Teoria do Aprendizado Estatístico - Prova 01

Nome: Ellen Ozores e Mel Yukari

DESCRIÇÃO DO DATASET

Nosso dataset aborda o desempenho escolar dos alunos no ensino secundário de duas escolas portuguesas. A coleta dos dados foi realizada por meio de relatórios escolares e questionários. A base foi usada para publicação de um trabalho na área da ciência da computação, que teve como título: "Using data mining to predict secondary school student performance", além disso, o dataset é licenciado pela Creative Commons Attribution 4.0 International (CC BY 4.0) license, dessa forma, a base se torna confiável, justamente por possuir uma verificação e ser utilizada em um trabalho acadêmico.

DICIONÁRIO DOS DADOS

Nome da Variável	Tipo	Unidade	Descrição	Observação
school	qualitativa nominal (character)	não se aplica	escola do estudante	binário: 'GP' = Gabriel Pereira ou 'MS' = Mousinho da Silveira
sex	qualitativa nominal (binary)	não se aplica	gênero do estudante	binário: 'F' = feminino ou 'M' = masculino
age	quantitativa discreta (integer)	não se aplica	idade do estudante	numérico: de 15 a 22
address	qualitativa nominal (character)	não se aplica	tipo de endereço residencial do aluno	binário: 'U' = urbano ou 'R' = rural
famsize	qualitativa nominal	não se aplica	tamanho da família	binário: 'LE3' = menos

	(character)			ou igual a 3 ou 'GT3' = maior que 3
pstatus	qualitativa nominal (character)	não se aplica	situação de coabitação dos pais	binário: 'T' = morando junto ou 'A' = separado
medu	qualitativa ordinal (integer)	não se aplica	educação da mãe	numérico: 0 = nenhuma, 1 = ensino fundamental, 2 = do 5° ao 9° ano, 3 = ensino médio, 4 = ensino superior
fedu	qualitativa ordinal (integer)	não se aplica	educação do pai	numérico: 0 = nenhuma, 1 = ensino fundamental, 2 = do 5° ao 9° ano, 3 = ensino médio, 4 = ensino superior
mjob	qualitativa nominal (character)	não se aplica	trabalho da mãe	nominal: 'professor, 'saúde' relacionado a cuidados, 'serviços' civis, 'em casa' ou 'outros'
fjob	qualitativa nominal (character)	não se aplica	trabalho do pai	nominal: 'professor, 'saúde' relacionado a cuidados, 'serviços' civis, 'em casa' ou 'outros'
reason	qualitativa nominal (character)	não se aplica	razão da escolha da escola	nominal: perto de 'casa', 'reputação' da escola, preferência do 'curso' ou 'outro'

guardian	qualitativa nominal (character)	não se aplica	guardião legal do estudante	nominal: 'mãe', 'pai' ou 'outro'
traveltime	qualitativa ordinal (integer)	não se aplica	tempo de viagem de casa para a escola	numérico: 1 = < 15 min, 2 = 15 a 30 min, 3 = 30 min a 1 hr ou 4 = > 1 hr
studytime	qualitativa ordinal (integer)	não se aplica	tempo de estudo semanal	numérico: 1 = < 2 hrs, 2 = 2 a 5 hrs, 3 = 5 a 10 hrs ou 4 = > 10 hrs
failures	qualitativa ordinal (integer)	não se aplica	número de reprovações anteriores	numérico: n se 1 <= n < 3, senão 4
schoolsup	qualitativa nominal (binary)	não se aplica	apoio educacional extra	binário: sim ou não
famsup	qualitativa nominal (binary)	não se aplica	apoio educacional da família	binário: sim ou não
paid	qualitativa nominal (binary)	não se aplica	aulas extras pagas dentro da disciplina do curso (Matemática ou Português)	binário: sim ou não
activities	qualitativa nominal (binary)	não se aplica	atividades extra curriculares	binário: sim ou não
nursery	qualitativa nominal (binary)	não se aplica	frequentou a creche	binário: sim ou não
higher	qualitativa nominal (binary)	não se aplica	quer fazer um curso superior	binário: sim ou não
internet	qualitativa nominal (binary)	não se aplica	acesso à internet em casa	binário: sim ou não
romantic	qualitativa	não se aplica	com um	binário:

