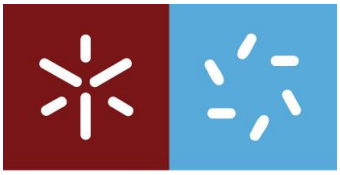


Universidade do Minho
Escola de Ciências

Aplicação de Métodos de Aprendizagem Estatística à Base de Dados *"Student Performance"*

Mestrado em Estatística para a Ciência de Dados
Teoria da Aprendizagem Estatística

Ano Letivo 2024/2025
Rui Miguel Pereira Alves - PG55577

Abril de 2025

Resumo Executivo

Este trabalho propõe-se a prever o sucesso académico dos estudantes na disciplina de Matemática, utilizando exclusivamente variáveis demográficas, sociais e escolares, sem recorrer diretamente às classificações anteriores. O problema abordado enquadra-se no domínio da aprendizagem estatística supervisionada, sendo tratado como uma tarefa de classificação binária (aprovação ou reprovação).

Foram utilizados três métodos principais ao longo do estudo: Regressão Lasso, Árvores de Decisão e *Random Forest*. A seleção de preditores e a construção dos modelos procuraram equilibrar a capacidade preditiva com a interpretabilidade, recorrendo a técnicas de reamostragem como validação cruzada (*10-fold cross-validation*) e divisão treino/teste (70%/30%).

As análises realizadas indicaram que os fatores mais relevantes para a previsão do desempenho académico dos alunos foram o número de reprovações anteriores (*failures*), o número de faltas (*absences*), a existência de apoio educativo extra (*schoolsup*), o desejo de prosseguir para o ensino superior (*higher*) e a frequência de saídas com amigos (*goout*). Estes preditores demonstraram maior consistência e impacto nos diversos modelos.

Em termos de desempenho, o método de *Random Forest* obteve a melhor taxa de acerto, com uma *accuracy* média de cerca de 75,4%. As Árvores de Decisão também apresentaram boas performances, com uma *accuracy* que atingiu 77.2% no treino e cerca de 75% no teste, mantendo ainda uma interpretabilidade razoável. A Regressão Lasso, focada em modelos mais parcimoniosos, registou uma *accuracy* média em torno dos 70%.

Conclui-se que é possível construir modelos preditivos eficazes e robustos para a previsão da aprovação escolar utilizando um número reduzido de variáveis, sensivelmente entre oito a dez preditores, mantendo um equilíbrio entre capacidade preditiva e compreensão dos fatores críticos para o sucesso académico.

Conteúdo

1 Introdução

O presente trabalho foi desenvolvido no âmbito da unidade curricular de Teoria da Aprendizagem Estatística, com o objetivo de aplicar os conceitos estudados durante o semestre a um problema real de previsão.

A capacidade de prever o sucesso académico dos estudantes, tendo por base fatores demográficos, sociais e escolares, assume particular importância na identificação precoce de alunos em risco de insucesso e na definição de estratégias de intervenção. Neste projeto, a previsão do desempenho foi realizada sem utilizar diretamente as classificações dos estudantes ao longo do ano letivo, procurando identificar quais os fatores que realmente tiveram preponderância no desempenho académico e que foram essenciais para a construção dos modelos de aprendizagem.

Este problema apresenta um nível acrescido de desafio, uma vez que se baseia exclusivamente em variáveis indiretas, sem recorrer a informações explícitas de desempenho anterior. Além disso, o elevado número de preditores disponíveis - maioritariamente variáveis qualitativas, binárias ou categóricas - aumenta a complexidade da tarefa, exigindo métodos rigorosos de seleção e modelação. A capacidade de desenvolver um modelo eficaz nestas condições é de particular interesse prático, dado que permite a identificação de alunos em risco de reprovação a partir de dados habitualmente recolhidos pelas instituições de ensino.

Ao longo do trabalho, foram exploradas diversas técnicas de aprendizagem estatística abordadas na unidade curricular, incluindo métodos de reamostragem, seleção de preditores e algoritmos de classificação. A construção dos modelos teve em consideração tanto a capacidade preditiva como a interpretabilidade das soluções, procurando alcançar um equilíbrio entre a precisão das previsões e a compreensão dos fatores que mais contribuem para o sucesso escolar.

O objetivo principal consistiu na previsão da aprovação ou reprovação dos estudantes, com base num conjunto alargado de variáveis que se acredita que podem influenciar o desempenho académico.

2 Descrição dos Dados

O presente estudo utiliza a base de dados *Student Performance Data Set*, disponível no repositório UCI Machine Learning (`http://archive.ics.uci.edu/dataset/320/student+performance`).

Os dados referem-se a alunos de duas escolas públicas da região do Alentejo, em Portugal, recolhidos durante o ano letivo de 2005/2006. A recolha foi realizada a partir de duas fontes: relatórios escolares em formato papel (contendo informações como notas e faltas) e questionários preenchidos em sala de aula. Os questionários incluíam questões demográficas, sociais e relacionadas com o ambiente escolar.

Foram inicialmente recolhidos dados de 788 estudantes, mas após um processo de limpeza (nomeadamente pela ausência de identificação em algumas respostas), resultaram dois conjuntos de dados finais: um para a disciplina de Matemática (395 observações) e outro para a disciplina de Português (649 observações). Neste trabalho apenas se considera o conjunto de dados relativo à disciplina de Matemática.

O conjunto de dados contém diversas variáveis:

- **Variáveis demográficas**: como o género e idade.
- **Variáveis familiares**: como o nível de educação dos pais, profissão dos pais e situação familiar.
- **Variáveis escolares**: como tempo de deslocação para a escola, tempo de estudo semanal, participação em atividades extracurriculares e acesso à internet.

- **Variáveis sociais e comportamentais**: como consumo de álcool nos dias úteis e fim de semana, relações familiares e número de reprovações anteriores.
- **Variáveis de desempenho académico**: como as notas do primeiro, segundo e terceiro períodos (G1, G2 e G3), e número de faltas.

As variáveis são de diferentes tipos: variáveis quantitativas (ex.: idade, número de faltas, notas) e variáveis qualitativas (ex.: género, profissão dos pais, situação familiar).

O objetivo final do estudo é utilizar este conjunto de dados para prever se um aluno foi aprovado ou reprovado a Matemática, com base nas suas características.

3 Análise Exploratória

Antes da modelação preditiva, foi realizada uma breve análise exploratória dos dados, com o objetivo de entender melhor o comportamento das variáveis disponíveis e a relação entre preditores e a variável resposta.

3.1 Análise dos Preditores

Para as variáveis quantitativas, foi calculada a matriz de correlação de *Spearman*, uma vez que esta mede associações de monotonia e a maior parte das variáveis quantitativas são categóricas. Já para as variáveis qualitativas, foram construídos boxplots do *Pass/Fail* (1/0) em função de algumas variáveis categóricas selecionadas.

