Novembro 2022/23

ISCTE IUL
Instituto Universitário de Lisboa

DADOS NA CIÊNCIA, GESTÃO E SOCIEDADE

Student Performance

Docentes:
Ana Maria Almeida
José Dias

Grupo 4:

Carolina Vitória Prior dos Santos Brunheta nº 110888

Marco Delgado Esperança nº 110451

Miguel Duarte Aguilar das Neves Correia nº 110786

Vitória de Mendonça Teixeira Correia nº 110871

Índice

Introdução

O nível de educação da população portuguesa tem apresentado melhorias significativas nas últimas décadas. No entanto, as estatísticas colocam Portugal entre os países da Europa com maior número de insucessos, particularmente em disciplinas como português e matemática, e mostram uma grande percentagem de desistências dos estudos.

Testes recentes, que incidiram sobre dados reais, revelam que existem fatores relevantes, para além do resultado de avaliações passadas, que ajudam a compreender melhor o desempenho final dos alunos. Os resultados de pesquisas deste género permitem o desenvolvimento de ferramentas de previsão eficientes, que melhoram a qualidade da educação e tornam possível às escolas proporcionar aos seus alunos um apoio mais personalizado e prevenir o abandono dos estudos.

Desta forma, o presente relatório, que incide sobre o tema “Student Performance", tem como objetivo a compreensão, preparação e análise de um conjunto de dados, de maneira a obter/criar conhecimento que torne possível um melhor entendimento acerca do desempenho de estudantes do ensino secundário na disciplina de matemática, em duas escolas portuguesas, bem como dos fatores que o influenciam.

Primeiramente, e utilizando os dados provenientes do dataset que nos foi disponibilizado, reunido através da recolha de relatórios escolares e inquéritos, procuramos contextualizar o problema proposto, identificando os objetivos expectáveis e elaborando um plano para o trabalho. Consideramos que seria mais relevante explorar aspetos relacionados com a nota final dos alunos(‘G3’), excluindo as notas do primeiro e segundo períodos(‘G1’ e ‘G2’), porque anulam a utilidade da análise, o desejo de cada aluno de prosseguir os estudos para o ensino superior(‘higher’), o número de chumbos(‘failures’) e o apoio proporcionado pelas escolas aos seus alunos ('school sup'). Para além disso, definimos as ferramentas (Orange) e técnicas que iríamos usar (método CRISP-DM).

Decidimos, então, responder às seguintes perguntas: “Será possível prever a nota final dos alunos?”; “Qual a probabilidade de um aluno

querer seguir um curso superior?"; "Quais são os fatores dos quais depende o número de reprovações?".

Data Understanding

Após estarmos por dentro do domínio do problema, procuramos compreender os dados, o significado de cada variável, bem como, para cada pergunta, escolher as variáveis que faziam mais sentido para aquilo que se pretendia prever.

Começámos por compreender os dados, o significado de cada variável, bem como quais é que iríamos escolher de entre as 33 disponíveis. Seguidamente, utilizamos a documentação fornecida pelos docentes com a descrição detalhada de cada uma das variáveis e procedemos à análise do dataset.

school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex - student's sex (binary: 'F' - female or 'M' - male)
age - student's age (numeric: from 15 to 22)
address - student's home address type (binary: 'U' - urban or 'R' - rural)
Pstatus - parent's cohabitation status (binary: 'T' - together or 'A' - apart)
Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g.administrative or police), 'at home' or 'other')
Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g.administrative or police), 'at home' or 'other')

guardian - student's guardian (nominal: 'mother', 'father' or 'other')
famsize - family size (binary: ≤3 or >3)
famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
traveltime - home to school travel time (numeric: 1 - <15 min; 2 - 15 to 30 min; 3 - 30 min to 1 hour or 4 - >1 hour)
studytime - weekly study time (numeric: 1 - <2 hours; 2 - 2 to 5 hours; 3 - 5 to 10 hours or 4 - >10 hours)
failures - number of past class failures (numeric: n if 1≤ n<3, else 4)
schoolsup - extra educational school support (binary: yes or no)
famsup - family educational support (binary: yes or no)
activities - extra-curricular activities (binary: yes or no)
paid - extra paid classes within the course subject (binary: yes or no)
internet - internet access at home (binary: yes or no)
nursery - attended nursery school (binary: yes or no)
higher - wants to take higher education (binary: yes or no)
romantic - with a romantic relationship (binary: yes or no)
freetime - free time after school (numeric: from 1 - very low to 5 - very high)
goout - going out with friends (numeric: from 1 - very low to 5 - very high)
Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
health - current health status (numeric: from 1 - very bad to 5 - very good)
absences - number of school absences (numeric: from 1 to 93)
G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20)

Verificamos que havia dados categóricos e numéricos, sendo que convertemos todos os dados numéricos para categóricos por sugestão dos

docentes e para facilitar a classificação e distribuição das previsões através da matriz de confusão. Para cada pergunta procuramos estabelecer um critério de seleção relacionado de acordo com o nosso target, que tínhamos definido como o objetivo da previsão. Assim, exploramos as 4 perguntas, relacionadas com o tema do nosso projeto, que nos permitiram tirar conclusões interessantes sobre diferentes fatores que influenciam o sucesso académico.

