ESTIMATIVA DE EVASÃO ESTUDANTIL EM UM CURSO DE CIÊNCIA DA COMPUTAÇÃO

SANTOS, Vinicius Teixeira[1]
BAÍA, Joás Weslei[2]

[1]Discente do curso de em Ciência da Computação do UNIFAGOC
[2]Mestre em Ciência da Computação pela UFV e Professor do curso de Ciência da Computação do UNIFAGOC

joasweslei@gmail.com

RESUMO

Este trabalho apresenta a aplicação de técnicas estatísticas e de aprendizado de máquina para a estimativa do rendimento acadêmico e a previsão de evasão em um curso de Ciência da Computação. A regressão linear foi utilizada para modelar o aproveitamento futuro dos estudantes com base no desempenho prévio, sendo validada por meio do Teste t de Student, que demonstrou boa aderência entre os valores previstos e os valores reais. Já a regressão logística foi empregada para estimar a probabilidade de evasão discente, alcançando uma acurácia de 76%. A avaliação da regressão logística incluiu a análise da matriz de confusão, que revelou maior eficiência na identificação de estudantes não evadidos e apontou limitações no reconhecimento de casos de evasão, atribuídas ao desequilíbrio das classes no conjunto de dados. Os resultados destacam o potencial dessas abordagens para subsidiar decisões estratégicas, como intervenções direcionadas e políticas de retenção, contribuindo para a redução das taxas de evasão e a melhoria do desempenho acadêmico.

Palavras-chave: Rendimento Acadêmico. Flipped Classroom. Aprendizagem de Máquina.

1 INTRODUÇÃO

O conceito de rendimento acadêmico refere-se ao desempenho do aluno em seu ambiente educacional, comumente medido por notas, pontuações em testes e avaliações de professores. É um indicador importante da competência e do progresso do estudante ao longo de sua jornada educacional. No entanto, é importante lembrar que o rendimento acadêmico não é o único indicador do valor de um aluno, e fatores como a motivação, a paixão pelo aprendizado e o desenvolvimento de habilidades interpessoais e/ou profissionais também desempenham um papel crucial na formação de um indivíduo (Rodrigues, 2013).

Essas questões educacionais, no Brasil, constantemente são temas de debates em diversos meios, sejam eles acadêmicos, sociais, filosóficos e até mesmo informais. Muitas questões estão relacionadas aos desafios encontrados quando se refere à educação, e muitas pesquisas visam avaliar a qualidade do ensino brasileiro e identificar os seus pontos de fragilidade, para, a partir disso, tomar decisões que possam auxiliar no melhor desenvolvimento educacional do País (Menolli; Neto, 2021, p. 5). Por exemplo, quais fatores afetam o rendimento acadêmico de um estudante e

como os métodos de estudo podem interferir nessa questão? Quais são os possíveis motivos de reprovação e/ou evasão do ensino superior?

O processo avaliativo dos cursos de Ciência da Computação revela que os estudantes enfrentam várias dificuldades durante a aprendizagem da programação. Segundo Castro (2003) e Aureliano *et al*. (2020), essas dificuldades ocorrem nas disciplinas que abordam Programação Introdutória, Algoritmos e Estruturas de Dados e reduzem a motivação dos alunos. Esse cenário os leva, muitas vezes, a desistirem do processo de ensino-aprendizagem.

Nesse contexto, este trabalho investigou a estimativa de rendimento acadêmico e o seu impacto na evasão no curso de Ciência da Computação, no qual a metodologia ativa *Flipped Classroom* é adotada. Neste texto, serão discutidas as seguintes estratégias:

- O método de estimativa dos resultados futuros de uma disciplina;
- O método de teste desse modelo de previsão de resultados;
- O método de estimativa de evasão discente;
- A comparação dos resultados da previsão futura com os dados consolidados.

Na próxima seção é apresentado o método de construção de um estimador de resultado aplicado à realidade educacional.

2 APRENDIZAGEM ARTIFICIAL

Segundo Russell e Norvig (2013, p. 607), a aprendizagem artificial supervisionada é baseada na seguinte tarefa: dado um conjunto de treinamento de N pares de exemplos de entrada e saída $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$, em que cada valor de y foi gerado por uma função desconhecida $y = f(x)$. Então, o resultado é uma função h que se aproxima da função verdadeira $y = f(x)$.

