Introdução à Aprendizagem Estatística - Fase 1

Membros: Luís Gomes (A78701), Joel Rodrigues (A79068), Elisa Valente (A79093)

Descrição do problema:

Através de um conjunto de dados relacionado com o consumo de álcool em estudantes do ensino secundário português pretende-se estudar as influências de vários fatores nas notas obtidas. Para além do consumo de álcool são estudadas características pessoais e socais de cada estudante. Assim, espera-se concluir quais destes têm uma influência significativa no desempenho escolar e de que forma afetam, positiva ou negativamente, a classificação final do aluno.

Descrição do conjunto de dados:

O *dataset* a ser utilizado diz respeito a 649 alunos inscritos à disciplina de português de 2 escolas secundárias. Existem 33 variáveis de interesse:

1. school – escola (binário: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
2. sex – sexo (binário: 'F' – feminino ou 'M' - masculino)
3. age – idade (numérico: de 15 a 22)
4. address – tipo de morada (binário: 'U' - urbana ou 'R' - rural)
5. famsize – tamanho da família (binário: 'LE3' - menor ou igual a 3 ou 'GT3' – maior que 3)
6. Pstatus – coabitação dos pais (binário: 'T' - vivem juntos ou 'A' – vivem separados)
7. Medu – escolaridade da mãe (numérico: 0 - nenhuma, 1 – escola primária (4º ano), 2 – 5º ao 9º ano, 3 – ensino secundário ou 4 – ensino superior)
8. Fedu – escolaridade do pai (numérico: 0 - nenhuma, 1 – escola primária (4º ano), 2 – 5º ao 9º ano, 3 – ensino secundário ou 4 – ensino superior)
9. Mjob – Emprego da mãe (nominal: 'teacher' - professora, 'health' – serviços ligados à saúde, 'services' – serviços públicos, 'at_home' – doméstica ou 'other' – outros)
10. Fjob - emprego do pai (nominal: 'teacher' - professor, 'health' – serviços ligados à saúde, 'services' – serviços públicos, 'at_home' – doméstico ou 'other' – outros)
11. reason – razão de escolha da escola (nominal: 'home' – proximidade de casa, 'reputation' – reputação, 'course' – disciplinas ou 'other' – outros)
12. guardian – encarregado de educação (nominal: 'mother', 'father' ou 'other')
13. traveltime – tempo de viagem casa-escola (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - >1 hora)
14. studytime – tempo de estudo semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, ou 4 - >10 horas)
15. failures – número de reprovações à disciplina (numérico: n se 1<=n<3, senão 4)
16. schoolsup – suporte extracurricular (binário: yes ou no)
17. famsup - apoio educacional familiar (binário: yes ou no)
18. paid – explicações (binário: yes ou no)
19. activities – atividades extracurriculares (binário: yes ou no)
20. nursery – frequentou o infantário (binário: yes ou no)
21. higher – pretende frequentar o ensino superior (binário: yes ou no)
22. internet – acesso à internet em casa (binário: yes ou no)
23. romantic – está num relacionamento (binário: yes ou no)

24. famrel – qualidade das relações familiares (numérico: de 1 – muito má a 5 - excelente)
25. freetime – tempo livre depois da escola (numérico: de 1 - muito baixo a 5 - muito alto)
26. goout – frequência de saídas com amigos (numérico: de 1 - muito baixa a 5 - muito alta)
27. Dalc – consumo de álcool à semana (numérico: de 1 - muito baixo a 5 - muito alto)
28. Walc – consumo de álcool ao fim-de-semana (numérico: de 1 - muito baixo a 5 - muito alto)
29. health – estado de saúde atual (numérico: de 1 - muito mau a 5 - muito bom)
30. absences – número de faltas à escola (numérico: de 0 a 93)

31. G1 – Nota do 1º período (numérico: de 0 a 20)
32. G2 – Nota do 2º período (numérico: de 0 a 20)
33. G3 – Nota final (numérico: de 0 a 20, variável de resposta)

Características do problema:

Trata-se de um problema supervisionado de regressão uma vez que existe uma variável de resposta (G3) e esta é quantitativa.

Temos como preocupação a correlação das notas finais de cada período uma vez que estas são calculadas tendo em conta a nota do período anterior.

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	13
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	13
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	no	no	yes	yes	yes	no	4	2	2	1	1	1	0	15	16	17
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	5	1	1	1	5	0	12	12	13
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes	no	no	yes	yes	yes	no	3	3	3	1	2	2	2	14	14	14
GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes	no	yes	yes	yes	yes	no	5	2	2	1	1	4	0	10	12	13
GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	no	yes	yes	yes	yes	no	4	3	3	1	3	5	0	12	13	12
GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	no	no	yes	yes	yes	no	5	4	3	1	2	3	0	12	12	13
GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	14	15
GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes	yes	yes	no	4	4	4	1	2	2	6	17	17	17

Figura 1 - Extrato de 16 instâncias do dataset.