Para a primeira pergunta ("Quais os fatores que mais influenciam a nota final dos alunos?" – target 'G3'), escolhemos as variáveis com base em 4 subcategorias:

- Educação / formação dos pais ('Medu'/'Fedu' - formação da mãe/pai; 'MJob'/'FJob' - profissão da mãe/pai);
- Tempo disponível ('freetime' - tempo livre, 'studytime' - tempo de estudo e 'traveltime' - tempo de viagem, 'goout'- saídas);
- Dados / recursos pessoais ('age' - idade, 'adress' - morada, 'guardian' - com quem vive, 'internet' – se tem acesso à internet);
- Desempenho escolar / apoio escolar ('failures'- número de chumbos, 'absenses' - faltas, 'schoolsup' - apoio das escolas, 'school' - escola).

Para a segunda pergunta ("Qual a probabilidade de um aluno seguir um curso superior?" - target 'higher'), escolhemos as variáveis 'failures', 'absences', 'reason', 'G3', 'activities', 'Walc', 'Dalc', 'studytime', 'goout', 'Medu' e 'Fedu'.

Para a terceira pergunta ("Quais são os fatores dos quais depende o número de reprovações?" - target 'failures'), foram escolhidas as variáveis 'Walc', 'Dalc', 'G3', 'traveltime', 'Fedu', 'Medu', 'Mjob', 'Fjob', 'internet', 'schoolsup', 'famsup', 'absences' e 'higher'.

Data Preparation

Finalizado o processo de compreensão e seleção das variáveis, começamos a fazer a preparação e tratamento dos dados no Orange.

Primeiramente, importamos o CSV disponibilizado pelos docentes na pasta do projeto através do widget “File”.

Em seguida, utilizamos “Feature Statistics” e verificamos que não existiam “missing values” no nosso dataset, o que acelerou o nosso trabalho e facilitou o processo. Para além disso, convertemos todos os dados que estavam numéricos para categóricos. Esta mudança é justificável por facilitar a análise, tendo em conta a distribuição de cada um dos dados, de forma a ser possível obter um modelo de previsão de qualidade.

Recorrendo ao widget “Select Columns” (um por cada pergunta) selecionamos as variáveis anteriormente mencionadas, sendo todas elas categóricas. Através de “Edit Domain”, subdividimos as seguintes variáveis:

- ‘G3’ em 2 grupos: fail (notas menores que 10) e pass (notas a partir de 10);
- ‘age’ em 2 grupos: menor (menores que 18 anos) e maior (maiores de idade);
- ‘studytime’ em 2 grupos: studies a lot e studies less;
- ‘absences’ em 3 grupos: no absences, 1-9 absences e 10+ absences
- ‘Medu’/’Fedu’ em 3 grupos: low, médium e high;
- ‘Walc’/’Dalc’ em 3 grupos: low, médium e high.

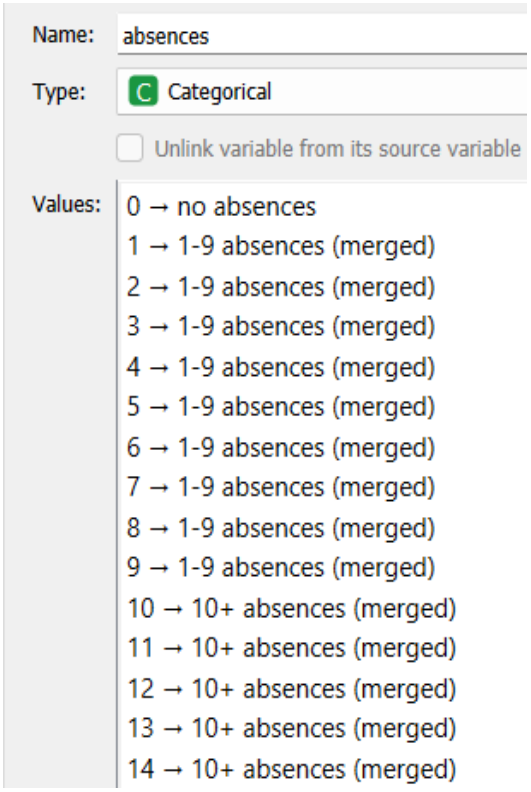

Figura 1 - Divisão do dado categórico absenses.

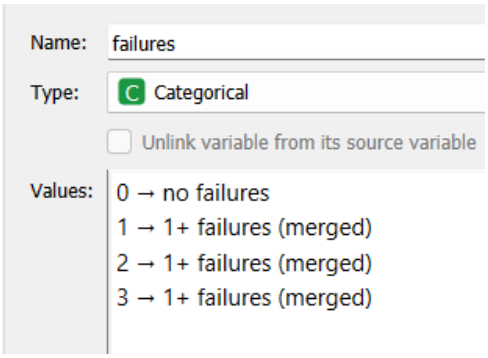

Figura 2 - Divisão do dado categórico failures.