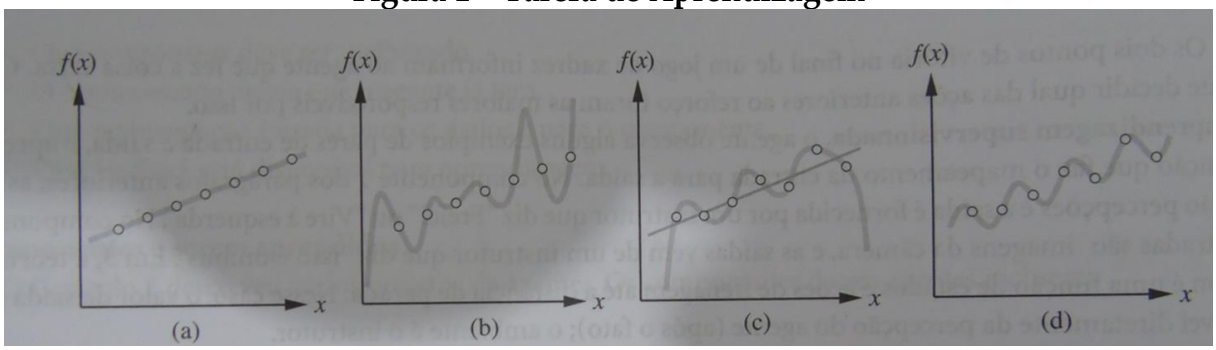
Por exemplo, seja y_1 o aproveitamento acadêmico de um discente na disciplina D_2 e x_1 o aproveitamento na disciplina com pré-requisito, D_1. A função h estimará o aproveitamento futuro dos alunos que farão a disciplina D_2, tendo cursado a disciplina D_1. Logo, usaremos N pares de exemplos de entrada e saída que permitirão a obtenção da referida função h. Ela será o modelo matemático de estimativa do aproveitamento discente na disciplina D_2.

A função h é uma hipótese, e a tarefa de aprendizagem é encontrar os parâmetros dessa função que façam com que ela se aproxime da função verdadeira. Para aferir a precisão de uma hipótese, é usado um conjunto de exemplos de testes que são *distintos do conjunto de treinamento*. Assim, uma hipótese generaliza bem se prevê corretamente o valor de y para novos exemplos. Por exemplo, para novos alunos que cursarão D_2.

É importante salientar que essa estimativa pode ser obtida antes que a disciplina D_2 seja ofertada, permitindo que ações sejam tomadas para evitar a futura evasão desses alunos. Por fim, a aplicação da aprendizagem supervisionada ao aproveitamento discente será chamada de *regressão* quando x e y forem números. Quando for uma variável qualitativa, assumirá a forma de uma *classificação* (Russel; Norvig, 2013, p. 607).

A Figura 1 ilustra a tarefa de aprendizagem supervisionada mencionada anteriormente. Em (a) são ilustrados alguns dados com ajuste exato por linha reta. Essa linha foi considerada consistente porque concordou com todos os *N* pares de treinamento. Essa situação seria a melhor possível, pois a função *h* consegue atender a todos pares (x, y).

Figura 1 – Tarefa de Aprendizagem



(a) Exemplo de pares (x, y) e uma hipótese linear consistente. (b) Hipótese de polinômio de grau 7 consistente para o mesmo conjunto de dados. (c) Conjunto de dados diferente que admite um ajuste de polinômio de grau 6 exato ou um ajuste linear aproximado. (d) Um simples ajuste senoidal exato para o mesmo conjunto de dados.

Fonte: Russell; Norvig (2013, p. 608).

Em (b) é apresentado um exemplo de pares de dados que formam um polinômio de grau alto, o que dificulta a tarefa de aprendizagem. Em (c) é um exemplo de pares de dados que não são explicados por uma linha reta. Por fim, em (d) foi apresentada uma função *h* senoidal que se ajusta aos mesmos pares de dados em (c).

Um bom modelo de aprendizagem de máquina deve considerar uma estratégia para determinar a melhor função que descreve o conjunto de dados analisado. Na regressão utilizada neste trabalho, foi considerada a reta que minimiza os quadrados dos erros. Assim, foi aplicado o método dos mínimos quadrados (Pianezzer, 2020, p. 68).

3 REGRESSÃO LINEAR

A regressão linear é uma técnica utilizada na análise estatística e no campo da aprendizagem de máquina. Ela desempenha um papel crucial na modelagem de relacionamentos entre variáveis, permitindo a previsão de valores com base em dados observados (Bingham; Fry, 2010, p. 9).