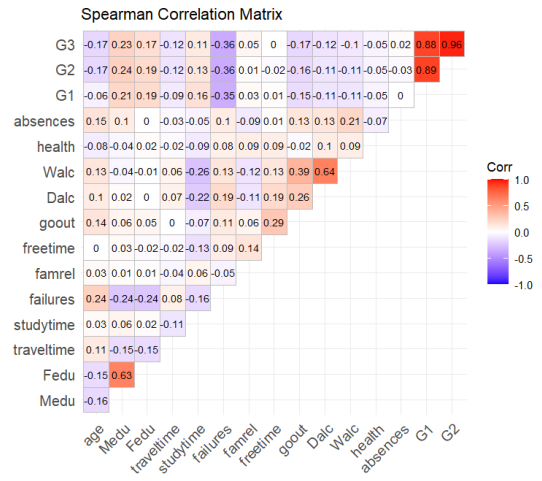


Figura 3.1: Matriz de correlação de Spearman entre preditores quantitativos.

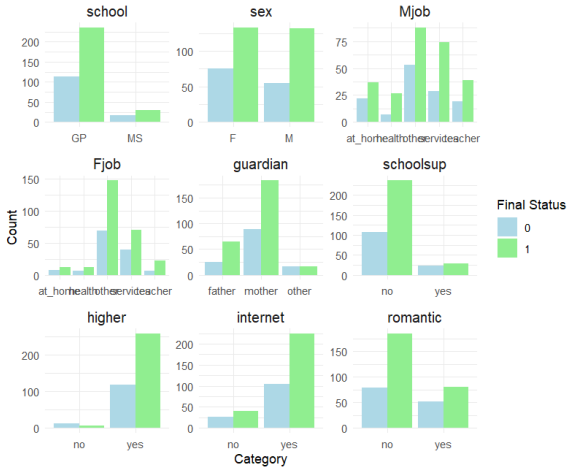


Figura 3.2: Histogramas da aprovação/reprovação (*PF*) em função de variáveis categóricas.

A matriz de correlação de Spearman mostra que as notas dos períodos anteriores (G1 e G2) apresentam forte correlação com a nota final (G3), como esperado. Além destas, observa-se que o número de reprovações anteriores (*failures*) e o número de faltas (*absences*) também estão moderadamente correlacionados com o desempenho final. Também se observam correlações positivas significativas para a escolaridade dos pais (*Medu e Fedu*) e negativas para idade (*age*), o tempo de deslocação (*traveltime*) e a frequência de saídas com amigos (*goout*), em relação à variável resposta.

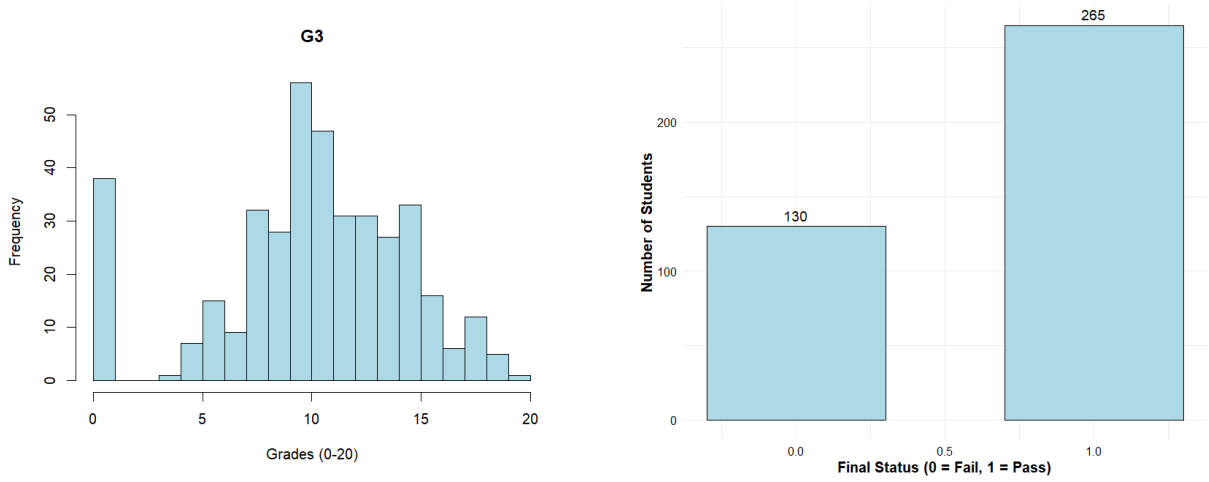
Relativamente às variáveis categóricas, nota-se que fatores como o tipo de apoio escolar (*schoolsup*), o desejo de prosseguir para o ensino superior (*higher*), e a situação romântica

(*romantic*) parecem influenciar a taxa de aprovação. A distribuição das respostas sugere tendências diferenciadas no sucesso académico, justificando a inclusão destas variáveis nas análises posteriores.

3.2 Análise da Variável Resposta

Para compreender melhor a construção da variável resposta *PF* (Pass/Fail), analisou-se inicialmente a distribuição da nota final (*G3*) dos alunos.

A Figura 3.3 apresenta, lado a lado, o histograma das notas finais (*G3*) e o histograma da variável binária *PF*. Observa-se que a variável *PF* resulta de uma discretização de *G3*, considerando como aprovação as notas iguais ou superiores a 10 valores.



(a) Histograma da nota final (G3)

(b) Histograma da variável binária (PF)

Figura 3.3: Distribuição da variável resposta: nota final (*G3*) e aprovação/reprovação (*PF*).

A Figura 3.3 apresenta a distribuição das notas finais (*G3*) dos alunos na disciplina de Matemática. Observa-se uma distribuição relativamente dispersa entre 0 e 20 valores, com uma concentração acentuada em torno dos 10 valores. Nota-se ainda uma quantidade considerável de alunos com nota 0, o que poderá indicar reprovações automáticas ou desistências.

Já a Figura 3.3 mostra a distribuição da variável binária *PF*, correspondente ao estado final de aprovação (1) ou reprovação (0). Verifica-se que cerca de dois terços dos estudantes foram aprovados, enquanto cerca de um terço reprovou. Esta distribuição moderadamente desbalanceada é relevante para a interpretação e avaliação dos modelos de classificação, especialmente para métricas como a sensibilidade/especificidade.

4 Seleção de Preditores

Como já enunciado nas secções acima, esta base de dados contém um vasto número de variáveis que se acredita que podem influenciar o desempenho académico e, consequentemente, a reprovação ou aprovação na disciplina de matemática. Dado isto, e na tentativa de responder à questão de quantos preditores serão necessários para construir um modelo de previsão adequado, recorremos a diferentes métodos de seleção de preditores para tentar perceber quais seriam os fatores críticos e quais teriam menor importância.