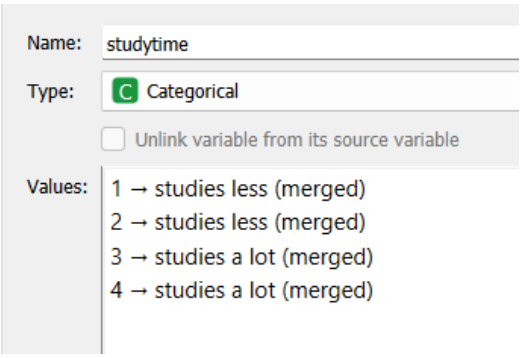

Figura 3 – Divisão do dado categórico studytime.

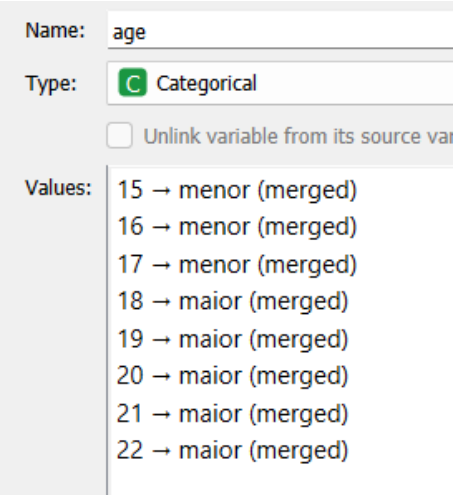

Figura 4 - Divisão do dado categórico age.

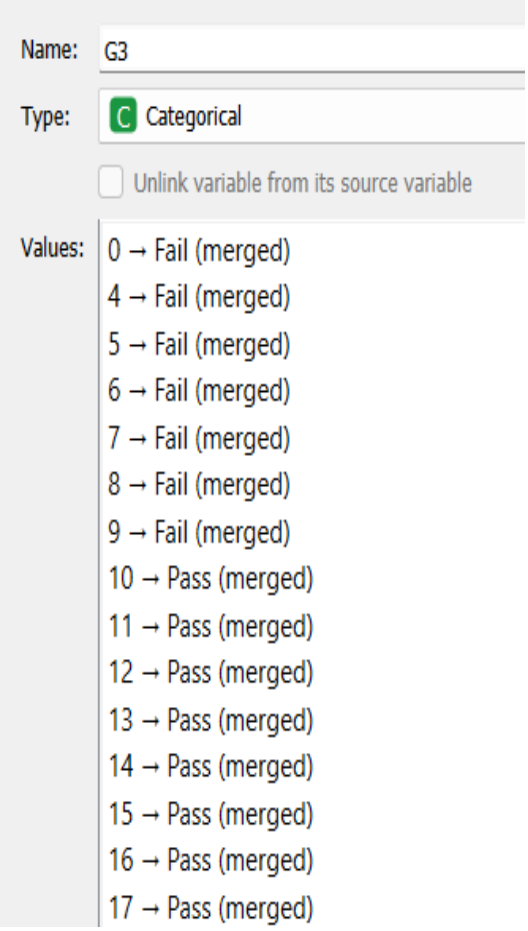

Figura 5 - Divisão do dado categórico G3.

Realizados estes passos tínhamos uma amostra de dados pronta para ser trabalhada e realizar os modelos preditivos.

Prosseguimos então para a próxima fase: data analysis.

Data Analysis

Para cada pergunta, escolhemos as variáveis recorrendo aos widgets “Rank” (melhor cotados) e “Feature Statistics” (menos dispersão).

De forma a facilitar a visualização dos diferentes gráficos utilizamos o widget “Colour” de forma a cada resposta ter uma cor diferente e intuitiva para o seu contexto e significado. A todas as variáveis categóricas binárias com “yes”, atribuímos a cor verde, e para “no” a cor vermelha. No ‘failures’, colocamos “no failures” a verde e “1+ failures” a vermelho. Nas variáveis com mais de duas respostas possíveis deixamos as cores default sugeridas pelo Orange. Na variável ‘absences’ colocamos a verde a categoria “no absences”, a amarelo “1-9 absences” e a vermelho “10+ absences”.

Tentamos perceber que fatores poderiam influenciar mais os estudantes que pretendem ou não seguir estudos no ensino superior (“**Qual a probabilidade de um aluno seguir um curso superior?**” – target ‘higher’).

Tendo em conta as relações estabelecidas anteriormente entre as variáveis e a nossa variável principal, começámos a analisar tendo em conta o quanto mais relacionada estava com o target “higher”.