Trata-se de um método estatístico para encontrar uma relação linear entre duas ou mais variáveis, frequentemente utilizada para entender como uma variável independente (ou preditora) afeta uma variável dependente (ou resposta). A forma mais singela de regressão linear é a regressão linear simples, que lida com apenas duas variáveis: uma variável independente X e uma variável dependente Y (Osborne, 2016, p. 10). A equação da regressão linear simples pode ser representada da seguinte forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Onde:

- Y é a variável dependente.
- X é a variável independente.
- β0 é o coeficiente de interceptação.
- β1 é o coeficiente de inclinação, que representa como Y muda em resposta a uma mudança em X.
- ε é o erro, que representa a variação não explicada em Y.

Quando a regressão é aplicada na aprendizagem de máquina, como foi feito neste trabalho, é utilizado um conjunto de treinamento de *N* pares de exemplos de entrada e saída $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$, em que cada valor de *y* foi gerado por uma função desconhecida $y = f(x)$. Então, o resultado é uma função *h,* que se aproxima da função verdadeira $Y = f(x)$. Portanto, a função *h* é a nossa hipótese de estimativa de rendimento acadêmico (Russel; Notvig, 2013, p. 607).

O principal objetivo da regressão linear é estimar os valores dos coeficientes β_0 e β_1 para melhor ajustar a linha aos dados. Isso é feito minimizando a soma dos quadrados dos resíduos, ou seja, a diferença entre os valores previstos pela linha de regressão e os valores reais dos dados (Osborne, 2016, p. 11).

A regressão linear é uma base importante na aprendizagem de máquina. Modelos mais avançados, como regressão linear múltipla e regressão logística, são amplamente utilizados em problemas de classificação e previsão. Ela é a base para compreender conceitos mais complexos, como regularização e seleção de recursos em modelos de aprendizado de máquina (Gomes *et al*., 2017).

A avaliação de um modelo de regressão linear é crucial para determinar a sua adequação, precisão e utilidade na análise futura. Existem várias métricas e métodos que podem ser utilizados para avaliar a qualidade de um modelo de regressão linear. Nesta pesquisa, foi utilizado o teste t (Zimmerman, 1987).

4 REGRESSÃO LOGÍSTICA

A regressão logística é uma técnica estatística utilizada para modelar a probabilidade de ocorrência de eventos binários (dicotômicos) com base em variáveis explanatórias (Press; Wilson, 1978). É oportuna para situações em que a variável dependente assume dois valores distintos, como **evadido** ou **não evadido**. Diferentemente da regressão linear, que prevê valores contínuos, a regressão logística estima a probabilidade de um evento ocorrer, com resultados variando entre 0 e 1.

A regressão logística se baseia na função logística, uma curva sigmoidal que transforma uma combinação linear das variáveis independentes em probabilidades dentro do intervalo de 0 a 1. A fórmula fundamental do modelo de regressão logística é:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$$

Os coeficientes (β_0, β_1,…,β_p) na regressão logística são estimados pelo método da máxima verossimilhança, que busca encontrar os valores dos parâmetros que maximizam a probabilidade de os dados observados serem gerados pelo modelo. Segundo Hosmer e Lemeshow (2000), o método da máxima verossimilhança é

utilizado para estimar os coeficientes de modelos de regressão logística, permitindo que a função de probabilidade seja maximizada com base nos dados disponíveis. Além disso, Menard (2002) destaca que a máxima verossimilhança é essencial para ajustar modelos que tratam de variáveis dependentes binárias.

A regressão logística tem aderência à análise de evasão discente, pois permite identificar padrões que contribuem para o abandono. No contexto educacional, variáveis como características socioeconômicas, notas, frequência e engajamento podem ser usadas para modelar a probabilidade de evasão.

5 TRABALHOS RELACIONADOS

A comunidade científica tem trabalhado na investigação dos desafios da aprendizagem. Nesse sentido, esta seção apresenta algumas pesquisas realizadas em relação ao rendimento acadêmico dos discentes das instituições de ensino superior.

Souza Miranda *et al.* (2018) analisaram as diferenças entre o rendimento escolar de estudantes beneficiados pelo programa de bolsa em relação aos não beneficiados. Inicialmente, os autores apresentaram o cenário da educação brasileira, as dificuldades das pessoas de baixa renda no acesso à educação superior e como as ações afirmativas visam sanar esse problema. Um elemento importante nessa pesquisa foi traçar o perfil dos alunos ingressantes no curso de Ciências Contábeis.