É importante referir que todos os métodos que vão ser apresentados neste projeto começaram com a utilização de todas as variáveis explicativas. Esta secção serve apenas para dar uma visão global sobre os preditores e a sua relevância.

4.1 Seleção com Regsubsets

Inicialmente foi utilizada a função *regsubsets* da biblioteca *leaps* para a construção dos melhores modelos utilizando de 1 até 20 preditores. Foram construídos modelos lineares da variável *G3* (notas dos alunos na disciplina de matemática no 3º Período) em função das variáveis explicativas existentes na base de dados (com a exceção de *G1* e *G2* uma vez que estão fortemente correlacionadas com a variável resposta e podem suprimir a influência dos outros preditores).

Através da análise dos modelos gerados, foi possível observar que algumas variáveis foram consistentemente selecionadas nos primeiros modelos, indicando a sua elevada relevância para explicar a variável resposta *G3*. Entre estas destacam-se, por exemplo, o número de reprovações anteriores (*failures*), o tempo semanal de estudo (*studytime*) e o estatuto de apoio educacional extra (*schoolsup*).

Por outro lado, algumas variáveis nunca foram incluídas nos melhores modelos com até 20 preditores, sugerindo uma fraca contribuição preditiva no contexto considerado. Entre estas encontram-se variáveis como a escola frequentada (*school*), o estado civil dos pais (*Pstatus*), a educação do pai (*Fedu*), o encarregado de educação do aluno (*guardian*), o tempo de deslocação até à escola (*traveltime*), se tem explicações privadas pagas (*paid*), a frequência de atividades extracurriculares (*activities*), a frequência da creche (*nursery*), o acesso à Internet em casa (*internet*), a qualidade das relações familiares (*famrel*), e os consumos de álcool diário e de fim de semana (*Dalc* e *Walc*).

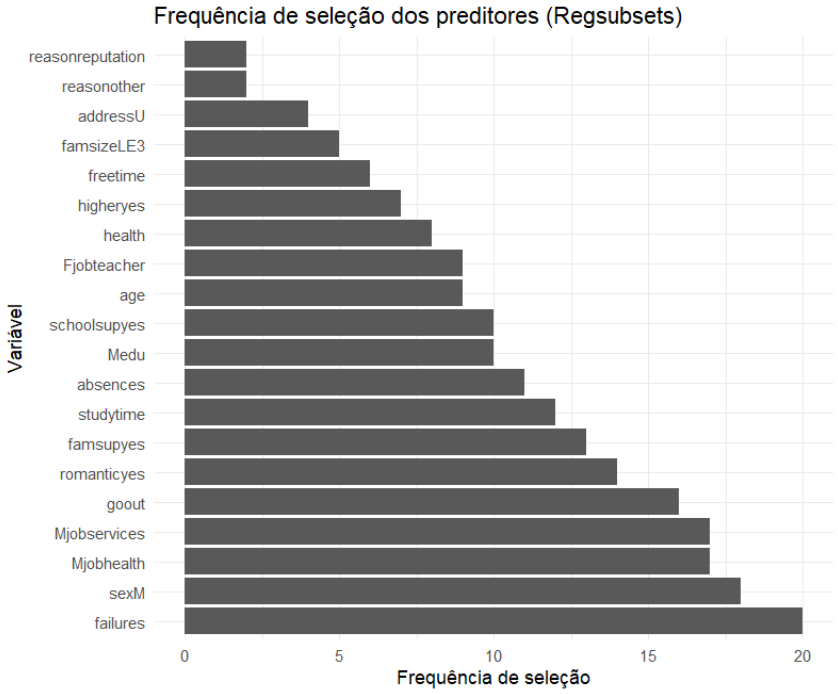


Figura 4.1: Frequência de seleção dos preditores através do método *regsubsets*.

Esta análise inicial permitiu identificar tendências sobre a relevância das variáveis, informação que será considerada nas metodologias de previsão a serem aplicadas nas secções seguintes.

4.2 Regressão Lasso: Ajuste Inicial

Para além da seleção exaustiva de subconjuntos de variáveis, foi também utilizada a regressão Lasso (*Least Absolute Shrinkage and Selection Operator*) como método alternativo de seleção

de preditores. Esta abordagem é conhecida pela sua capacidade de realizar simultaneamente regularização e seleção de variáveis, através da penalização da soma dos valores absolutos dos coeficientes no modelo.

Para aplicar esta metodologia, foi utilizada a função `glmnet` do R. A variável resposta considerada foi a variável binária *PF* (Pass/Fail - indicando aprovação ou reprovação), e a matriz de preditores foi construída a partir da base de dados, considerando todas as variáveis disponíveis, excetuando as notas *G1* e *G2*.

A regressão Lasso foi ajustada para uma sequência de valores de penalização λ, variando de 10^{10} a 10^{-2}. A Figura 4.2 apresenta a evolução dos coeficientes à medida que o parâmetro de regularização λ varia. Observa-se que, para valores maiores de λ, mais coeficientes são encolhidos para zero, evidenciando o efeito de seleção de variáveis inerente ao método.

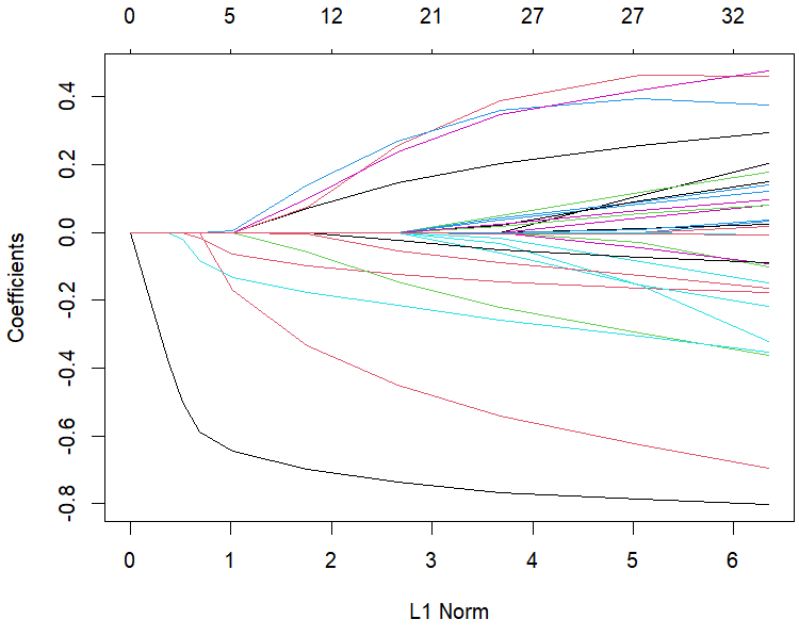


Figura 4.2: Evolução dos coeficientes na regressão Lasso para diferentes valores de λ.

