Vanessa Berwanger Wille - 211708026

Trabalho de Modelagem Estatística:
Predição do desempenho de alunos do Ensino Secundário em Portugal

Brasil

Junho de 2023

Vanessa Berwanger Wille - 211708026

Trabalho de Modelagem Estatística:
Predição do desempenho de alunos do Ensino Secundário em Portugal

Trabalho elaborado para a disciplina de Modelagem Estatística com o objetivo de analisar os fatores que influenciam desempenho de alunos do Ensino Médio de duas escolas Portuguesas e realizar previsões em relação às notas das avaliações finais dos alunos.

Fundação Getulio Vargas - RJ
Escola de Matemática Aplicada

Professor: Luiz Max F. Carvalho

Brasil
Junho de 2023

Sumário

1 Introdução

1.1 Problema e relevância

A educação exerce um papel fundamental na formação dos indivíduos e no desenvolvimento da sociedade, sendo, assim, importante compreender fatores que influenciam o sistema educacional. Entre eles, destacam-se aspectos ambientais, econômicos, sociais e familiares, que podem interferir negativamente ou positivamente no processo de aprendizagem do aluno (1).

Dito isso, o objetivo desse trabalho é explorar estatisticamente dados reais sobre alunos e procurar identificar fatores que, possivelmente, levam a uma melhor performance (ou não) dos mesmos e prever, de acordo com a circunstâncias na qual um aluno específico se encontra, qual seria seu desempenho. É válido observar que a abordagem se dá em um contexto bastante específico e os resultados obtidos não devem ser generalizados para a educação como um todo, mas podem trazer percepções interessantes para auxiliar, por exemplo, os educadores, família e órgãos públicos a ajustar melhor seus estilos de trabalho e fornecer ajuda e incentivo aos alunos de maneira adequada.

É conveniente ressaltar que um enfoque para esse cenário é relevante para que a sociedade entenda que existem disparidades entre os alunos ocasionados por fatores externos a escola e é necessário lidar com isso da melhor maneira possível, buscando uma educação mais inclusiva.

Ainda, para esse trabalho, focaremos em dados relacionados a escolas portuguesas. Segundo um estudo da Direção-Geral de Estatísticas da Educação e da Ciência, DGEEC, até 2019 apenas 52% das pessoas entre os 25 e os 64 anos terminaram o secundário ou tiraram um curso superior em Portugal (8). Isso faz com a análise seja ainda mais pertinente, pois vamos avaliar fatores de um país por vezes considerado um dos menos educados da União Europeia e que mais do que outros, precisa encarar e empenhar-se na busca de melhorias.

1.2 Dados

Os dados para esse trabalho estão disponíveis no repositório de *Machine Learning* da Universidade da Califórnia, Irvine, titulado como "Student Performance"(2), abordando o desempenho dos alunos no ensino secundário de duas escolas portuguesas. São fornecidos dois conjuntos de dados relativos ao desempenho em duas disciplinas distintas: Matemática e Língua Portuguesa. O último, por apresentar mais amostras, foi o escolhido.

Os dados brutos, coletadas por meio de relatórios escolares e questionários, contêm 649 observações e 33 variáveis, relacionados à escola (como apoio educacional extra escolar), relacionados ao aluno (como desempenho no curso anterior, idade, tempo de estudo, desejo de prosseguir com o ensino superior) e relacionados à família (como status dos pais, qualidade da relação familiar, escolaridade e trabalho dos pais).

Dessas variáveis, o atributo alvo será a nota final "G3". É observado no próprio conjunto de dados que esse atributo possui forte correlação com os atributos G1 e G2, que correspondem às notas do 1º e 2º período, sendo assim mais difícil prever G3 sem G2 e G1. Contudo, essa previsão é muito mais útil na busca de explicações de influências de outros fatores nas notas dos alunos, que é um dos objetivos desse trabalho.

2 Metodologia

Neste tópico, serão explicadas as abordagens escolhidas para examinar fatores que influenciam no desempenho acadêmico de alunos e para realizar as previsões.

