRIM_EMP	TESTE	EDUC
25 - 45	casado	média	propria	Sim	Micro par: 70 Micro impar: 82	superior

b) Qual a taxa de erro estimada com o método de cross validation?

2) Considere o arquivo "dilei senha" no

P: EAESP Abraham_Laredo MBA... (senha: apocalipse16)

Rode a regressão logística considerando apenas as variáveis TIPORESID+PRIM_EMP+TESTE+EDUC
A variável alvo é STATUS (transforme em ALVO conforme abaixo especificado). Não divida a amostra
em duas partes e não discretize as variáveis.

a) qual o valor do intercepto do modelo?

b - **MICRO IMPAR**) qual a probabilidade de que o funcionário X1012 seja "mau"?

c) Selecione as variáveis utilizando uma função apropriada do R.

3) Os outputs estimados com uma rede neural foram discretizados em 4 categorias. A tabela seguinte mostra a quantidade de bons e maus em cada categoria. Considere a tabela especificada pelo número de seu micro

MICRO PAR			
	bom	mau	
A	1100	730	
B	1240	610	
C	1450	390	
D	1600	240	

MICRO IMPAR			
	bom	mau	
A	100	1730	
B	240	1610	
C	450	1390	
D	600	1240	

a) Com base nesses resultados calcule o valor do KS correspondente.

b) qual a probabilidade estimada que um indivíduo da segunda classe (B) seja "bom"?