Este procedimento fornece uma primeira visão sobre quais preditores são mais relevantes para a previsão do sucesso académico. Em secções posteriores, serão aprofundadas as escolhas de λ com base em critérios de validação cruzada, assim como a comparação de desempenho preditivo dos modelos selecionados.

4.3 Cross-Validation e Escolha do Melhor Modelo

Após a aplicação do método de regressão Lasso, foi realizada validação cruzada para selecionar o valor de λ que proporciona o melhor relação entre complexidade e interpretabilidade do modelo.

Para tal, utilizou-se a função `cv.glmnet`, com uma divisão dos dados em 10 folds. O gráfico da validação cruzada (Figura 4.3) ilustra a evolução do erro de validação para diferentes valores de λ.

O λ que minimizou o erro de validação cruzada foi:

$$\lambda_{\min} = 0.0231 \quad (\log(\lambda_{\min}) \approx -3.77),$$

enquanto o λ mais simples, dentro de um erro padrão do mínimo, foi:

$$\lambda_{1\text{se}} = 0.0936 \quad (\log(\lambda_{1\text{se}}) \approx -2.37).$$

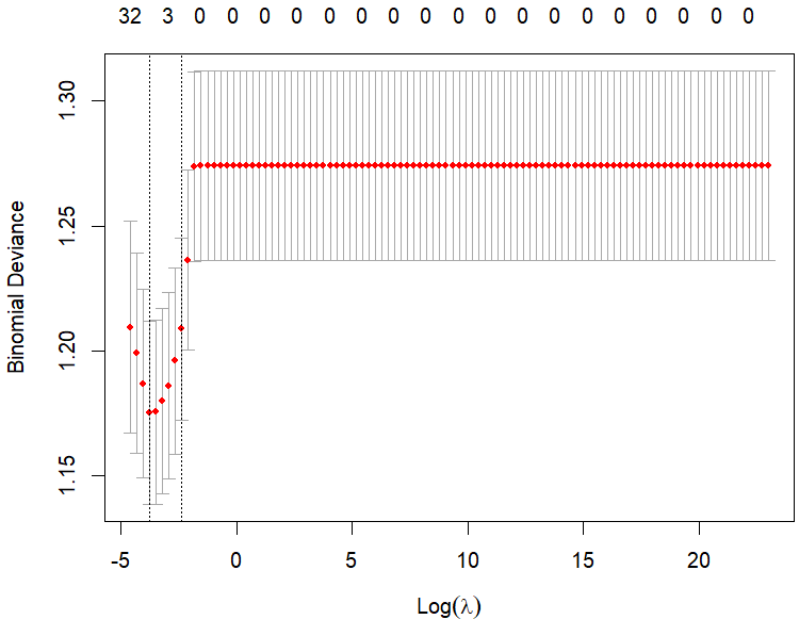


Figura 4.3: Erro de validação cruzada em função de $\log(\lambda)$ para a regressão Lasso.

Optou-se por considerar $\lambda_{\min}$, permitindo um modelo mais preditivo mesmo que ligeiramente mais complexo.

Com base no valor ótimo de λ obtido por validação cruzada, estão de seguida apresentados os coeficientes estimados pelo modelo Lasso. As variáveis anuladas (i.e., com coeficiente exatamente igual a zero) encontram-se assinaladas com um ponto ".".

Variável	Coeficiente
school	.
sexM	0.147
age	-0.124
address	.
famsize	.
Pstatus	.
Medu	.
Fedu	.
Mjobhealth	0.257
Mjobother	.
Mjobservices	0.269
Mjobteacher	.
Fjob	.
reason	.
guardian	.
traveltime	.
studytime	.
failures	-0.738
schoolsupyes	-0.452
famsupyes	-0.145
paid	.
activities	.
nursery	.
higheryes	0.239
internet	.
romanticyes	-0.054
famrel	.
freetime	.
goout	-0.215
Dalc	.
Walc	.
health	-0.024
absences	-0.001

Tabela 4.1: Coeficientes estimados pelo modelo Lasso com λ ótimo. As variáveis anuladas estão representadas por um ponto.

Esta análise inicial permitiu identificar as variáveis mais relevantes para a previsão da aprovação ou reprovação dos alunos, informação que será tida em conta nas metodologias preditivas desenvolvidas nas secções seguintes.

5 Modelação Preditiva

Nesta secção vão ser abordados métodos diferentes de aprendizagem estatística, cujo objetivo principal é a previsão da aprovação ou reprovação dos alunos na disciplina de matemática, no 3º Período, associado à interpretabilidade do mesmo modelo.

Vão ser experimentados 3 métodos diferentes: Regressão Lasso, Árvores de Decisão e Floresta Aleatória (*Random Forest*). A métrica de avaliação comum aos modelos irá ser a taxa de acerto (*accuracy*) e os métodos de reamostragem para treino/teste serão validação cruzada com 10 *folds* (*10-fold cross validation*) e, por vezes, uma divisão aleatória simples de 70% das observações para

treino e 30% para teste.

5.1 Regressão Lasso

A regressão Lasso é uma extensão da regressão logística tradicional, mantendo a mesma função *link* (logit), mas incorporando uma penalização Ł$_1$ sobre os coeficientes. Em vez de apenas maximizar a verosimilhança, como na regressão logística clássica, a regressão Lasso maximiza a verosimilhança penalizada, promovendo simultaneamente a seleção automática de variáveis e o controlo da complexidade do modelo.

Ao aplicar o Lasso ao problema de previsão de aprovação/reprovação, o objetivo foi obter um modelo que não apenas selecionasse um subconjunto reduzido de variáveis relevantes, mas também estimasse os seus coeficientes de forma a otimizar a capacidade preditiva, minimizando o risco de *overfitting*. O parâmetro de regularização λ foi escolhido através de validação cruzada, equilibrando o compromisso entre complexidade e desempenho preditivo.

Desta forma, o modelo resultante da regressão Lasso é utilizado diretamente para prever novos casos, esperando-se que apresente uma boa capacidade de generalização para novos alunos.

5.1.1 Ajuste do Modelo

Para o ajuste do modelo preditivo, foi utilizada a regressão Lasso com o parâmetro de regularização λ previamente selecionado através de validação cruzada ($\lambda_{\text{min}} = 0.0231$). O modelo foi treinado utilizando a totalidade dos dados disponíveis, recorrendo à matriz de preditores construída na secção anterior.

Adicionalmente, foi também considerado o modelo obtido com o parâmetro mais simples (λ_{1se}), que inclui um número reduzido de preditores (apenas a variável *failures* permaneceu ativa), com o objetivo de comparar a capacidade preditiva entre modelos mais complexos e mais parcimoniosos.