Foram realizados testes estatísticos das médias de 11 disciplinas do curso de Ciências Contábeis. Tomaram-se como referência as disciplinas cursadas em dois semestres e utilizaram-se quatro variáveis: médias dos alunos e classificação no vestibular, mescla desses a partir do agrupamento de classificação no vestibular, e a diferença em relação à quantidade de aprovações. Nesta última etapa, foi utilizado um teste com 95% de confiança, o teste de médias não paramétrico, de Mann-Ehitney (McKnight; Najab, 2010). Além disso, utilizou-se também o teste de Qui-quadrado.

Como resultado desta pesquisa, foi observado que não há diferença considerada significativa em relação ao rendimento acadêmico. Ou seja, o fato de o aluno vir de escola pública não afeta a sua capacidade de obter um bom rendimento na faculdade (Souza Miranda; Lima; Marinello, 2018, p. 201).

Meurer (2018) analisou o estilo de aprendizagem e o rendimento acadêmico na universidade. Diferentemente da pesquisa de Souza Miranda *et al*. (2018), esta observou o rendimento acadêmico sob a ótica do modelo de aprendizagem para observar o comportamento do rendimento. Tomaram como base o modelo de KOLB (1984), para o qual existem quatro estilos de aprendizagem: acomodado, assimilador, convergente e divergente.

Os discentes considerados nesta pesquisa foram os matriculados no curso de Ciências Contábeis de uma IES pública no ano de 2015 e os seus professores. Foi observado como o estilo de aprendizagem do discente-docente interfere no seu desempenho (Meurer, 2018, p. 28). Os dados desta pesquisa foram coletados por meio de questionários, compostos por dois blocos, um com características pessoais e outro com o inventário de Kolb (1984). Além disso, foram utilizados os CRAS e suas notas do ENEM.

A partir de uma análise estatística descritiva, concluíram que "o estilo que dedica maior tempo de estudo extraclasse é justamente aquele que possui

característica acentuada de análise lógica e detalha os conhecimentos teóricos com a aplicação prática" (Meurer, 2018, p. 38). Ressaltou-se a importância de se realizar esse estudo em outras regiões do país, instituições de ensino públicas e privadas, com o intento de agregar mais na linha de pesquisa. Essa constatação está diretamente relacionada com a metodologia ativa *Flipped Classroom.*

Silva (2013) examinou o desempenho acadêmico de 1000 alunos ingressantes no 1º semestre de 2010 até o 1º semestre do ano de 2012. As variáveis analisadas foram: matrícula, curso escolhido, rendimento no primeiro semestre, gênero, status civil, escola onde cursou o ensino médio (p. 23). A partir disso, foi realizada a análise descritiva dos dados, utilizando a análise de regressão com o modelo Probit (p. 28).

Como resultado, observou-se que o estado civil não foi um dado estatisticamente significante em relação aos demais. Além disso, foram apresentadas as proporções de falha ou sucesso a partir do gênero, da idade de conclusão do ensino médio e de ingresso no ensino superior, exercer ou não alguma atividade remunerada.

6 MÉTODO DE DESENVOLVIMENTO DO TRABALHO

O projeto desenvolvido é classificado como uma pesquisa de natureza aplicada, tendo em vista que propõe-se a aplicar uma tecnologia disponível para resolver um problema que atinge a comunidade acadêmica. Quanto ao método de desenvolvimento, é quantitativo. Nesse sentido, ele seguiu as seguintes atividades:

- Foram utilizadas duas disciplinas com registro de notas para compor os N pares de dados para a obtenção da função h capaz de estimar o aproveitamento futuro na disciplina D_2, a partir do comportamento na disciplina D_1. As disciplinas escolhidas foram *Algoritmos e Técnicas de Programação*, D_1, e *Estruturas de Dados*, D_2, do curso de Ciência da Computação de uma Instituição de Ensino Superior do Brasil. Pois, possuem pré-requisito e principalmente devido ao fato de a disciplina D_2 apresentar alta taxa de reprovação e contribuir com a evasão.
- A coleta e tratamento dos dados de aproveitamento nessas disciplinas seguiu as etapas do processo utilizado por Souza (2021, p. 3).
- A limpeza dos dados consistiu em descartar as notas ausentes em D_2 e D_1. Além disso, os valores do aproveitamento acadêmico foram *transformados em uma escala distinta da usada*, tendo em vista o sigilo dos dados utilizados.
- O modelo de aprendizagem de máquina empregado foi a regressão linear para identificar o aproveitamento futuro, pois as variáveis analisadas são quantitativas. Para obter a probabilidade de evasão, foi aplicada a regressão logística com dados de quatro disciplinas desse curso.
- Foi medida a capacidade de generalização da função h obtida na fase de treinamento do modelo. Foi utilizado o Teste *t de Student* (Zimmerman, 1987). Trata-se de um procedimento estatístico usado para determinar se há uma diferença significativa entre as médias de dois grupos de dados. Ele foi desenvolvido por William Sealy Gosset em 1908 e é utilizado em análises estatísticas, especialmente em situações em que há amostras pequenas. Ele é usado para comparar as médias de duas amostras para verificar se a diferença entre elas é estatisticamente significativa. Ele leva em consideração a variabilidade dos dados e o tamanho da amostra. Existem várias versões do teste t, incluindo o teste t de duas amostras independentes, o teste t