2.1 Modelos

O modelo selecionado para seguir esse trabalho foi o de regressão linear multivariada, amplamente aplicado na área de análise estatística. Essa escolha se baseia, especialmente, na suposição de que existe uma relação linear entre as notas finais dos alunos (variável dependente) e os diversos atributos (variáveis independentes), o que é fundamentado por teorias prévias.

Uma das principais vantagens da regressão linear múltipla é a interpretabilidade dos resultados. Os coeficientes estimados para cada variável fornecem informações sobre o impacto médio delas na avaliação acadêmica, controlando-se os demais fatores. Essa interpretabilidade permite a identificação dos fatores mais influentes no desempenho dos alunos.

A construção dos modelos de regressão múltipla foi baseada na correlação da variável dependente em relação às outras variáveis, na análise de variância (ANOVA) (7), realizado para descobrir a relação entre os diferentes grupos de dados categóricos, na análise exploratória de dados e, ainda, na remoção da variável menos significativa nos testes, isso é, com o maior p-valor. Além disso, considerações foram feitas em relação à ameaça de multicolinearidade na utilização de variáveis preditivas para construir cada modelo.

2.2 Ajuste

Para o ajuste do modelo, optou-se pela abordagem frequentista (ao invés da bayesiana), onde busca-se entender os coeficientes β que minimizam a soma dos quadrados residuais, isso é, as diferenças quadráticas entre os valores conhecidos (y) e as saídas do modelo previsto ($\hat{y}$) (3). A soma residual dos quadrados é uma função dos parâmetros do modelo:

$$RRS(\beta) = \sum_{i=1}^{N}(y_i - \hat{y})^2 = \sum_{i=1}^{N}(y_i - \beta^T x_i)^2.$$

Esta equação tem uma solução de forma fechada para os parâmetros β do modelo que minimizam o erro. Isso é conhecido como a estimativa da máxima verossimilhança de β porque é o valor que é mais provável, dadas as entradas X e saídas y.

O que obtemos da regressão linear frequentista é uma única estimativa para os parâmetros do modelo com base apenas nos dados de treinamento. Isso é, o modelo é completamente informado pelos dados: nessa visão, tudo o que precisamos saber para o nosso modelo é codificado nos dados de treinamento que temos disponíveis.

Nesse sentido, a escolha por essa abordagem deu-se pela quantidade considerável de dados a dispor, objetividade dos resultados e por questões de simplicidade.

As análises serão feitas no *Python*, utilizando principalmente a biblioteca *statsmodels.api* e dividindo os dados em uma proporção de 80% para treino e 20% para testes.

2.3 Avaliação

Avaliar os modelos e a bondade dos ajustes é uma etapa importante para verificar se as previsões efetuadas se adéquam aos dados observados. Para isso, serão considerados os seguintes tópicos:

- Akaike Information Criterion (AIC): Métrica que mensura a qualidade de um modelo estatístico visando também a sua simplicidade. Leva em consideração a maximização da verossimilhança (l_n) do modelo e penaliza a inclusão de mais parâmetros (d):

$$AIC = 2d - 2l_n,$$

- Raiz do Erro Quadrático Médio (RMSE): Essa métrica mede a diferença média, respectivamente, entre os valores previstos pelo modelo e os valores reais observados. Quanto menor o valor, melhor é o desempenho do modelo.

- Coeficiente de determinação (R^2): O R^2 mede a proporção da variabilidade total na variável dependente que é explicada pelo modelo. Um valor de R^2 mais próximo de 1 indica que o modelo é capaz de explicar uma porcentagem maior da variação nos dados.

- Gráficos e análises de resíduos (5): Portanto, para um modelo bem-sucedido, é muito crucial validar essas suposições no conjunto de dados fornecido. Esses gráficos ajudam a identificar padrões nos resíduos, como heteroscedasticidade ou não linearidade, que são premissas da regressão linear, e caso não validados, podem indicar deficiências no modelo.

- Coeficiente e seu erro padrão, p-valor e intervalo de confiança: A partir desses valores é possível encontrar uma associação significativa entre as variáveis independentes e a variável dependente, com base nos valores de p significativos (geralmente, p $<$ 0,05) e nos intervalos de confiança que não incluem zero.