5.1.2 Avaliação da Performance

A avaliação dos modelos foi realizada utilizando dois procedimentos de reamostragem:

- **Validação Cruzada 10-Fold**: Os dados foram divididos em 10 subconjuntos (*folds*), utilizando 9 para treino e 1 para teste em cada iteração. O procedimento foi repetido para todos os folds.
- **Divisão Aleatória 70%/30%**: Foi realizada uma divisão aleatória simples, reservando 70% dos dados para treino e 30% para teste.

Resultados da Validação Cruzada 10-Fold

Modelo com λ_{min}

		Observado	
		0	1
Predito	0	30	16
	1	100	249

Métrica	**Valor**
Accuracy	70.63%
Sensibilidade	23.08%
Especificidade	93.96%

Tabela 5.1: Matriz de confusão e principais métricas do modelo Lasso com λ_{min} (10-Fold).

Modelo com $\lambda_{\mathbf{1se}}$

		Observado	
		0	1
Predito	0	12	4
	1	118	261

Métrica	Valor
Accuracy	69.11%
Sensibilidade	9.23%
Especificidade	98.49%

Tabela 5.2: Matriz de confusão e principais métricas do modelo Lasso com $\lambda_{1\text{se}}$ (10-Fold).

Resultados da Divisão 70%/30%

- **Modelo com $\lambda_{\mathbf{min}}$**: A precisão no conjunto de teste foi de **63.87%**.
- **Modelo com $\lambda_{\mathbf{1se}}$**: A precisão no conjunto de teste foi de **62.18%**.

Os valores de *accuracy* obtidos na divisão 70%/30% foram ligeiramente inferiores aos da validação cruzada, mas mantêm a mesma ordem de magnitude, demonstrando alguma consistência dos modelos. As respetivas matrizes de confusão e métricas associadas encontram-se nos anexos (A.1).

Apesar de uma precisão global razoável (cerca de 70%), destaca-se a baixa capacidade dos modelos em identificar corretamente os alunos reprovados (classe 0), refletida nas baixas sensibilidades registadas (23.08% para λ_{min} e 9.23% para $\lambda_{1\text{se}}$). Isto comprova que, apesar da taxa global de acerto ser próxima, o modelo mais simples revela bastante dificuldade em acertar na previsão de alunos que reprovam, indicando um ponto negativo nesta abordagem.

Este comportamento sugere que os modelos Lasso ajustados tendem a favorecer a classe maioritária (aprovação), priorizando uma elevada especificidade em detrimento da sensibilidade. Tal viés pode ser consequência do desequilíbrio das classes no conjunto de dados (maior número de alunos aprovados) e da própria natureza da penalização Lasso.

5.2 Árvores de Decisão

As árvores de decisão são métodos de aprendizagem supervisionada conhecidos pela sua elevada interpretabilidade. A sua estrutura hierárquica permite visualizar de forma intuitiva as regras de decisão que levam à classificação de um novo indivíduo. Além disso, são técnicas versáteis, capazes de lidar com dados mistos (categóricos e numéricos) e de identificar a possível relevância das variáveis.

5.2.1 Construção da Árvore

Inicialmente, foi ajustada uma árvore de decisão utilizando todas as observações disponíveis. A variável resposta foi a aprovação ou reprovação (PF), enquanto os preditores foram todas as restantes variáveis (exceto *G1* e *G2*). O modelo completo, sem qualquer restrição de complexidade, apresentou uma taxa de acerto (*accuracy*) de aproximadamente **77.22%**. De notar que, apesar de estarem disponíveis todas as variáveis, apenas foram realmente utilizadas na construção da árvore as variáveis: *failures*, *Mjob*, *goout*, *absences*, *guardian*, *studytime*, *sex*, e *famrel*. A árvore gerada encontra-se representada na Figura 5.1.

De seguida, foi treinada uma árvore utilizando apenas 70% dos dados para treino e 30% para teste. O modelo treinado foi depois avaliado no conjunto de teste, tendo-se obtido os resultados disponíveis na Tabela 5.3.

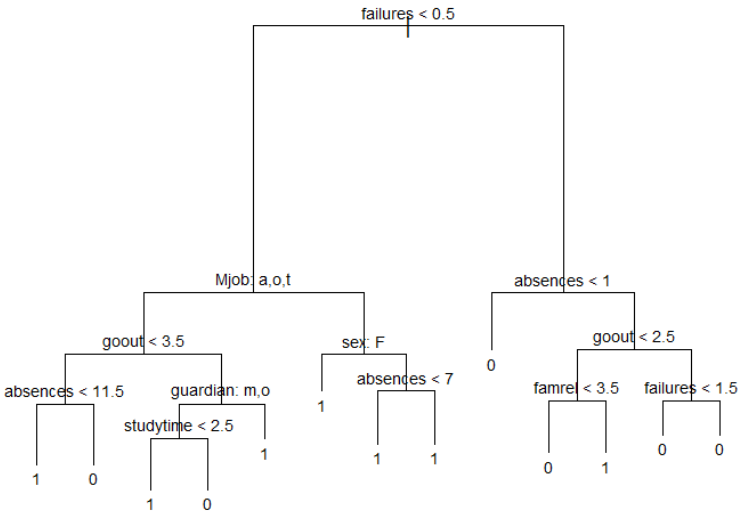


Figura 5.1: Árvore de decisão treinada com todas as observações.

		Observado	
		0	1
Predito	0	21	10
	1	25	63

Métrica	**Valor**
Accuracy	70.59%
Sensibilidade	45.65%
Especificidade	86.30%

Tabela 5.3: Matriz de confusão e principais métricas da árvore de decisão (70%/30% divisão)

5.2.2 Poda e Otimização

Para melhorar a capacidade de generalização do modelo e evitar *overfitting*, foi aplicada a técnica de poda (*pruning*) baseada no erro de classificação (*misclassification rate*). A validação cruzada com 10 *folds* foi usada para identificar o número ótimo de folhas.

A evolução do erro de validação em função do número de folhas é ilustrada na Figura 5.2.