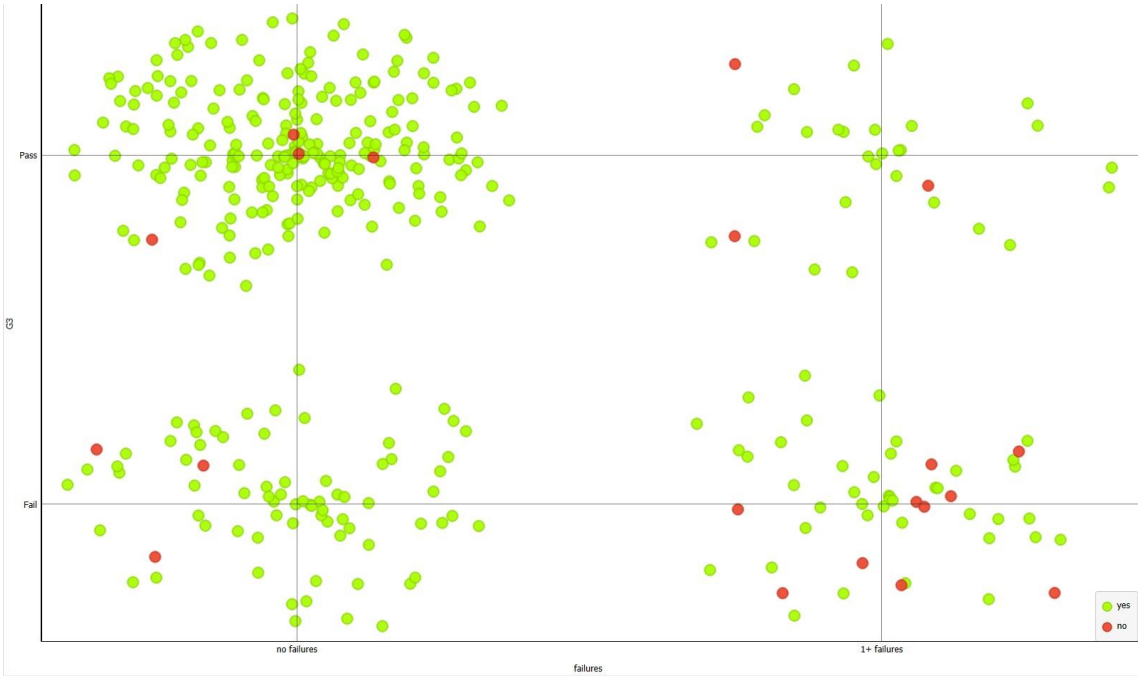

Figura 6 – Nota final ('G3') em função dos chumbos (‘failures’) com a distinção para a variável 'higher'.

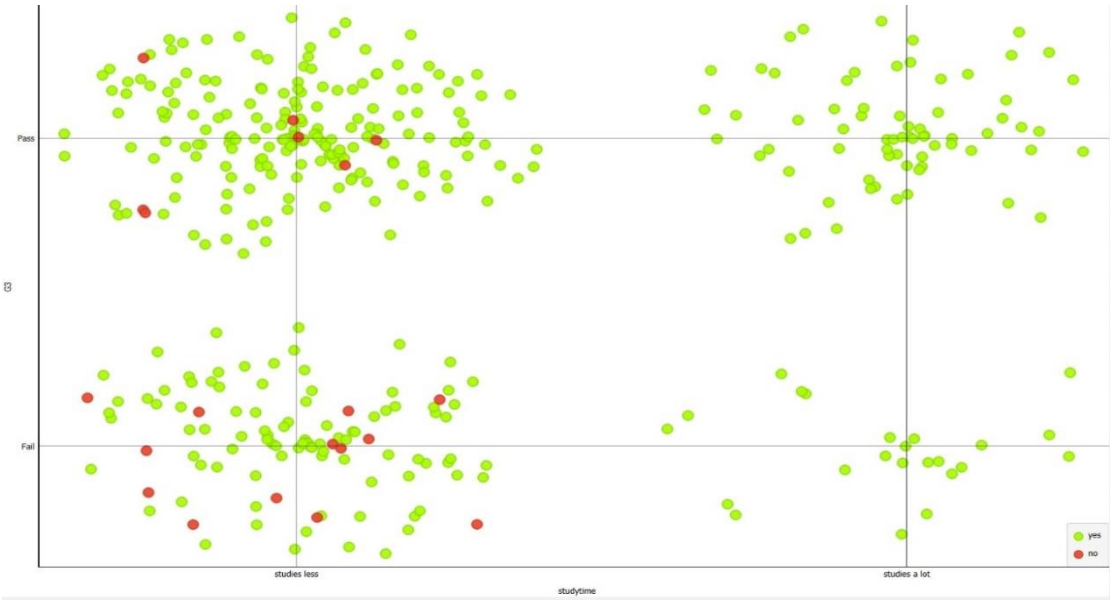

Figura 7 – Nota final ('G3') em função do tempo de estudo ('studytime') com a distinção para a variável 'higher'.