pareado (ou teste t de amostras emparelhadas), e o teste t de uma amostra, entre outros. Ele é baseado na distribuição t de *Student*, que é semelhante à distribuição normal, mas leva em consideração a incerteza introduzida pelo uso de amostras pequenas. O teste t calcula uma estatística t, que é uma medida da diferença entre as médias das amostras em relação à variabilidade dos dados. Com base nessa estatística e no número de graus de liberdade, pode-se determinar se a diferença entre as médias é estatisticamente significativa. Assim, as notas obtidas no oferecimento da disciplina foram comparadas com as estimadas pelo modelo de previsão.

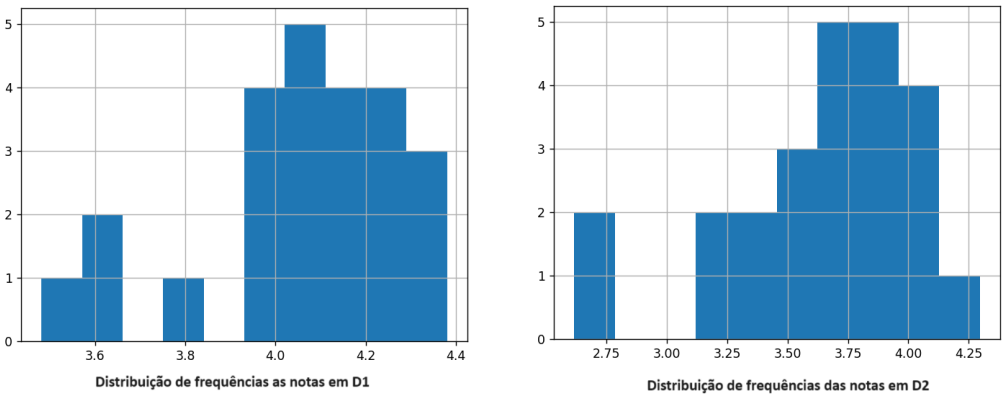
- Por fim, foram usadas as ferramentas da Estatística descritiva para aferir a aderência da hipótese *h* para fazer estimativas futuras do rendimento acadêmico em D_2 e evasão.

Com a execução dessas atividades, foi atingido o objetivo desta pesquisa: estimar o aproveitamento discente e sua probabilidade de evasão, a partir da aprendizagem de máquina.

7 RESULTADOS

A partir do método de desenvolvimento escolhido, os dados do aproveitamento acadêmico foram coletados e tratados. A Figura 2 ilustra o histograma que *resume* esses valores. No eixo x foram apresentadas as medidas do aproveitamento discente e, no eixo y, suas frequências.

Figura 2 - Histograma do aproveitamento acadêmico nas duas disciplinas



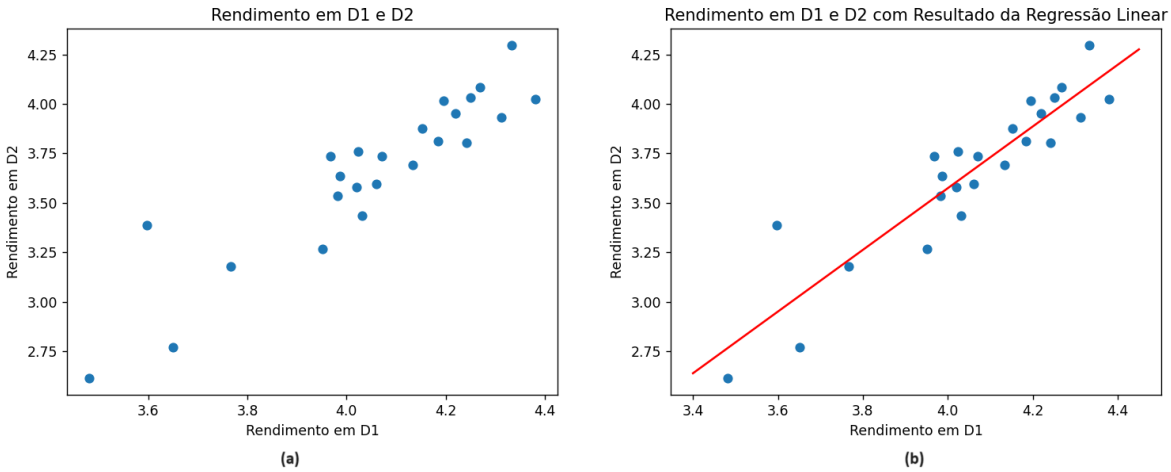
Fonte: dados da pesquisa.