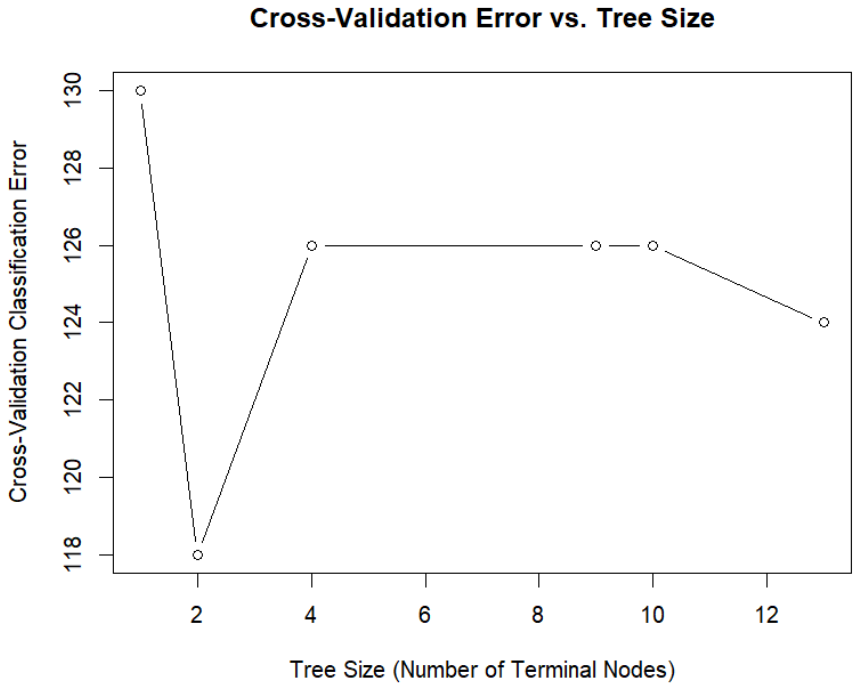


Figura 5.2: Validação cruzada para determinar o número ótimo de folhas na árvore.

Com base na análise, foram consideradas três podas específicas:

- **2 folhas**: modelo extremamente simples, focado apenas na variável *failures*.
- **4 folhas**: modelo com complexidade moderada, incorporando também o número de faltas (*absences*) e a frequência de saídas extra-escola (*goout*).
- **10 folhas**: modelo mais complexo, incluindo mais variáveis como *Mjob*, *famrel*, *guardian* e *studytime*.

As árvores resultantes para cada caso são ilustradas nas Figuras 5.3, 5.4 e 5.5.

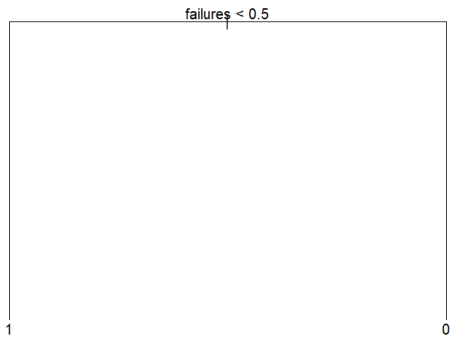


Figura 5.3: Árvore de decisão com 2 folhas.

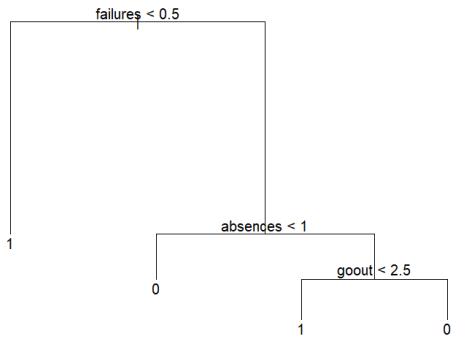


Figura 5.4: Árvore de decisão com 4 folhas.

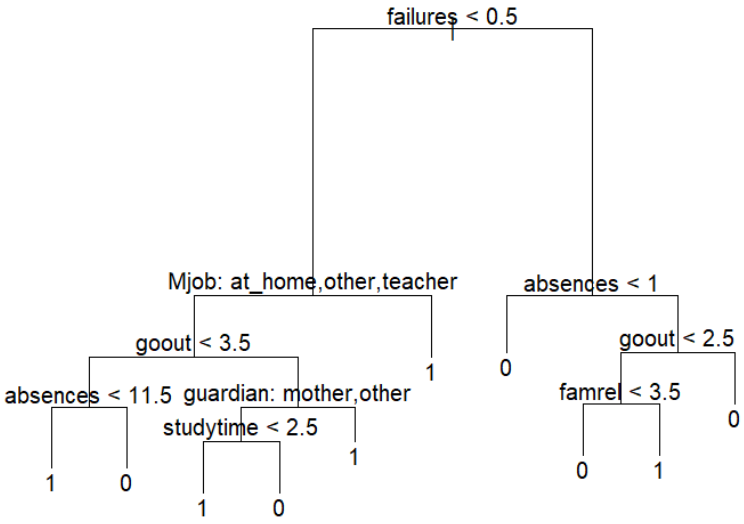


Figura 5.5: Árvore de decisão com 10 folhas.

5.2.3 Resultados e Avaliação

Nesta secção apresentam-se as métricas de desempenho obtidas para cada árvore podada.

		Observado	
		0	1
Predito	0	52	31
	1	78	234

Métrica	**Valor**
Accuracy	72.41%
Sensibilidade	40.00%
Especificidade	88.30%

Tabela 5.4: Matriz de confusão e métricas - árvore com 2 folhas.

		Observado	
		0	1
Predito	0	49	20
	1	81	245

Métrica	**Valor**
Accuracy	74.43%
Sensibilidade	37.69%
Especificidade	92.45%

Tabela 5.5: Matriz de confusão e métricas - árvore com 4 folhas.

		Observado	
		0	1
Predito	0	68	28
	1	62	237

Métrica	**Valor**
Accuracy	77.22%
Sensibilidade	52.31%
Especificidade	89.43%

Tabela 5.6: Matriz de confusão e métricas - árvore com 10 folhas.

Análise Comparativa

- À medida que o número de folhas aumenta, a **accuracy** também aumenta, evidenciando que aumentar o número de variáveis presentes no modelo origina uma maior capacidade de predição.
- A **especificidade** permanece alta em todos os modelos, indicando boa capacidade de identificar corretamente os aprovados.
- A **sensibilidade**, por outro lado, não atinge percentagens elevadas nos 3 modelos, apenas ultrapassando 50% no modelo de maior complexidade, revelando uma maior dificuldade preditiva nos casos de reprovação.
- O modelo com 10 folhas apresenta o melhor compromisso entre **accuracy** e **sensibilidade**.
- A variável ***failures*** está na raiz de todas as árvores podadas, alem da mesma também ***absences*** e ***goout*** estão presentes nas 3 árvores, indicando assim relevância na previsão da aprovação/reprovação.
- Todas as árvores de decisão apresentaram capacidades preditivas elevadas, no entanto é necessário referir que o teste foi realizado com os dados utilizados também para treino, daí haver possibilidade de overfitting e de os resultados serem demasiado otimistas.

Para resolver a questão de sobreajuste das árvores de classificação, uma vez que está a testar com observações usadas para treino, foi dividida a base de dados em 70% para treino e 30% para teste e foram obtidos os seguintes resultados:

		Observado	
		0	1
Predito	0	17	5
	1	29	68

Métrica	**Valor**
Accuracy	71.43%
Sensibilidade	36.96%
Especificidade	93.15%

Tabela 5.7: Matriz de confusão e métricas - árvore com 2 folhas (70%/30% divisão).