3 Resultados

Neste capítulo serão expostos os resultados da análise exploratória do dados, que buscou identificar padrões, tendências e relações entre as variáveis para a construção dos modelos preditivos, bem como o desempenho deles e a relevância de cada atributo nas notas acadêmicas dos alunos.

3.1 Análise exploratória de dados

Inicialmente, avaliamos que a base de dados é bem completa e não existem valores faltantes. Contudo, para a variável "G3"(a ser prevista), existem 15 *outliers* com nota zero. Essas avaliações são, possivelmente, de pessoas que não realizaram a prova ou saíram da escola durante o período e, por acabarem sendo prejudiciais para as análises e modelos, as observações desses alunos foram removidas.

Como citado em métodos, a correlação entre as variáveis influenciou a escolha das mesmas. Para entender melhor essas correlações, plotou-se a seguinte matriz de correlações com *pandas reporting*:

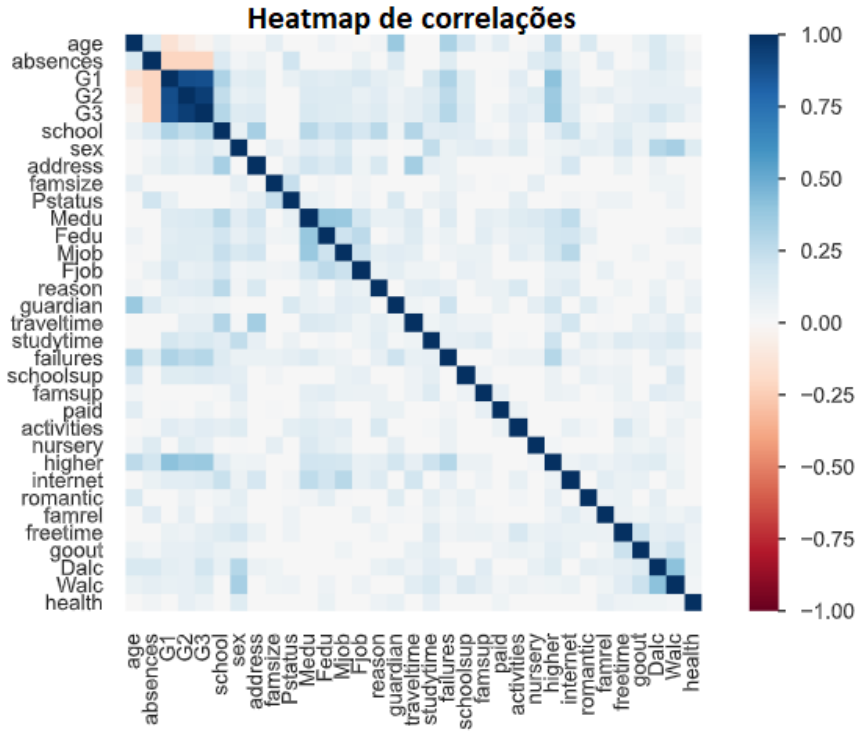


Figura 1 – Heatmap de correlações

Como já mencionado, não se fará uso das variáveis G1 e G2. Contudo, é possível analisar outra váriaveis que apresentem um grau considerável de correlação com G3, entre as quais:

- "studytime": tempo de estudo semanal.

- "school": escola do aluno.
- "higher": quer cursar o ensino superior.
- "failures": número de reprovações anteriores.
- "Medu"e "Fedu": escolaridade da mãe e do pai, respectivamente. Esses dois atributos possuem alta correlação e, portanto, usou-se apenas a escolaridade da mãe, que apresentava uma correlação mais alta, no modelo linear.
- "Dalc"e "Walc": respectivamente, consumo diário e semanal de álcool. Esses dois atributos possuem alta correlação e, portanto, usou-se a média no modelo linear, criando uma nova coluna "alcool".
- "schoolsup": apoio educacional extra.