É perceptível a diferença nas duas distribuições de frequências apresentadas na Figura 2. Em D_1, 67% dos resultados estão acima de *4.0 unidades de medida*. Em D_2, a variável dependente no modelo de previsão, apenas 20% atingiram esse valor. Observa-se, portanto, que a disciplina D_2 apresenta *menor desempenho acadêmico.*

Os discentes precisam aplicar os conhecimentos adquiridos em D_1 e absorver os inerentes a D_2. Portanto, D_2 possui um nível de dificuldade maior. Ela é a variável dependente, enquanto D_1 assume o papel de independente, porque, a partir dos rendimentos acadêmicos da disciplina D_1, serão estimados os aproveitamentos em D_2.

A Figura 3 apresenta os N pares de treinamento usados para obter a nossa hipótese de estimativa de rendimento. Além disso, é apresentado o resultado da função de previsão ajustada pelo processo de regressão linear.

Figura 3 - (a) N pares de aproveitamento usados na regressão linear. (b) a função h, nossa hipótese de estimativa para D_2



Fonte: dados da pesquisa.

Analisando o gráfico *(a)*, os eixos x e y são os rendimentos acadêmicos no intervalo de [0, 5]. Podemos perceber que há uma relação entre os rendimentos nas disciplinas D_1 e D_2. Aqueles que tiveram dificuldades na aprendizagem em D_1 também tiveram em D_2. Em *(b)*, foi apresentado o resultado do ajuste da função h que será usada para prever o rendimento acadêmico dos discentes no oferecimento futuro de D_2.

A Tabela 1 apresenta o resultado da regressão linear apresentada em (b), bem como os parâmetros que nos permitem analisar sua aderência para fazer estimativas de rendimento acadêmico. Portanto, nossa função h assume a seguinte forma: $\boldsymbol{h = 1{,}56x - 2{,}67}$. Ela é a hipótese de previsão do aproveitamento futuro na disciplina de Algoritmos e Estruturas de Dados a partir do aproveitamento futuro em Algoritmos e Técnicas de Programação.

Tabela 1 - Estatísticas da regressão apresentada na Figura 05 *(b)*

Estatística de regressão	
R múltiplo	0,91
R-Quadrado	0,84
R-quadrado ajustado	0,83
Erro padrão	0,07
Coeficiente angular	1,56
Intercept	-2,67

Fonte: dados da pesquisa.

O coeficiente de correlação R na regressão linear é uma medida fundamental que avalia a força e a direção da relação linear entre variáveis. O coeficiente de

correlação R varia de -1 a 1, indicando uma correlação positiva perfeita quando próximo a 1, uma correlação negativa perfeita quando próximo a -1 e nenhuma correlação linear quando próximo a 0. Esta medida desempenha um papel crucial na análise estatística, auxiliando na compreensão das relações entre variáveis em um modelo de regressão linear (Inman, 1994).

Pela análise da Tabela 1, observa-se que o valor de R se distanciou do valor zero. Pode-se perceber pela Figura 3 *(b)* que a linha foi ajustada aos dados de treinamento e as diferenças foram minimizadas.

Uma tarefa importante realizada foi o teste da hipótese ***h = 1,56x – 2,67*** para aferir se realmente ela é capaz de prever o comportamento futuro. Nesse sentido, os rendimentos acadêmicos obtidos noutro oferecimento de D_2 foram comparados com os valores estimados por *h*, conforme apresentado na Tabela 2 e pelos histogramas da Figura 4.