Alguns aspetos que consideramos importante analisar, através do widget "Scatter Plot", foram a relação entre os chumbos ('failures'), o tempo de estudo ('studytime') e o desejo dos alunos de ir para a faculdade ('higher'). Na nossa perspetiva, o facto de um aluno não ter chumbado nenhuma vez e estudar bastante seriam bons indicadores de que pretende prosseguir os estudos para o ensino superior. Efetivamente, a maior parte dos estudantes motivados a prosseguir estudos para o ensino superior não têm qualquer registo de chumbos e passaram de ano, o que foi de encontro às nossas expectativas. Contudo, e contrariamente ao que pensávamos, há um maior número de estudantes que querem seguir estudos no ensino superior que se encontram na categoria de estudar menos ('study less') do que aqueles na categoria de estudar bastante ('studies a lot').

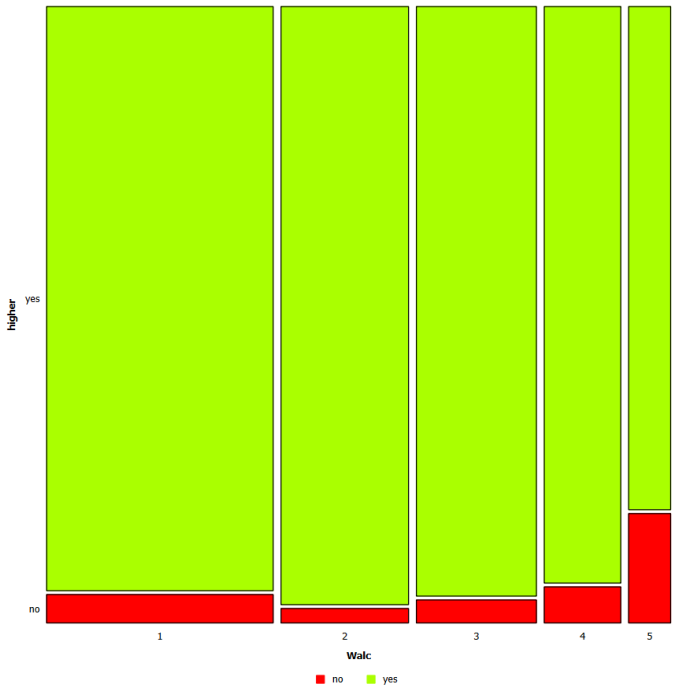

Figura 8 – Distinção do Consumo alcoólico ao fim de semana para a variável 'higher'.

Analisamos, também, a relação entre o consumo de álcool (‘Walc’) e o desejo de ingressar no ensino superior (‘higher’), através de um “Mosaic display”.

Pudemos verificar que, quanto maior o consumo alcoólico ao fim de semana, menor a proporção de estudantes que pretendem seguir estudos no ensino superior. Isto pode ser justificado pelo facto de, talvez, este alto consumo de álcool, de forma regular e irresponsável, poder constituir uma barreira aos estudos, revelando desinteresse pelos mesmos.

Decidimos não colocar o gráfico em função do consumo diário de álcool (‘Dalc’), uma vez que não agregava valor à análise.

Seguidamente, tentamos perceber alguns dos fatores dos quais depende o número de reprovações (**“Quais são os fatores dos quais depende o número de reprovações?”** – target ‘failures’).

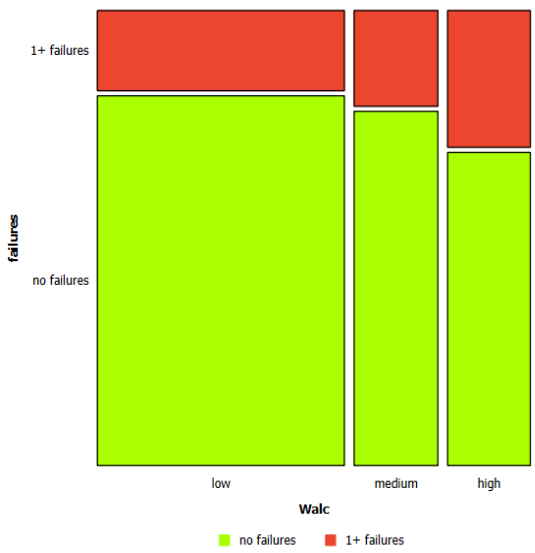

Figura 9 – Distinção do Consumo alcoólico ao fim de semana para a variável ‘failures’.

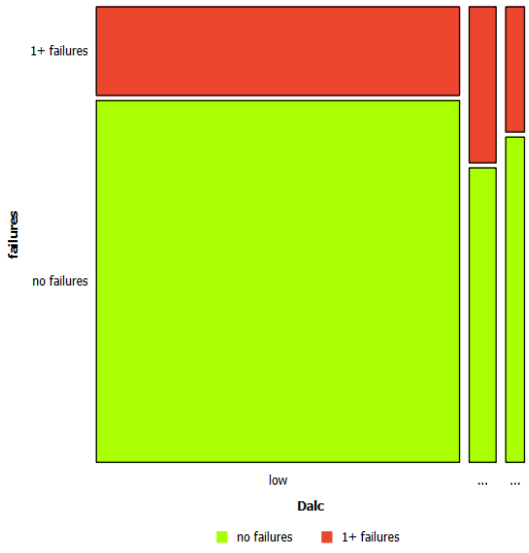

Figura 10 – Distinção do Consumo alcoólico diário para a variável ‘failures’.