As variáveis "school", "higher"e "schoolsup"são binárias e, portanto, para fins de modelagem, foram designadas por 1 e 0 (por exemplo, variável "school_GP"recebe 1 se o aluno for da escola Gabriel Pereira e 0 caso contrário). As demais variáveis preditoras categóricas (e que não possuem sentido ordinal) também não podem ser inseridas diretamente em um modelo de regressão e ser interpretadas de forma significativa, sendo necessário recodificá-las em variáveis dummy (foram usadas em testes de modelos).

Ainda, para confirmar as correlações apresentadas no Heatmap, as variáveis categóricas foram avaliadas com boxplots, pelos quais podemos ver que existem tendências de alunos que estudam na escola GP, possuem desejo de ir para a universidade e não recebem apoio escolar extra (o que pode parecer curioso) receberem notas mais altas.Também realizou-se o teste ANOVA, que, no geral, apresentava um valor alto de teste F e um pequeno valor de p, indicando correlação entre a variável alvo e a variável categórica. Um exemplo é para "higher", onde obteve-se valor para o teste F de 80.242 e um p-valor de 3.499e-18.

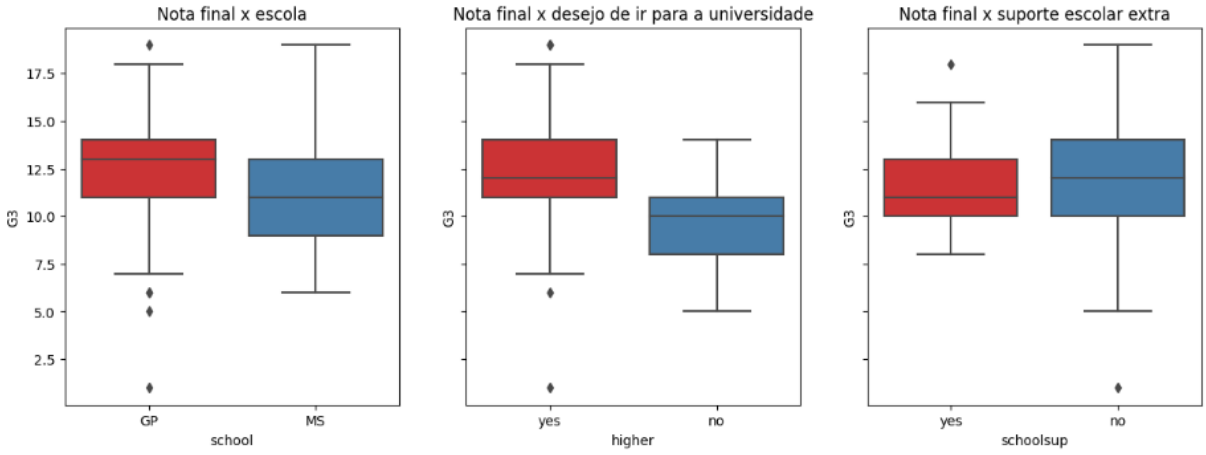


Figura 2 – Boxplots

Ainda, para variáveis ordinais e quantitativas foram plotados gráficos de dispersão, confirmando uma relação linear entre as variáveis, como pode ser visto a seguir.