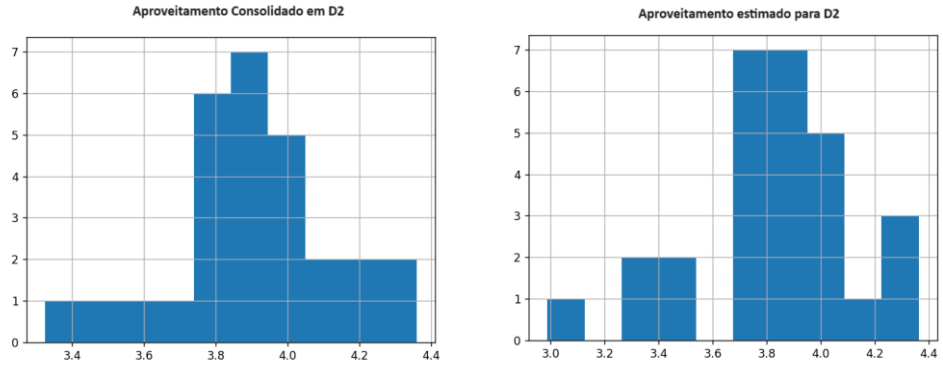
Tabela 2 - Estatística comparativa entre valores consolidados e estimados

ESTATÍSTICAS	VALORES CONSOLIDADOS	VALORES ESTIMADOS
Observações	28	28
Média	3,90	3,82
Desvio padrão	0,22	0,30
Mínimo	3,32	2,99
25%	3,81	3,74
50%	3,86	3,87
75%	4,02	3,98
Máximo	4,36	4,36

Fonte: dados da pesquisa.

A coluna de valores consolidados apresenta as medidas obtidas pelos alunos que cursaram o oferecimento da disciplina D_2. Por exemplo, a média da turma com 28 alunos foi de 3,90.

Figura 4 - Aproveitamento Consolidado e Estimado

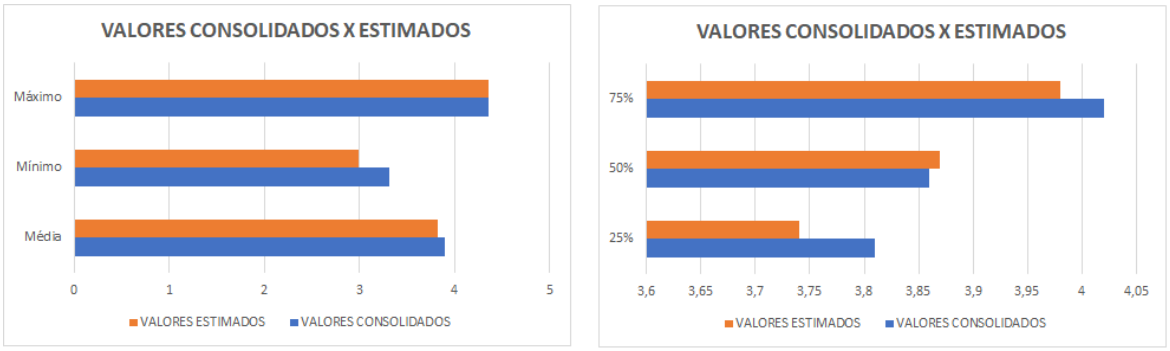


À esquerda temos os rendimentos que aconteceram no oferecimento de D_2. À direita temos o aproveitamento estimado pela hipótese ***h = 1,56x – 2,67***.

Fonte: dados da pesquisa.

A Figura 5 ilustra os comparativos dessas medidas. Os valores estimados foram menores que os consolidados em cinco das seis medidas, acentuando-se a diferença no quartil 75%.

Figura 5 - Valores consolidados x estimados para as medidas máximo, média, mínimo e quartis



Fonte: dados da pesquisa.

Para analisar se houve evidência estatisticamente significativa da diferença nas médias entre o rendimento acadêmico consolidado e o estimado, foi aplicado o Teste *t de Student*. Os valores médios foram usados para testar o desempenho da função *h*. Como o valor desse *teste* foi *menor* que *o valor crítico de t tabelado*, não há evidência estatisticamente significativa de que houve diferença nas médias das notas consolidadas e estimadas. Portanto, a hipótese *h* passou no teste.

A regressão logística foi utilizada para modelar a probabilidade de evasão dos estudantes a partir de seus desempenhos em três disciplinas: Algoritmos e Técnicas de Programação), Raciocínio Lógico e Matemático, e Algoritmos e Estruturas de Dados. Sua aplicação alcançou uma acurácia de 76%, o que significa que ele classificou corretamente 76% dos estudantes como evadidos ou não evadidos, conforme mostra a Tabela 3.