	nominal (binary)		relacionamento amoroso	sim ou não
famrel	qualitativa ordinal (integer)	não se aplica	qualidade das relações familiares	numérico: 1 = muito ruim a 5 = excelente
freetime	qualitativa ordinal (integer)	não se aplica	tempo livre depois da escola	numérico: 1 = muito baixo a 5 = muito alto
goout	qualitativa ordinal (integer)	não se aplica	sair com os amigos	numérico: 1 = muito baixo a 5 = muito alto
dalc	qualitativa ordinal (integer)	não se aplica	consumo de álcool durante o dia de trabalho	numérico: 1 = muito baixo a 5 = muito alto
walc	qualitativa ordinal (integer)	não se aplica	consumo de álcool no fim de semana	numérico: 1 = muito baixo a 5 = muito alto
health	qualitativa ordinal (integer)	não se aplica	estado de saúde atual	numérico: 1 = muito ruim a 5 = muito boa
absences	quantitativa discreta (integer)	não se aplica	número de faltas escolares	numérico: de 0 a 93
g1	quantitativa discreta (integer)	não se aplica	nota do primeiro período	numérico: de 0 a 20
g2	quantitativa discreta (integer)	não se aplica	nota do segundo período	numérico: de 0 a 20
g3	quantitativa discreta (integer)	não se aplica	nota final	numérico: de 0 a 20

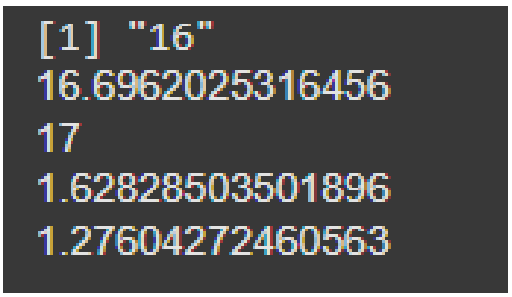
TRANSFORMAÇÃO DOS DADOS

Para carregar e visualizar o dataset, o primeiro passo foi instalar o pacote “readr”, que possibilita a leitura dos dados da base por meio do comando “head” e “names”. Visando evitar complicações, todas as variáveis passaram a ser chamadas com letras minúsculas, usando a função “tolower”. Em seguida, os dados que a princípio eram qualitativos nominais, foram transformados em factor e os qualitativos ordinais em factor ordenado.

CÁLCULOS ESTATÍSTICOS

Realizou-se o cálculo da moda, média, mediana, variância e desvio padrão entre todas as variáveis quantitativas discretas do dataset.

Figura 01 - Cálculos Estatísticos para a variável Idade



Fonte: Elaborado pelos autores, 2025.

No exemplo da figura, estão os resultados obtidos sobre a variável idade, nela, a faixa-etária coletada vai de 15 a 22 anos. Dessa forma, partindo para a interpretação dos valores, a moda = 16, média ≈ 16.7, mediana = 17, variância ≈ 1.63 e o desvio padrão ≈ 1.27.

A tabela de frequência também foi construída, como mostra a figura a seguir, que apresenta uma exemplificação dos resultados obtidos com a variável sexo.

Figura 02 - Tabela de Frequência para a variável Sexo

A matrix: 3 × 3 of type dbl

	freq_absoluta	freq_relativa	p_freq_rel
F	208	0.53	52.66
M	187	0.47	47.34
Total	395	1.00	100.00