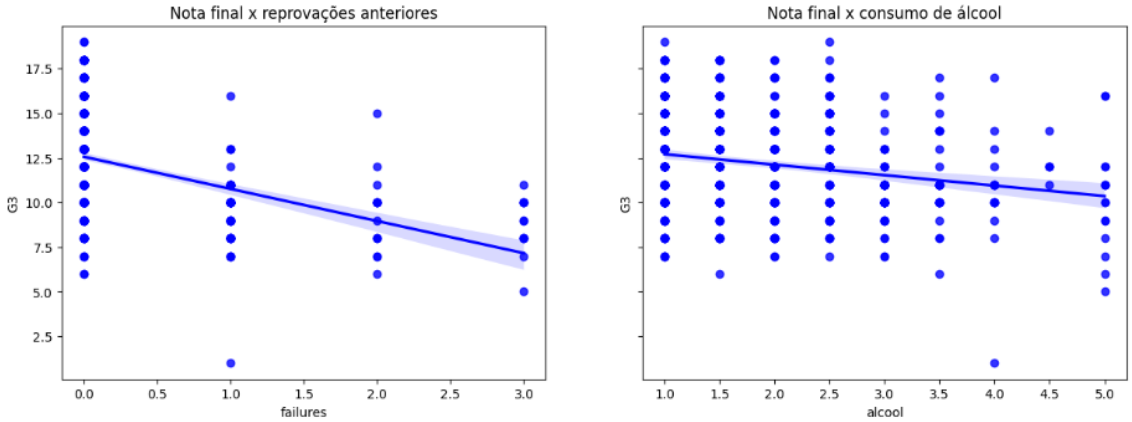


Figura 3 – Relações lineares entre nota e reprovações/consumo de álcool

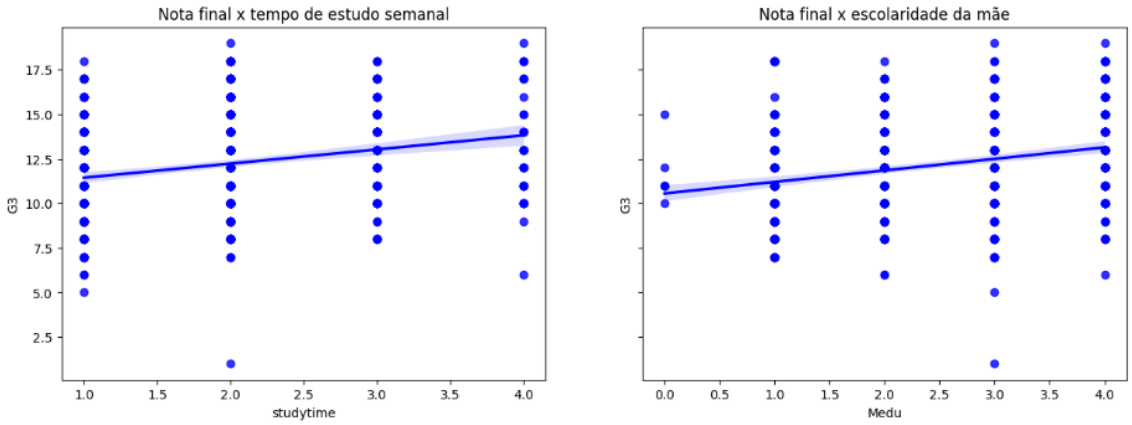
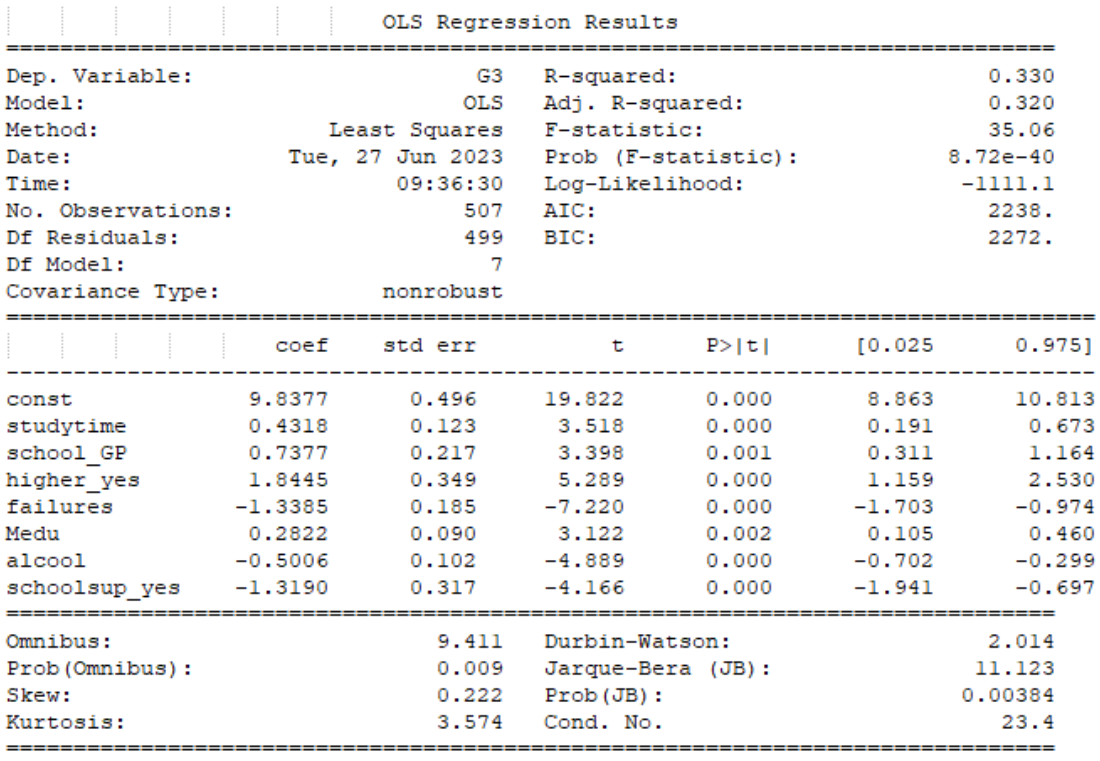


