Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

João Paulo da Costa Wolff

Análise de Fatores Relacionados na Performance Acadêmica dos Alunos: Um Estudo de Caso com Regressão Linear Simples e Múltipla

Curitiba
2024

João Paulo da Costa Wolff

Análise de Fatores Relacionados na Performance Acadêmica dos Alunos: Um Estudo de Caso com Regressão Linear Simples e Múltipla

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. José Luiz Padilha da Silva

Curitiba
2024

Análise de Fatores Relacionados na Performance Acadêmica dos Alunos: Um Estudo de Caso com Regressão Linear Simples e Múltipla

Analysis of Factors Related to Student Academic Performance: A Case Study with Simple and Multiple Linear Regression

João Paulo da Costa Wolff[1], José Luiz Padilha da Silva[2]

[1]Aluno do programa de Especialização em Data Science & Big Data, joao.pcw@gmail.com
[2]Professor do Departamento de Estatística - DEST/UFPR, jlpadilha@ufpr.br

Este trabalho busca investigar se fatores socioeconômicos e comportamentais estão relacionados ao desempenho acadêmico de alunos de uma universidade sul-africana situada na cidade de *Stellenbosch*. A pesquisa contou inicialmente com uma população de 406 indivíduos, distribuídos em 15 variáveis independentes e uma variável dependente, que é a Média Acadêmica do Ano de 2023. No entanto, após a exclusão de 73 estudantes que não estavam matriculados no ensino superior em 2023, a amostra final consistiu em 333 indivíduos. Para investigar a relação entre as variáveis independentes e a variável dependente, foram aplicados modelos de regressão linear simples e múltipla. Na primeira etapa, a regressão linear simples foi utilizada para avaliar o impacto isolado de cada variável independente sobre a média acadêmica. Esse procedimento permitiu identificar quais variáveis possuíam uma relação significativa com o desempenho acadêmico individualmente. Posteriormente, foi empregado o modelo de regressão linear múltipla para avaliar o efeito conjunto das variáveis independentes. Essa abordagem permitiu controlar a influência de cada variável enquanto se analisava o impacto das outras, oferecendo uma visão mais abrangente das interações complexas entre os diferentes fatores. A análise de variância foi utilizada para avaliar o efeito das variáveis independentes no modelo múltiplo. Essa análise ajudou na seleção das variáveis significativas que compõem o modelo final, permitindo a construção de um modelo robusto para explicar as variações na média acadêmica dos estudantes. Apesar dos indícios, o estudo explica pouco quais fatores socioeconômicos e comportamentais são relevantes ou não para o desempenho do aluno, sugerindo a necessidade de novos estudos.
Palavras-chave: Performance, Fatores Socioeconômicos, Fatores Comportamentais, Regressão Linear Simples, Regressão Linear Múltipla

This study seeks to investigate whether socioeconomic and behavioral factors are related to the academic performance of students at a South African university located in the city of Stellenbosch. Initially, the research involved a population of 406 individuals, distributed across 15 independent variables and one dependent variable, which is the Academic Average of the Year 2023. However, after excluding 73 students who were not enrolled in higher education in 2023, the final sample consisted of 333 individuals. To investigate the relationship between the independent variables and the dependent variable, simple and multiple linear regression models were applied. In the first stage, simple linear regression was used to evaluate the isolated impact of each independent variable on the academic average. This procedure allowed for identifying which variables had a significant relationship with academic performance individually. Subsequently, a multiple linear regression model was employed to evaluate the joint effect of the independent variables. This approach allowed for controlling the influence of each variable while analyzing the impact of others, providing a more comprehensive view of the complex interactions among different factors. Analysis of variance was used to assess the effect of the independent variables in the multiple model. This analysis helped select the significant variables that make up the final model, allowing the construction of a robust model to explain variations in the students' academic averages. Despite the indications, the study provides little explanation of which socioeconomic and behavioral factors are relevant or not to student performance, suggesting the need for further studies.
Keywords: Performance; Socioeconomic Factors; Behavioral Factors; Simple Linear Regression; Multiple Linear Regression

1. Introdução

Entende-se que desempenho acadêmico é uma preocupação central tanto para educadores quanto para estudantes, uma vez que é um dos principais indicadores de sucesso no ambiente educacional. A compreensão dos fatores que influenciam o desempenho acadêmico é fundamental para a elaboração de estratégias que visem melhorar a qualidade do ensino e a aprendizagem dos alunos. Além disso, tal qualidade do ensino superior é uma questão complexa, e discutir os métodos para avaliar a aquisição de conhecimentos e habilidades dos alunos ao longo da graduação é extremamente importante, tanto no contexto brasileiro quanto mundial (Melguizo; Wainer, 2016, p. 381-401).