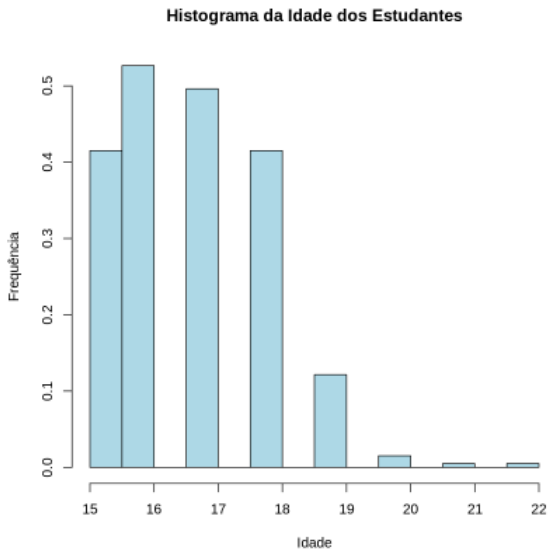
Fonte: Elaborado pelos autores, 2025

A tabela mostra a frequência absoluta e relativa (tanto em decimal, quanto em porcentagem) do gênero dos estudantes. Dito isso, de 395 alunos registrados (100%), 208 são do público feminino, se tornando a maioria nas escolas, com aproximadamente 52,6%, já o público masculino, apresenta a soma de 187 estudantes, cerca 47,34% sobre o valor total.

HISTOGRAMA

Na construção do histograma, utilizou-se a função “hist( )”, que possibilita a criação desse gráfico, por meio da variável Idade, permitindo observar a distribuição de frequência e analisar os padrões de faixa etária nas escolas.

Figura 03 - Histograma da Idade dos Estudantes



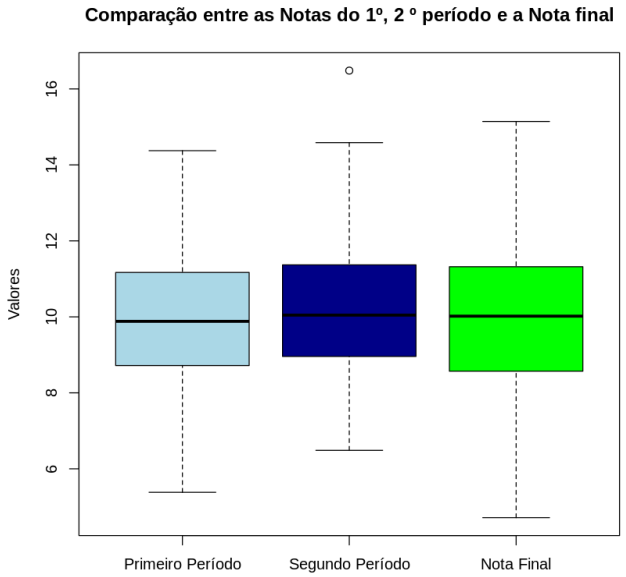
Fonte: Elaborado pelos autores, 2025

Esse histograma demonstra no seu eixo x, as idades, e no eixo y, a frequência relativa. Diante disso, podemos observar que a maioria dos alunos possuem 16 anos, valor apontado pelo algoritmo ao calcularmos a moda, citada anteriormente.

BOX-PLOT

É um tipo de gráfico que mostra o mínimo, o primeiro quartil (Q1), a mediana (Q2), o terceiro quartil (Q3) e o máximo de um conjunto de dados. Ele foi aplicado com as variáveis g1, g2 e g3, que apresentam a nota do primeiro período, segundo e a final, respectivamente.

Figura 04 - Box-Plot das Variáveis g1, g2 e g3



Fonte: Elaborado pelos autores, 2025.

Antes de analisar o gráfico, é importante destacar que as escolas são portuguesas, então o sistema de notas neste país vai de 0 a 20, no qual de 0 a 9 o aluno é reprovado, nota igual a 10 é o suficiente para ser aprovado e valores maiores indicam um bom desempenho estudantil.

No primeiro período (azul claro), a nota mínima atingida foi em torno de 6 e a máxima 14, o quadro azul representa a concentração das notas, indicando que 50% delas estão entre 9 e 11, aproximadamente. Por fim, a linha preta representa a mediana das notas, resultando em 10.

O segundo período de notas (azul escuro), apresenta 7 como valor mínimo e um pouco mais que 14 como máximo. A mediana das notas se concentra em torno de 10 e há um outlier acima de 16, neste caso, provavelmente é um estudante que teve desempenho excepcionalmente alto no 2º período.