Figura 4 – Relações lineares entre nota e tempo de estudo semanal/escolaridade da mãe

Tendo um entendimento melhor sobre os dados, parte-se para o ajuste dos modelos e a análise dos resultados.

3.2 Ajustes, coeficientes e predições

Os métodos indicados anteriormente foram aplicados para diversos modelos, levando em conta o nível de significância dos coeficientes (intervalo de confiança não incluía o zero), bem como as apurações dos ajustes. Com isso, chegamos ao modelo indicado a seguir (com os dados de treino):



OLS Regression Results

Dep. Variable:	G3	R-squared:	0.330
Model:	OLS	Adj. R-squared:	0.320
Method:	Least Squares	F-statistic:	35.06
Date:	Tue, 27 Jun 2023	Prob (F-statistic):	8.72e-40
Time:	09:36:30	Log-Likelihood:	-1111.1
No. Observations:	507	AIC:	2238.
Df Residuals:	499	BIC:	2272.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.8377	0.496	19.822	0.000	8.863	10.813
studytime	0.4318	0.123	3.518	0.000	0.191	0.673
school_GP	0.7377	0.217	3.398	0.001	0.311	1.164
higher_yes	1.8445	0.349	5.289	0.000	1.159	2.530
failures	-1.3385	0.185	-7.220	0.000	-1.703	-0.974
Medu	0.2822	0.090	3.122	0.002	0.105	0.460
alcool	-0.5006	0.102	-4.889	0.000	-0.702	-0.299
schoolsup_yes	-1.3190	0.317	-4.166	0.000	-1.941	-0.697

Omnibus:	9.411	Durbin-Watson:	2.014
Prob(Omnibus):	0.009	Jarque-Bera (JB):	11.123
Skew:	0.222	Prob(JB):	0.00384
Kurtosis:	3.574	Cond. No.	23.4

Figura 5 – Regressão linear - OLS

Isso é, a predição do desempenho final dos alunos a partir desse modelo pode ser dada pela seguinte fórmula:

$$\begin{aligned}(\hat{G3}) =& 9.8377 + 0.4318(studytime) + 0.7377(school\_GP) + 1.844(higher\_yes) \\ & - 1.338(failures) + 0.2822(Medu) - 0.5(alcool) - 1.319(schoolsup\_yes)\end{aligned}$$

Podemos observar que todos os p-valores são menores que 0.05 e que nenhum intervalo de confiança inclui o zero (apesar de alguns estarem bastante próximos), o que é bastante positivo:

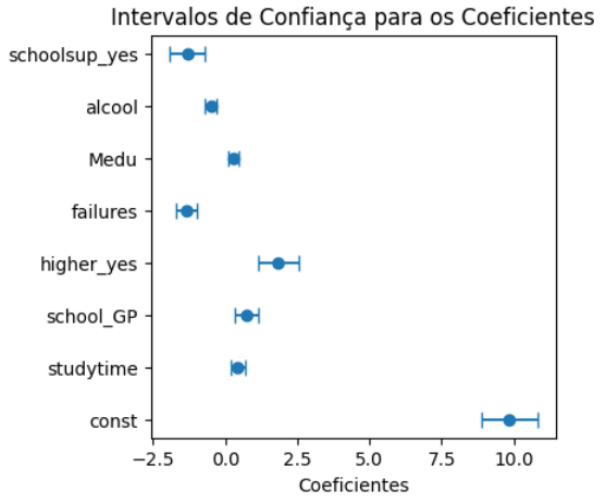


Figura 6 – Intervalos de confiança de 95% ($\beta \pm 1.96(std\_err)$)

Também, na tabela de regressão já são apresentados alguns valores de ajuste, como o AIC (-1111.1) e o R^2 que indica que nosso modelo explica apenas 33% da variância dos dados, o que é um valor baixo, mas pode ser justificado pelo fato de que provavelmente existem muitas outras variáveis que podem explicar a nota final do estudantes além das presentes na base de dados. Além disso, calculando o RMSE (avaliando as predições do modelo nos dados de teste), encontra-se o valor de 2.55, o que é um valor razoavelmente alto considerando o intervalo de 0 a 20 das notas, mas permite que tenhamos uma base plausível.

Por fim, além de verificar se esse modelo faz sentido do ponto de vista dos ajustes, devemos verificar se ele realmente atende aos pressupostos de um modelo linear, o que realmente ocorre, como mostram os resultados seguintes:

- A média dos resíduos é aproximandamente zero ($6.839e^{-15}$).
- Os dados são homocedásticos e os resíduos que descrevem a diferença entre o valor observado e o predito da variável dependente são normalmente distribuídos (p-value do teste de Shapiro-Wilk resultou em 0.000167).

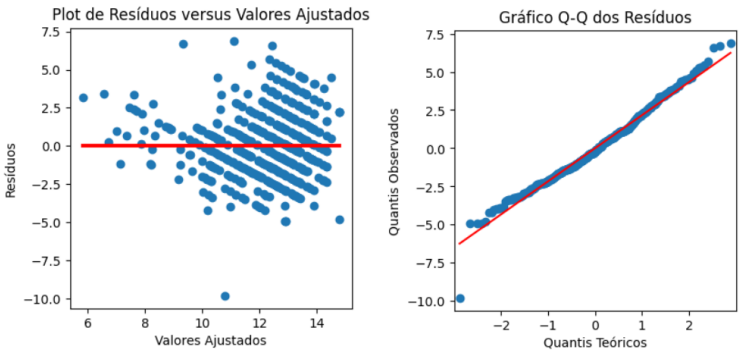


Figura 7 – Homocedacidade dos resíduos e normalidade dos dados

4 Conclusão

4.1 O que aprendemos?

Em suma, ao realizar a análise exploratória, identificamos correlações significativas entre a nota final (variável dependente) e várias variáveis independentes, como tempo de estudo semanal, escola do aluno, desejo de ingressar no ensino superior, número de reprovações anteriores, escolaridade da mãe, consumo de álcool e suporte escolar extra. Essas descobertas nos ajudaram a selecionar as variáveis mais relevantes para a construção dos modelos preditivos.

Ao ajustar o modelo de regressão linear multivariada, encontramos coeficientes estimados significativos para cada uma das variáveis selecionadas. Isso nos permitiu construir uma equação de previsão que incluiu essas variáveis e seus respectivos coeficientes. Além do intercepto 9.48, a análise dos coeficientes revelou que alunos que estudam mais tempo, estudam na escola Gabriel Pereira, pretendem entrar para universidade e possuem mães que tiveram um nível mais alto de ensino, desempenham melhor na avaliação acadêmica final, enquanto que reprovações, consumo de álcool e suporte escolar extra tendem a trazer relações negativas.

Contudo, apesar de conseguirmos atingir um dos objetivos e entender melhor os fatores que influenciam a performance dos alunos, sendo os coeficientes dos modelos preditivos significativos, análises de bondade do ajusteas e as predições alcançadas não são foram tão boas, com um RMSE e R^2 razoavelmente altos.