Em um artigo escrito por Silva e Junior (2016, p. 424-426) no qual são analisados 6 estudos, viu-se que o vestibular, por exemplo, distingue os alunos mais bem preparados daqueles menos preparados para ingressar no ensino superior. No entanto, durante o curso, outros fatores como a família, o emprego, a formação básica e a identificação com o curso, entre outros, influenciam o desempenho desses estudantes. Ademais, características domiciliares favoráveis e a conclusão do ensino superior pelos pais estão diretamente associadas ao desenvolvimento acadêmico. Além disso, estudantes que não trabalham obtêm melhores resultados do que aqueles que trabalham. Por fim, os alunos cujos pais estão empregados apresentam um rendimento acadêmico mais satisfatório em comparação aos que têm pais desempregados.

Diante desse cenário, resolveu-se entender mais da relação desempenho acadêmico versus fatores socioeconômicos e comportamentais. Para isso utilizou-se de uma base de dados online que hospeda competições de ciência de dados, projetos de código aberto e conjuntos de dados de alta qualidade (Kaggle, 2024). A base foi disponibilizada por alunos do departamento de Estatística e Ciência Atuarial da universidade de *Stellenbosch*, situada na cidade de mesmo nome na África do Sul. Por fim, utilizou-se o *Rstudio*, um *software* livre de ambiente de desenvolvimento integrado para R para analisar os dados e aplicar os modelos.

2. Discussão

Na população da pesquisa em questão identificam-se 406 indivíduos distribuídos em 15 variáveis independentes e uma variável dependente sendo ela a média acadêmica do ano de 2023. Para melhor compreensão, optou-se por traduzir ambas as variáveis da língua inglesa para língua portuguesa, tomando cuidado para manter a ideia original da base de dados. Optou-se também por traduzir as características das variáveis, conforme ilustrado na Figura 1. Entretanto e para fins do estudo, na base dados encontram-se 73 estudantes que não estavam matriculados no ensino superior no ano de 2023, ou seja, não elegíveis. Portanto a base passou de 406 indivíduos para 333.

Variáveis Independentes	Características das Variáveis
Sexo	Feminino (48), Masculino (52)
Média no Último Ano do Ensino Médio	Média no Último Ano do Ensino Médio
Ano Anterior	1º Ano (38), 2º Ano (46), 3º Ano (12), 4º Ano (2), Pós-graduação (2)
Departamento	Ciências Agrárias (6), Artes e Ciências Sociais (12), Ciências Econômicas e de Gestão (54), Educação (2), Engenharia (8), Direito (3), Medicina e Serviços de Saúde(3), Ciências (13)
Tipo de Acomodação	Acomodação na universidade (14), Acomodação particular/ficar com família/amigos (86)
Ajuda de Custo Mensal	R** 4.001,00- R 5.000,00 (42), R 5.001,00 - R 6.000,00 (28), R 6.001,00 - R 7.000,00 (14), R 7.001,00 - R 8.000,00 (8), R 8.000,00+ (8)
Bolsa de Estudos	No (89), Yes (NSFAS, etc...) (11)
Horas Extras de Estudo Semanais	0 (5), 1-2 (21), 3-5 (23), 5-8 (21), 8+ (30)
Confraternizações Semanais	0 (4), 1 (27), 2 (22), 3 (15), 4+ (4), Apenas aos finais de semana (28)
Consumo por Confraternização	0 (7), 1-2 (22), 3-5 (23), 5-8 (28), 8+ (21)
Faltas Semanais por Motivos de Consumo de Álcool	0 (51), 1 (20), 2 (17), 3 (6), 4+ (7)
Disciplinas Reprovadas	0 (58), 1 (17), 2 (9), 3 (7), 4+ (8)
Relacionamento Afetivo	Sim (58), Não (42)
Aprovação dos Pais com Consumo de Álcool	Sim (14), Não (86)
Relacionamento com os Pais	Muito próximo (66), Próximo (24), Distante (1), Regular (10)

Figura 1: Variáveis e Características

Considerando que o foco principal do estudo é a performance diante de fatores socioeconômicos e comportamentais, ou seja, a variável dependente pode assumir valores que vão de 0 a 100%, vê-se que a média das notas é de 66,28%, enquanto a mediana é exatamente 65% e por fim o desvio padrão fica em 9,15%. Além disso, viu-se que a nota mais baixa é de 30% ao passo que a maior nota é de 95,22%.