A nota final abrange uma diferença maior entre número máximo e mínimo de de notas, porém a mediana continua próxima de 10 e apresenta uma distribuição semelhante com o primeiro período.

A maioria dos dados ficam entre o limite inferior e superior. No caso do primeiro período e a nota final, praticamente 100% dos dados estão dentro dos limites, já no segundo período, como existe um outlier, a porcentagem diminui para 99% aproximadamente. Em suma, é notório que a maioria dos alunos tiram a nota mínima de aprovação nas escolas presentes no dataset.

TABELA CRUZADA

A tabela cruzada é uma ferramenta de análise de dados que apresenta a relação entre duas ou mais variáveis, normalmente categóricas, organizando os dados numa grade de linhas e colunas para mostrar a frequência e os padrões das combinações.

Figura 05 - Tabela Cruzada das Variáveis school e address

	R	U	Total_coluna
GP	63	286	349
MS	25	21	46
Total_linha	88	307	395

Fonte: Elaborado pelos autores, 2025.

Conforme a figura 05, a tabela mostra que no dataset, há 349 alunos da escola Gabriel Pereira (GP), entre eles, 63 vêm da zona rural de Portugal e 286 vivem em cenários urbanos.

Apenas 46 estudantes na base estão matriculados na escola Mousinho da Silveira (MS), dentre eles, 25 moram em zonas rurais e 21 nas urbanas.

Mesmo com pouca diferença, Mousinho da Silveira tem mais alunos da zona rural, enquanto a escola Gabriel Pereira, possuí um grupo majoritário de estudantes dos centros urbanos.

CORRELAÇÃO POR QUI-QUADRADO

A correlação por qui-quadrado é um teste estatístico que avalia a associação significativa entre duas variáveis categóricas. Ele verifica se as frequências observadas nos dados diferem significativamente das frequências esperadas, indicando uma relação real entre as variáveis de estudo.

Figura 06 - Correlação por Qui-Quadrado entre as variáveis failures e famsup

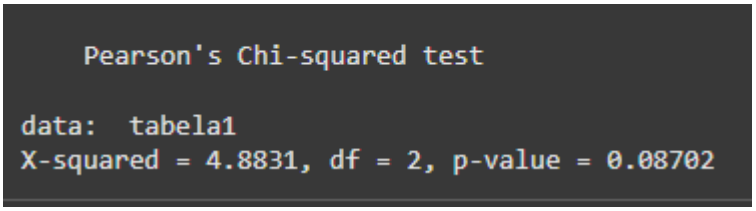
```
    Pearson's Chi-squared test

data:  tabela1
X-squared = 4.6147, df = 3, p-value = 0.2023
```

Fonte: Elaborado pelos autores, 2025

A variável “failures” no dataset, significa as reprovações dos alunos, enquanto “famsup”, aponta se tais recebem apoio educacional da família. Dito isso, foi calculado o qui-quadrado dessas duas variáveis para analisar se há correlação, ou seja, verificar se para os estudantes que não recebem auxílio familiar, têm reflexo nas suas reprovações.

No teste obteve-se um valor de $X^2 = 4.6147$, com 3 graus de liberdade e p-valor de 0.2023. Considerando o nível de significância de 5%, observa-se que o p-valor é superior ao limiar adotado, indicando ausência de associação estatisticamente significativa. Assim, o apoio familiar (famsup) não mostrou influência relevante sobre o número de reprovações dos alunos (failures) nesta amostra.