		Observado	
		0	1
Predito	0	17	4
	1	29	69

Métrica	**Valor**
Accuracy	72.27%
Sensibilidade	36.96%
Especificidade	94.52%

Tabela 5.8: Matriz de confusão e métricas - árvore com 4 folhas (70%/30% divisão).

		Observado	
		0	1
Predito	0	24	7
	1	22	66

Métrica	**Valor**
Accuracy	75.63%
Sensibilidade	52.17%
Especificidade	90.41%

Tabela 5.9: Matriz de confusão e métricas - árvore com 10 folhas (70%/30% divisão).

Os resultados obtidos com a divisão da base de dados estão em conformidade com os resultados anteriores. Os valores de accuracy são ligeiramente inferiores, como seria de esperar, mas mantêm-se bastante próximos. O mesmo acontece para as outras métricas usadas, indicando que não está a haver overfitting de forma a afetar a capacidade de previsão dos modelos.

5.3 *Random Forest*

A *Random Forest* é um método de aprendizagem estatística baseado na construção de múltiplas árvores de decisão e na combinação dos seus resultados. Ao contrário de uma única árvore de decisão, que pode sofrer de elevado erro de variância, a *Random Forest* gera várias árvores sobre subconjuntos aleatórios dos dados e das variáveis, e agrega as previsões individuais (por votação maioritária, no caso de classificação).

Este procedimento tende a reduzir o *overfitting* e a melhorar substancialmente a capacidade preditiva do modelo. No entanto, perde-se alguma interpretabilidade relativamente a modelos mais simples, como árvores únicas ou regressões Lasso, sendo a *Random Forest* mais orientada para otimizar a precisão de previsão.

5.3.1 Treino e Validação Cruzada

A *Random Forest* foi treinada utilizando a função `train` do pacote `caret`, com validação cruzada de 10 folds para escolher o melhor número de variáveis (*mtry*) a considerar em cada divisão dos nós.

A evolução da *accuracy* média em função de *mtry* foi registada, e é apresentada na Figura 5.6.

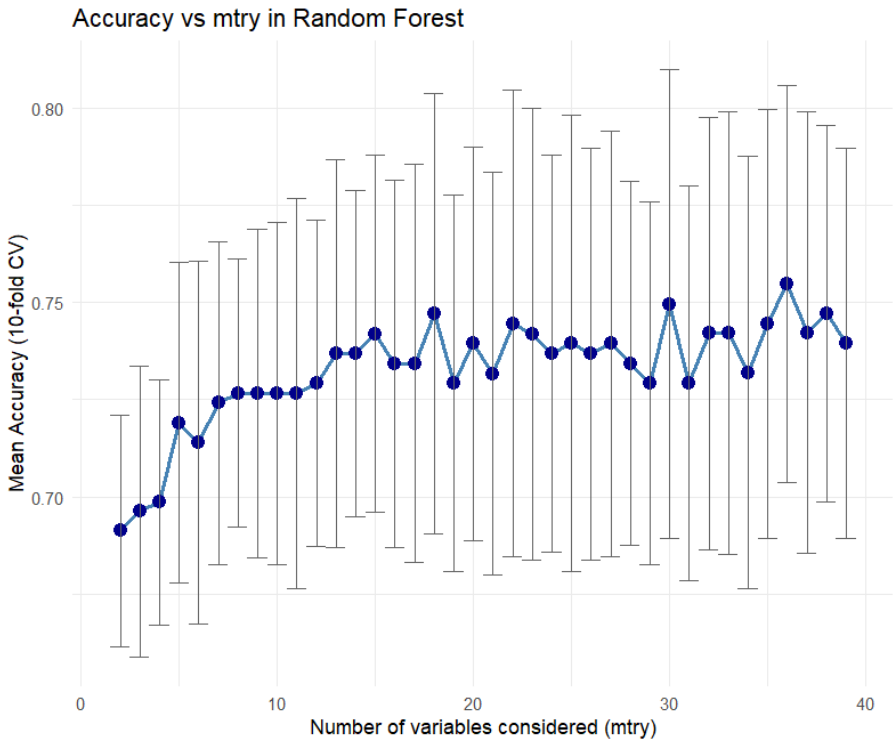


Figura 5.6: Evolução da Accuracy em função do número de variáveis (*mtry*) consideradas em cada split.

Observa-se que os melhores desempenhos foram alcançados com:

- $mtry = 18$: Accuracy média de 74.72%;
- $mtry = 30$: Accuracy média de 74.96%;
- $mtry = 36$: Accuracy média de 75.48%.

Dado o ligeiro aumento de desempenho, o modelo final escolhido foi o correspondente a $mtry = 36$.

5.3.2 Importância das Variáveis

Após o treino, foi analisada a importância das variáveis baseada na contribuição média para a precisão do modelo (*Mean Decrease Accuracy*). As variáveis mais relevantes identificadas foram:

- **failures** (número de reprovações anteriores) - manteve-se em 100% das árvores;
- **absences** (número de faltas) - manteve-se em 43.37% das árvores;
- **schoolsupyes** (apoio educativo extra) - manteve-se em 39.81% das árvores;
- **guardianother** (outro E.E diferente dos pais) - manteve-se em 35.48% das árvores;
- **goout** (frequência de saídas com amigos) - manteve-se em 35.23% das árvores.

O gráfico de importância das variáveis encontra-se na Figura 5.7.

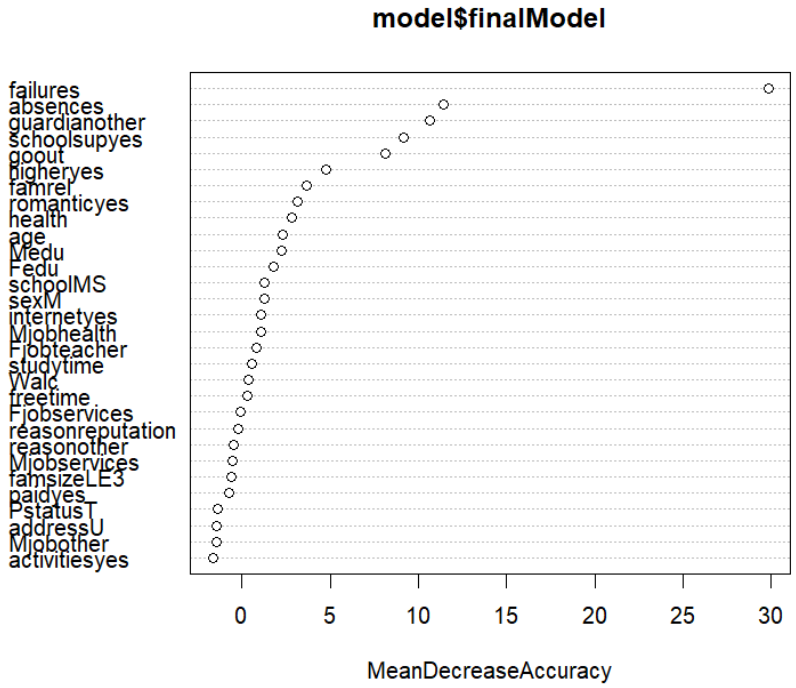


Figura 5.7: Importância das variáveis segundo o impacto médio na Accuracy.