Ponderando a questão do sexo, nota-se que a média para os homens fica em 65,09% à medida que para as mulheres é levemente superior em 67,64%. Esse equilíbrio é também visto na população, ou seja, 52% dos estudantes são homens enquanto 48 são mulheres.

3. Materiais e métodos

Para investigar os fatores relacionados ao desempenho acadêmico, estabeleceu-se uma análise bivariada utilizando regressão linear simples. Em seguida, conduzimos uma análise multivariada por meio da regressão linear múltipla, considerando significativas as associações quando $p<0,05$. A análise de variância foi

	Regressão Linear Simples	
Características das Variáveis	**Estimativa**	**Valor-p**
Sexo		
Feminino	-	-
Masculino	-2,55	0,013*
Média no Último Ano do Ensino Médio	-	-
Média no Último Ano do Ensino Médio	0,41	<0,001
Ajuda de Custo Mensal		
Ajuda de Custo R 4001 - R 5000	-	-
Ajuda de Custo R 6001 - R 7000	3,48	0,033*
Disciplinas Reprovadas		
Disciplinas Reprovadas 0	-	-
Disciplinas Reprovadas 1	-7,17	<0,001
Disciplinas Reprovadas 2	-6,52	<0,001
Disciplinas Reprovadas 3	-11,39	<0,001
Disciplinas Reprovadas 4+	-10,37	<0,001
Relacionamento com os Pais		
Relacionamento com os Pais Muito próximo	-	-
Relacionamento com os Pais Distante	21,97	<0,001

Figura 2: Regressão Linear Simples

	Regressão Linear Múltipla	
Características das Variáveis	**Estimativa**	**Valor-p**
Sexo		
Feminino	-	-
Masculino	-1,77	<0,001
Média no Último Ano do Ensino Médio		
Média no Último Ano do Ensino Médio	0,36	<0,001
Ajuda de Custo Mensal		
Ajuda de Custo R 4001 - R 5000	-	-
Ajuda de Custo R 5001 - R 6000	-0,09	0,942
Ajuda de Custo R 6001 - R 7000	3,56	0,022*
Ajuda de Custo R 7001 - R 8000	1,18	0.542
Ajuda de Custo R 8000+	1,68	0.387
Disciplinas Reprovadas		
Disciplinas Reprovadas 0	-	-
Disciplinas Reprovadas 1	-7,03	<0,001
Disciplinas Reprovadas 2	-5,72	<0,001
Disciplinas Reprovadas 3	-9,33	<0,001
Disciplinas Reprovadas 4+	-11,90	<0,001
Relacionamento com os Pais		
Relacionamento com os Pais Muito próximo	-	-
Relacionamento com os Pais Distante	23,11	<0,001

Figura 3: Regressão Linear Múltipla

utilizada para a seleção das variáveis significativas no modelo final (James et al., 2023, p. 59-121).

4. Resultados

Em um primeiro momento, após aplicar a análise bivariada utilizando regressão linear simples, poucos fatores sobressaíram-se em relação aos demais. Identifica-se na Figura 2 que os fatores mais significativos considerando o valor p são a média no último ano do ensino médio, disciplinas reprovadas em todas as faixas e relacionamento distante com os pais. Menos significativos são o sexo masculino em detrimento ao sexo feminino e aqueles alunos cuja ajuda de custo fica no intervalo entre R 6.001,00 e R 7.000,00.

Pela ótica do coeficiente estimado e com relação aos valores p menos significativos, percebe-se que com relação a variável ajuda de custo mensal, vê-se que o coeficiente estimado é de 3,48. Esse valor indica que, em média, os alunos que recebem uma ajuda de custo mensal entre R 6001 e R 7000 têm notas 3,48 pontos mais altas do que os alunos que recebem uma ajuda de custo mensal entre R 4001 e R 5000 (grupo de referência).