Figura 07 - Correlação por Qui-Quadrado entre as variáveis pstatus e guardian



Fonte: Elaborado pelos autores, 2025

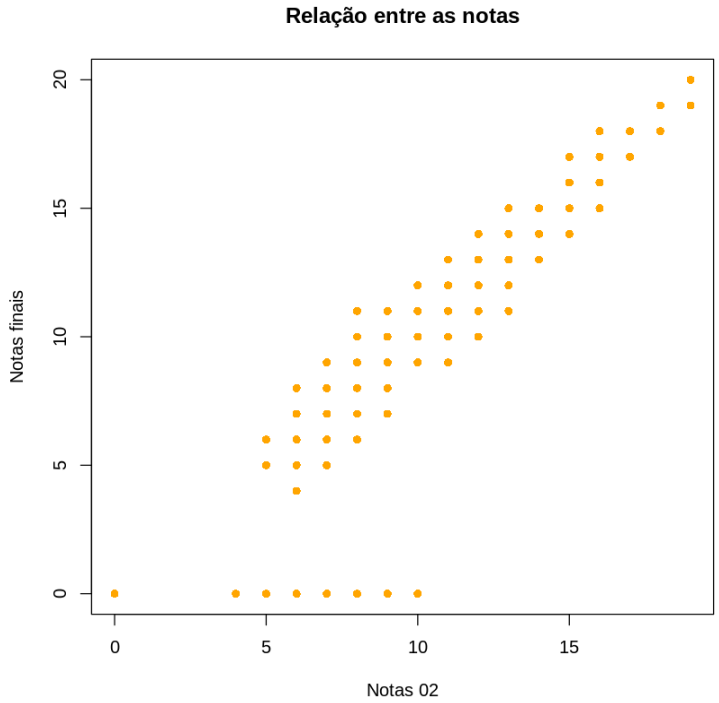
Seguindo a mesma lógica, agora com as variáveis “pstatus” e “guardian”, que indicam a situação de coabitação dos pais e a guarda legal do estudante, respectivamente, a aplicação do teste Qui-Quadrado entre as variáveis resultou em $X^2 = 4.8831$, com 2 graus de liberdade e p-valor de 0.08702. Adotando o nível de significância de 5%, verifica-se que o p-valor obtido é maior do que o valor de referência estabelecido, os valores achados mostram que não há suporte estatístico consistente para afirmar uma associação entre a coabitação dos pais e o responsável legal do estudante. Contudo, observa-se que o p-valor se aproxima do limiar de significância, o que pode indicar uma possível inclinação de dependência entre as variáveis.

Se fosse adotado um nível de significância de 10% ($\alpha = 0,10$), o resultado seria significativo, indicando possível associação entre a coabitação dos pais e a guarda legal do estudante, sugerindo que a relação merece investigação em estudos futuros.

GRÁFICO DE DISPERSÃO

O gráfico de dispersão visualiza a relação entre duas variáveis numéricas, plotando pontos em um plano para mostrar a tendência ou correlação entre elas, como mostra a figura 07.

Figura 07 - Gráfico de Dispersão em relação às notas (g1 e g3)



Fonte: Elaborado pelos autores, 2025.

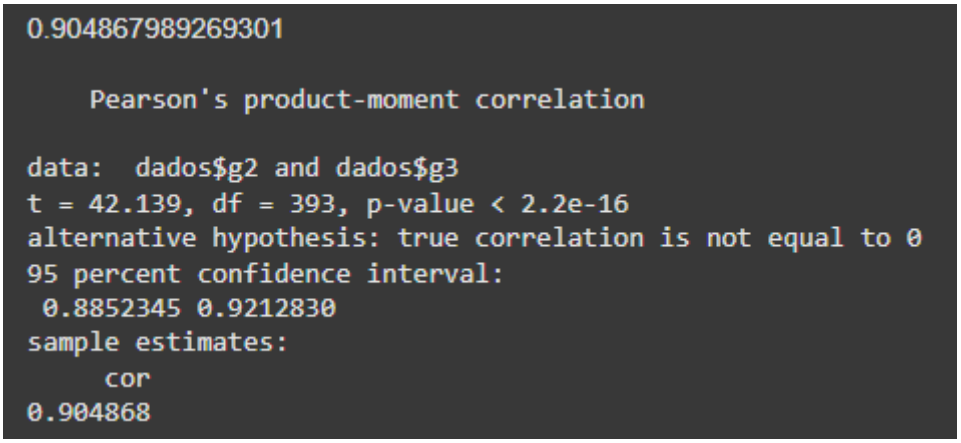
A figura 07 apresenta o resultado do gráfico de dispersão criado a partir das notas do segundo período e as finais. Desse modo, é notório que os pontos não se dispersam muito, formando uma semelhança a uma reta, esse fator indica uma relação forte entre as variáveis, além disso, elas tendem a uma correlação positiva.