Essa baixa precisão é explicada por não considerarmos um dos principais fatores que são as notas antigas e que certamente trariam um resultado melhor. Em poucos testes realizados foi possível chegar a valores de, por exemplo, R^2 e RMSE de aproximadamente 0.9.

4.2 Limitações

Como já citado, uma das principais limitações abordadas nesse trabalho referiu-se a não utilização das variáveis de notas anteriores para prever a nota final dos alunos, o que, como citado, resultaria em predições mais ajustadas. Entretanto, além disso, é necessário considerar que essa baixa precisão também decorre devido a outros aspectos, especialmente pela complexidade do sistema educacional.

Isso é, existem inúmeros outros fatores que temos que considerar para a previsão do desempenho em um exame final, como a ansiedade do aluno ou seu nível de conhecimento sobre o assunto especificamente avaliado ou, até mesmo, o componente sorte. Ainda assim, nosso modelo fornece bons *insights* sobre a chance de sucesso ou fracasso dos alunos em relação à diferentes variáveis que identificamos.

Também, a metodologia abordada é bastante específica, e como abordado no próximo tópico, outro métodos devem ser averiguados para que possam ser tiradas conclusões ainda mais significativas.

4.3 Trabalhos futuros

Em pesquisas futuras, o conjunto de dados pode ser ampliado, buscando mais dados sobre o desempenho do aluno de mais escolas e mais algumas variáveis, como características socioeconômicas dos alunos, motivação e habilidades individuais, a fim de obter um modelo mais abrangente e preciso.

Além disso, como a metodologia empregada de prever a real nota do aluno é bastante complexa, uma possível alternativa seria aplicar a regressão logística (6) para prever a aprovação ou não dos alunos, onde notas acima de 10 (4) seriam classificadas como 1 e abaixo como 0, por exemplo. Isso traria mais simplicidade e um modelo satisfatório. Ou, ainda, recorrer à utilização de abordagens mais avançadas, como modelo com perspectivas bayesianas (especialmente agora que temos mais conhecimento sobre os dados), não lineares e de aprendizado de máquina, também podem ser exploradas para melhorar ainda mais as previsões.

Isso significa que, apesar de alcançar percepções relevantes, existem muitas direções a serem seguidas e investigadas, afim de trazer resultados ainda melhores.

Referências

1 Manoel Messias Gomes. Fatores que facilitam e dificultam a aprendizagem. Revista Educação Pública. 2018. Disponível em:https://educacaopublica.cecierj.edu.br/artigos/18/14/fatores-que-facilitam-e-dificultam-a-aprendizagem 1

2 P. Cortez e A. Silva. Student Performance. UC Irvine Machine Learning Repository. Disponível em: https://archive.ics.uci.edu/dataset/320/student+performance 1

3 Gabriela Entringe. Introdução à Regressão Linear Bayesiana. 2019. Disponível em: https://medium.com/@gabrielaentringe/introdu%C3%A7%C3%A3o-%C3%A0-regress%C3%A3o-linear-bayesiana-d3ccc7bddfb0 3

4 Wikipedia. Nota escolar. Disponível em:https://pt.wikipedia.org/wiki/Nota_escolar 11

5 Selva Prabhakaran. Assumptions of Linear Regression. 2016-17. Disponível em: http://r-statistics.co/Assumptions-of-Linear-Regression.html 4

6 Lee Min Hua Mohamad Najmudin. Prediction of Student Performance. 2017. Disponível em: https://rpubs.com/mhlee/student-performance-prediction 11

7 Bruno Oliveira. Como interpretar uma Análise de Variância (ANOVA)?. 2019. Disponível em: https://statplace.com.br/blog/como-interpretar-analise-de-variancia-anova/ 3

8 Francisco Oliveira. PORTUGAL É O PAÍS MENOS EDUCADO DA EUROPA E ESTE ANO REPROVARAM 15.000 ALUNOS NO 12º. Inspiring Future. 2020. Disponível em: https://inspiring.future.pt/stories/portugal-e-o-pais-menos-educado-da-europa-e-este-ano-reprovaram-15-000-alunos-no-12 1