Em seguida, após aplicar a regressão linear múltipla, a variável independente sexo passou a desempenhar um papel de protagonismo em relação as demais variáveis. Além disso, a variável disciplinas reprovadas, em todas as faixas, voltou a demonstrar importância semelhante à variável sexo. Ademais, nota-se na Figura 3 que estudantes com relacionamento distante mantiveram-se relevantes no modelo. Por fim, alunos cuja ajuda de custo fica no intervalo entre R 6.001,00 e R 7.000,00 repetiram o mesmo desempenho do modelo anterior, isto é, moderadamente significativo. Frente à regressão linear múltipla, em média as notas dos homens diminuíram 1,77 pontos em relação às das mulheres, enquanto o coeficiente estimado para ajuda de custo apresentou pouca alteração.

Por fim, ao comparar os R^2 entre os modelos, observamos que as variáveis independentes explicam 40,63% da variação na performance dos alunos no modelo múltiplo, enquanto no modelo final o R^2 é de 34,48%. Pode-se atribuir essa diferença a outros fatores não investigados ou não disponíveis na base.

Ao observar as variáveis significativas e para melhor intepretação dos resultados, sugere-se que é preciso rever a população de maneira mais detalhada. Olhemos a variável relacionamento com os pais, por exemplo. Dos 333 alunos, apenas 2 marcaram a opção relacionamento distante e ainda assim essa variável foi uma das mais relevantes. Ademais e com relação a variável disciplinas reprovadas entende-se que se a nota do aluno é baixa existe um indício de que ele tenha reprovado no ano anterior. No entanto, próximo de 60% dos alunos não reprovaram no ano anterior. Por outro lado, nota-se uma normalidade ao verificar a variável ajuda de custo, mas especificamente na faixa entre R 6.001,00 e R 7.000,00, visto que 14% dos alunos encontram-se esse faixa. Por último temos a média no último ano do ensino médio. Novamente vê-se uma relação natural, isto é, tende-se a pensar que o desempenho no ensino superior seja parecido com a performance no ensino médio.

5. Conclusão

Apesar dos indícios, o estudo pouco explica quais fatores socioeconômicos e comportamentais são relevantes ou não para a performance do aluno, além disso, sugere-se um estudo orientado para o curso, considerando que os cursos podem ter dificuldades diferentes e um perfil de aluno específico. Outro ponto a considerar é o modelo da universidade ou faculdade em que o aluno está matriculado. Será que o desempenho dos estudantes em instituições particulares é superior ao das universidades públicas. No Brasil, o Exame Nacional de Desempenho dos Estudantes (ENADE) (2004, p.3) é um importante instrumento de avaliação educacional. Além da prova, os participantes respondem a um questionário socioeconômico e comportamental. Esse questionário abrange uma variedade de informações sobre características pessoais, socioeconômicas e acadêmicas dos estudantes. Esses dados são essenciais para análises complementares que não apenas ajudam na interpretação dos resultados do exame, mas também na formulação de políticas públicas educacionais. Essa abordagem integrada fornece uma visão abrangente do desempenho dos estudantes e das condições que podem influenciar seu sucesso acadêmico, promovendo assim uma educação mais equitativa e eficaz.

6. Agradecimentos

Gostaria de expressar meus agradecimentos a todos os docentes do curso de especialização em Data Science e Big Data que contribuíram para a realização do meu Trabalho, em especial ao meu orientador José Luiz Padilha e o avaliador Cesar Augusto Taconeli . Agradecer também, meus amigos e família que me ajudaram e incentivaram a continuidade dos meus estudos, em especial minha esposa Alessandra Campos Arduini.

Referências

[1] SILVA, H.; JÚNIOR, A. Fatores determinantes do desempenho acadêmico no ensino superior: estado da arte - Revista Plurais – Virtual, Anápolis - Go, vol.6, n. 2 – jul./dez. 2016.

[2] MELGUIZO, T.; WAINER, J. Toward a set of measures of student learning outcomes in higher education: evidence from Brazil. Higher Education, Amsterdam, v. 72, n. 3, p. 381-401, Sept. 2016.

[3] BRASIL. Lei nº 10.861, de 14 de abril de 2004. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES e dá outras providências. Diário Oficial da União, Brasília, 15 abr. 2004. Seção 1, p. 3.

[4] JAMES, Gareth et al. An Introduction to Statistical Learning: with applications in r. 2. ed. [S. L.]: Sprinter, 2023. Disponível em: https://www.statlearning.com/. Acesso em: 01 mar. 2024.

[5] Kaggle (2024) [datarepository] Effects of Alcohol on Student Performance. Disponível em: <https://www.kaggle.com/datasets/joshuanaude/effects-of-alcohol-on-student-performance>. Acesso em: 01 de mar. 2024.