Portanto, esse gráfico sugere que o desempenho do segundo período pode ser um bom preditor do desempenho final, ou seja, quem foi melhor no 2º período, em geral, teve melhores notas finais.

Os pontos que se encontram na parte inferior, mostram algumas notas finais iguais a 0, mesmo com as do segundo período diferentes, isso pode acontecer ao cogitar um desempenho ruim na prova final, falta ou reprovação.

Teste de Correlação

Figura 08 - Teste de Correlação em relação às notas (g2 e g3)



```
0.904867989269301

    Pearson's product-moment correlation

data:  dados$g2 and dados$g3
t = 42.139, df = 393, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8852345 0.9212830
sample estimates:
     cor
0.904868
```

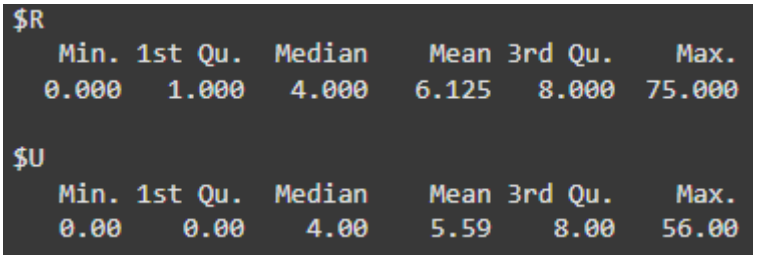
Fonte: Elaborado pelos autores, 2025.

O teste de correlação resultou em 0,90 aproximadamente, apontando que as notas do segundo período e as notas finais estão fortemente e positivamente correlacionadas, portanto, alunos com notas mais altas no segundo período tendem, de maneira consistente, a ter notas finais mais altas, ademais, esse teste confirma o que o gráfico de dispersão já mostrava visualmente.

RELAÇÃO ENTRE VARIÁVEL QUALITATIVA E QUANTITATIVA

A análise entre variáveis qualitativas e quantitativas ajuda a identificar como características do contexto dos alunos podem impactar seu desempenho ou comportamento. Neste caso, foram usadas as variáveis “address” (tipo de endereço: urbano ou rural) e “absences” (número de faltas escolares) que permitem observar se o local de moradia influencia na frequência escolar dos estudantes.

Figura 09 - Relação entre as variáveis qualitativa (address) e quantitativa (absences)



```
$R
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   4.000   6.125   8.000  75.000

$U
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    4.00    5.59    8.00   56.00
```

Fonte: Elaborado pelos autores, 2025.

Os resultados indicam que tanto os alunos residentes em área urbana quanto os de área rural apresentam a mesma mediana de 4 faltas, sugerindo comportamento semelhante em termos de padrão central, no entanto, a média de faltas dos estudantes da zona rural (6,12) é levemente superior à dos alunos da zona urbana (5,59), além disso, nota-se maior dispersão entre os residentes da zona rural, cujo número máximo de ausências atinge 75 faltas, enquanto no grupo urbano, o valor máximo observado foi 56.

Esses resultados apontam que, embora não haja diferenças relevantes no padrão central, alunos da zona rural tendem a apresentar casos mais extremos de ausência, possivelmente relacionados a fatores de deslocamento ou dificuldade de acesso à escola.

REFERÊNCIAS

CORTEZ, Paulo; SILVA, Alice. *Student Performance*. UCI Machine Learning Repository, 2008. Disponível em: https://archive.ics.uci.edu/dataset/320/student+performance

CORTEZ, Paulo; SILVA, Alice. *Using Data Mining to Predict Secondary School Student Performance*. Braga: Universidade do Minho, 2008.

GitHub com o trabalho completo: https://github.com/melyukari/Prova-1---R.git