5.3.3 Resultados e Avaliação

A avaliação da *Random Forest* foi realizada através da média dos resultados obtidos nos 10 *folds* da validação cruzada. A matriz de confusão média é apresentada na Tabela 5.10.

		Observado	
		0	1
Predito	0	14.9%	6.6%
	1	18.0%	60.5%

Métrica	**Valor**
Accuracy	75.44%

Tabela 5.10: Matriz de confusão e Accuracy média da *Random Forest* (10-*fold* CV).

O modelo demonstrou um bom desempenho preditivo, com uma *accuracy* média de aproximadamente 75.44%, em *cross-validation* com 10 *folds*, beneficiando da melhor capacidade de previsão deste método em detrimento da interpretabilidade.

6 Conclusão

O presente trabalho teve como objetivo principal a aplicação de diversas técnicas de aprendizagem estatística a um problema de previsão do desempenho académico, recorrendo apenas a variáveis demográficas, sociais e escolares. Para tal, foram explorados métodos de seleção de preditores e de classificação binária, nomeadamente a Regressão Lasso, Árvores de Decisão e *Random Forest*, sempre com especial atenção ao equilíbrio entre capacidade preditiva e interpretabilidade dos modelos.

Em relação à seleção de variáveis, diferentes abordagens evidenciaram de forma consistente alguns preditores como sendo particularmente relevantes para a previsão da aprovação ou reprovação dos alunos. Entre estes, destacam-se o número de reprovações anteriores (*failures*), o número de faltas (*absences*), o apoio educativo extra (*schoolsup*), a vontade de prosseguir para o ensino superior (*higher*), o encarregado de educação (*guardian*), e o tempo de lazer com amigos (*goout*). Estes fatores surgiram repetidamente como preditores chave nos diferentes modelos, o que corrobora a sua relevância prática no contexto académico analisado.

Quanto ao desempenho preditivo, a ***Random Forest*** foi o método que apresentou melhor *accuracy* média (cerca de 75.4%), seguido das Árvores de Decisão podadas com 10 folhas (cerca de 77% no treino e 75% no teste) e, por fim, da Regressão Lasso (cerca de 70%). Este resultado é coerente com as características intrínsecas dos métodos: enquanto a *Random Forest* prioriza a capacidade preditiva, mesmo à custa da interpretabilidade, o Lasso favorece modelos mais parcimoniosos e interpretáveis.

No que respeita à questão de quantas variáveis são estritamente necessárias para obter um modelo de previsão adequado, observou-se que modelos simples, como a árvore de decisão com apenas 2 folhas (baseada praticamente só na variável *failures*), já permitiam obter *accuracies* razoáveis (cerca de 72%). No entanto, a inclusão de um conjunto mais alargado de variáveis - cerca de 8 a 10 preditores - permitiu melhorar significativamente o desempenho, sugerindo que, para este problema, não é necessário utilizar a totalidade das variáveis disponíveis para atingir bons resultados.

De forma geral, os métodos aplicados demonstraram boa capacidade de previsão do sucesso académico com base em fatores indiretos, validando a importância de uma seleção criteriosa de preditores e da utilização de técnicas de aprendizagem estatística para apoiar decisões educacionais.

7 Referências

Referências

[1] UCI Machine Learning Repository, *Student Performance Data Set*. Disponível em: `http://archive.ics.uci.edu/dataset/320/student+performance`. Acesso em abril de 2025.

[2] Cortez, P., & Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance*. Universidade do Minho. Disponível em: `https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf`. Acesso em abril de 2025.

[3] Material de Apoio da Unidade Curricular Teoria da Aprendizagem Estatística, *Fichas de Trabalho em R*. Departamento de Matemática, Universidade do Minho (ano letivo 2024/2025).

[4] Slides Teóricos da Unidade Curricular Teoria da Aprendizagem Estatística, *Teoria da Aprendizagem Estatística*. Departamento de Matemática, Universidade do minho(ano letivo 2024/2025).

A Anexos

A.1 Matriz de Confusão - Modelo Lasso com λ_{min} (Divisão 70%/30%)

Nesta secção encontram-se as matrizes de confusão obtidas a partir da divisão aleatória dos dados em 70% para treino e 30% para teste, tanto para o modelo Lasso com λ_{min} como para o modelo mais simples com λ_{1se}.

Prediction \ Reference	**0**	**1**
0	6	3
1	40	70

Tabela A.1: Matriz de confusão do modelo Lasso com λ_{min} (divisão 70%/30%).

Métricas principais:

- **Accuracy**: 63.87%
- **Sensibilidade**: 13.04%
- **Especificidade**: 95.89%

A.2 Matriz de Confusão - Modelo Lasso com λ_{1se} (Divisão 70%/30%)

Prediction \ Reference	**0**	**1**
0	2	1
1	44	72

Tabela A.2: Matriz de confusão do modelo Lasso com λ_{1se} (divisão 70%/30%).

Métricas principais:

- **Accuracy**: 62.18%
- **Sensibilidade**: 4.35%
- **Especificidade**: 98.63%

A.3 Resultados Complementares da *Random Forest*

Random Forest treinada com 500 árvores e $mtry = 36$ (melhor parâmetro identificado via validação cruzada):

- **Número de árvores**: 500
- **Número de variáveis consideradas em cada split**: 36
- **Erro fora-da-amostra (OOB error)**: 28.35%

Matriz de confusão associada ao erro OOB:

		Observado	
		0	1
Predito	0	51	79
	1	33	232

Tabela A.3: Matriz de confusão da *Random Forest* baseada no erro OOB.

Importância das Variáveis (Mean Decrease Accuracy)

Importância relativa das variáveis para a melhoria da *Accuracy* na *Random Forest*:

Variável	**Importância**
failures	100.00
absences	43.37
schoolsupyes	39.81
guardianother	35.48
goout	35.23
higheryes	24.49
famrel	22.51
romanticyes	20.13
health	19.00
sexM	16.96
schoolMS	15.01
Medu	15.00
age	14.73
Fedu	14.09
studytime	13.73
Fjobteacher	13.19
Mjobhealth	12.45
Walc	11.87
internetyes	11.77
Fjobservices	10.93

Tabela A.4: Importância das 20 variáveis mais relevantes na *Random Forest*.

Anexo Técnico

O código R desenvolvido para a realização deste trabalho encontra-se entregue em ficheiro separado (projeto_student_RuiAlves.R).