Através dos “Mosaic display” apresentados, podemos concluir que, quanto maior o consumo alcoólico, maior é a quantidade de chumbos, tal como esperaríamos, uma vez que o álcool em excesso, especialmente se consumido em idades precoces, provoca problemas de concentração, fadiga, entre outros, o que pode comprometer o desempenho dos alunos e, consequentemente, levar a que, quando consumido de uma forma regular, provoque o chumbo de um maior número de alunos.

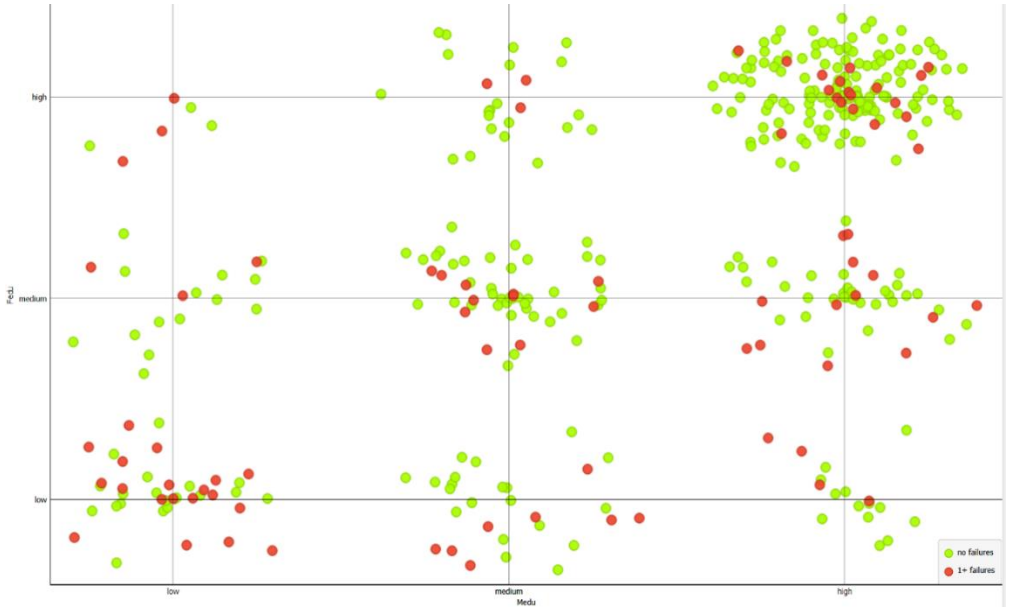

Figura 11 – Formação do pai ('Fedu') e formação da mãe ('Medu') com a distinção para a variável 'failures'.

Pelo "Scatter plot" apresentado, podemos concluir que quanto maior o nível de formação académica do pai e da mãe ('Medu'/'Fedu'), maior a proporção de alunos que nunca chumbaram, o que era de esperar, uma vez que, geralmente, quanto mais formação têm os pais, mais formação também vão querer que os seus filhos tenham e mais podem investir no seu sucesso e acompanhamento académico, prevenindo que chumbem.

Data Modelling

Escolhemos as variáveis com base na exclusão daquelas que tinham menos relief no "Rank", em articulação com as variáveis que estavam mais diretamente relacionadas com o target de cada pergunta. Por exemplo, escolhemos somente a variável G3 (em vez de incluir também as restantes notas: G1 e G2), pois são bastante semelhantes ao G3 e não acrescentavam informação útil ao modelo em termos de resultados de previsão.

Nesta fase, construímos os modelos de previsão com base nas variáveis que escolhemos para cada pergunta. Como modelos de aprendizagem selecionamos “Random Forest”, “Neural Network”, “Logistic Regression” e “SVM”, escolhendo para cada pergunta o modelo de aprendizagem que apresentou melhores resultados, ou seja, aquele que teve melhor *classification accuracy.*

Primeiramente, interligamos as variáveis escolhidas através de “Select Columns” e o modelo de aprendizagem ao widget “Test and Score” para verificar os resultados do nosso modelo.

Seguidamente, definimos como target a variável alvo de cada pergunta. Assim, o modelo tenta prever quais os fatores que mais a influenciam.

Para obter o ROC utilizamos o Widget “ROC Analysis” e para a *classification accuracy* e *precision* usamos o widget “Test and Score”.

Para cada modelo apresentamos aquele que teve melhor *classification accuracy* e, por uma questão de simplificação de visualização, apresentamos apenas o gráfico do ROC para esse modelo.

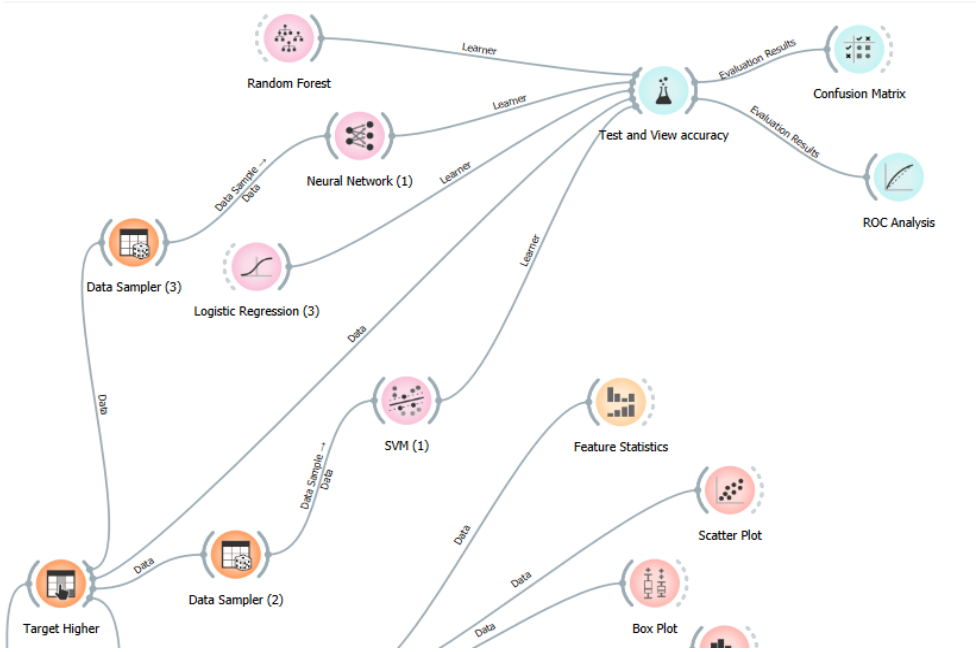

Figura 12 – Fase de data modelling no Orange.

Em relação à primeira pergunta (**“Quais os fatores que mais influenciam a nota final dos alunos?”** - target ‘G3’), o melhor modelo que

obtivemos foi o “Random Forest”. Conseguimos uma *area under ROC curve* de 0.690, uma *classification accuracy*, de, aproximadamente, 72% e uma *precision* de, aproximadamente, 70%. Cada um dos possíveis resultados para o ‘G3’ tinha uma curva única e, da sua análise, concluímos que a melhor curva era a que estava associada ao ‘Pass’ (figura 14).

		Predicted		
		Fail	**Pass**	**∑**
Actual	**Fail**	62.2 %	26.2 %	**130**
	Pass	37.8 %	73.8 %	**265**
	∑	**74**	**321**	**395**

Figura 13 – Confusion Matrix associada ao modelo cuja target variable é ‘G3’.

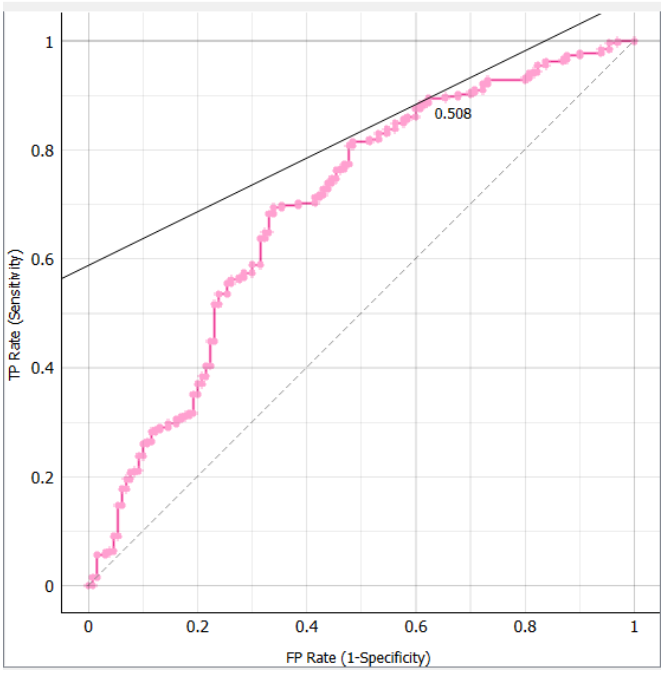

Figura 14 - ROC analysis relativa à resposta ‘Pass’ da variável ‘G3’.

Relativamente à segunda pergunta (**“Qual a probabilidade de um aluno seguir um curso superior?”** – target ‘higher’), o modelo com melhores conclusões foi o “Random Forest”, com uma *area under ROC* curve de 0.759, uma *classification accuracy* de, aproximadamente, 95% e uma *precision* de 90%.

Cada um dos resultados binários para o ‘higher’ tinha uma curva única e, da sua análise, concluímos que a melhor curva era a que estava associada ao ‘yes’’ (figura 16).

		Predicted		
		no	**yes**	**∑**
Actual	**no**	NA	5.1 %	**20**
	yes	NA	94.9 %	**375**
	∑	**0**	**395**	**395**

Figura 15 – Confusion Matrix associada ao modelo cuja target variable é ‘higher’.

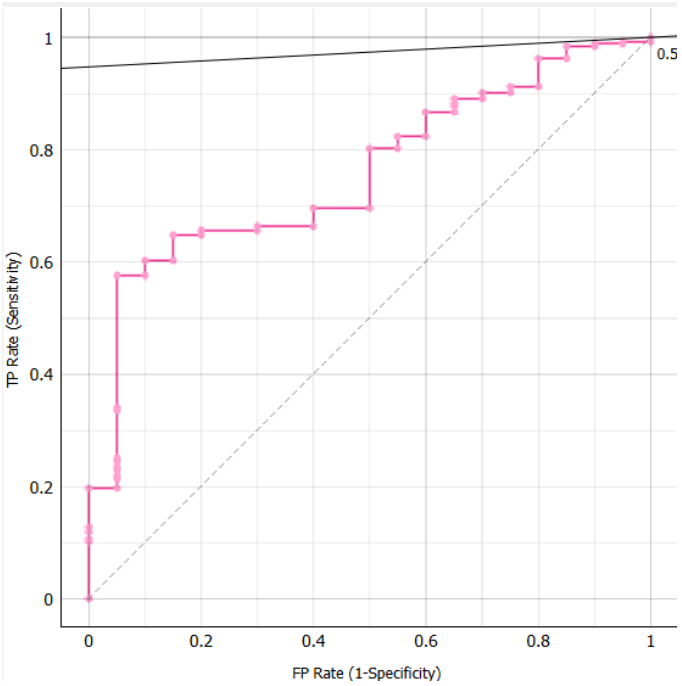

Figura 16 - ROC analysis relativa à resposta ‘Yes’ da variável ‘higher’.

Na última pergunta (**“Quais são os fatores que mais influenciam o número de reprovações?”** – target ‘failures’), “Logistic Regression” obteve os melhores resultados, com uma *area under ROC curve* de 0.755, uma *classification accuracy* de, aproximadamente, 81% e uma *precision* de, aproximadamente, 79%.

Cada um dos possíveis resultados para o ‘failures’ tinha uma curva única e, da sua análise, concluímos que a melhor curva era a que estava associada ao ‘no failures’ (figura 18).

		Predicted		
		no failures	**1+ failures**	**∑**
Actual	**no failures**	84.0 %	41.3 %	**312**
	1+ failures	16.0 %	58.7 %	**83**
	∑	**349**	**46**	**395**

Figura 17 – Confusion Matrix associada ao modelo cuja target variable é ‘failures’.

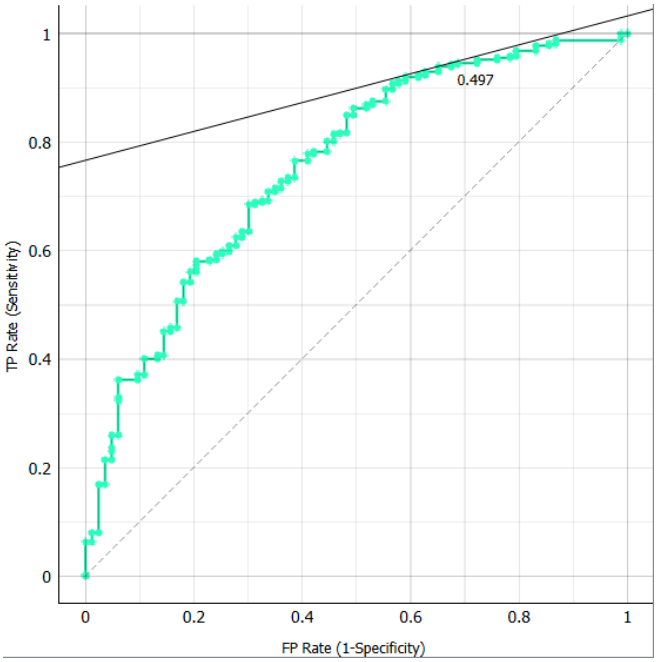

Figura 18 - ROC analysis relativa à resposta ‘no failures’ da variável ‘failures’.

Conclusão

Em suma, chegamos à conclusão que, nestas duas escolas portuguesas, os alunos do ensino secundário motivados a prosseguir estudos para o ensino superior, na sua maioria, são alunos que não chumbam, tal como esperávamos, e que não estudam muitas horas por semana, o que nos surpreendeu. Além disso, observamos que um consumo de álcool e uma formação académica dos pais elevados influenciam os chumbos, tal como previsto. Conseguimos obter, para duas das nossas perguntas, bons modelos de previsão. Quanto à previsão da nota final de matemática e dos fatores que a influenciam, não conseguimos chegar a conclusões muito relevantes e significativas.

Todas as conclusões foram suportadas por gráficos de visualização e pelos resultados obtidos pelos modelos de previsão.

Com vista à melhoria do trabalho, poderíamos ter conseguido encontrar um maior número de relações significativas entre variáveis, particularmente no caso em que o target era a variável ‘G3’.

Comprovamos, assim, que existem fatores, para além dos resultados académicos, que influenciam o desempenho escolar dos alunos.

Webgrafia

https://orangedatamining.com/widget-catalog/
http://www3.dsi.uminho.pt/pcortez/student.pdf