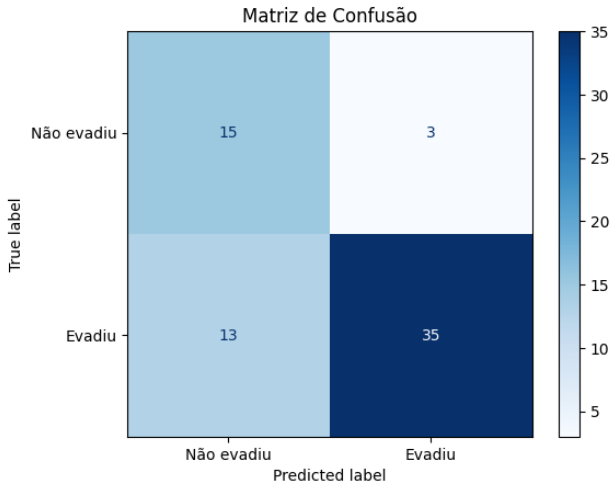
Tabela 3 - Avaliação do modelo de regressão logística para previsão de evasão

Métrica	Descrição	Não Evadido	Evadido	# Média
Precisão	Proporção de previsões positivas corretas	0,54	0,92	0,73
Revocação	Proporção de casos positivos corretamente ident	0,83	0,73	0,78
F1-score	Média harmônica de precisão e revocação	0,65	0,81	0,73
Suporte	Número total de observações	18	48	66,00
Acurácia	Proporção total de previsões corretas	-	-	0,76

Fonte: dados da pesquisa.

A matriz de confusão gerada a partir dos resultados da regressão logística fornece uma visão detalhada do desempenho do modelo. A análise da matriz exibida na Figura 6 revela que o modelo estimou que 35 alunos evadiram e **realmente eles evadiram**. Porém o modelo erroneamente previu 13 alunos como **não evadiram**, mas eles evadiram.

Figura 6 - Matriz de confusão



Fonte: dados da pesquisa.

A precisão na identificação de alunos evadidos foi menor (54%), o que significa que uma parcela significativa dos estudantes que evadiram não foi corretamente identificada pelo modelo. Essa diferença de precisão entre as duas classes pode ser influenciada por diversos fatores, como o desbalanceamento das classes no conjunto de dados, a escolha das variáveis preditoras e a complexidade do fenômeno da evasão, que pode ser influenciado por uma série de fatores além do desempenho acadêmico.

CONSIDERAÇÕES FINAIS

Este projeto foi uma pesquisa de natureza aplicada que se concentrou em estimar o desempenho acadêmico na disciplina de Algoritmos e Estruturas de Dados de um curso de Ciência da Computação por meio da aplicação de técnicas de aprendizado de máquina e análise estatística. As principais etapas do projeto incluíram a coleta e o tratamento de dados, a criação de um modelo de regressão linear para estimar o desempenho dos alunos e a avaliação da capacidade de generalização desse modelo por meio do Teste t de *Student*.

A pesquisa demonstrou a viabilidade da aplicação de técnicas de aprendizado de máquina na previsão do desempenho acadêmico e da probabilidade de evasão em cursos de Ciência da Computação. A regressão linear mostrou-se eficaz na estimativa de notas, enquanto a regressão logística apresentou resultados promissores na identificação de alunos em risco de evasão. A realização de análises adicionais, incluindo a avaliação da influência de outras variáveis, como características socioeconômicas e indicadores de engajamento dos alunos, seria fundamental para aprimorar a capacidade preditiva dos modelos e fornecer subsídios para a implementação de medidas eficazes de prevenção à evasão.

REFERÊNCIAS

AURELIANO, Viviane Cristina Oliveira; TEDESCO, Patricia Cabral de Azevedo Restelli; GIRAFFA, Lúcia Maria Martins. Desafios e oportunidades aos processos de ensino e de aprendizagem de programação para iniciantes. *In:* WORKSHOP SOBRE EDUCAÇÃO EM

COMPUTAÇÃO, 24., 2020, Porto Alegre. **Anais do XXIV Workshop sobre Educação em Computação.** Porto Alegre: SBC, 2020. p.71-80.

BINGHAM, Nicholas H.; FRY, John M. **Regression:** linear models in statistics. New York: Springer Science & Business Media, 2010.

CASTRO, Thais Helena Chaves de *et al*. Utilizando programação funcional em disciplinas introdutórias de computação. *In:* **Workshop sobre Educação em Computação (WEI)**, 2003, Campinas. Anais do WEI. Campinas: SBC, 2003.

GOMES, André Oliveira *et al*. **Uma abordagem de aprendizado de máquina para o auxílio no diagnóstico de evasão escolar em turmas do ensino médio.** 2017.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. 2. ed. New York: Wiley, 2000.

INMAN, Henry F. Karl Pearson and R. A. Fisher on statistical tests: a 1935 exchange from *Nature*. **The American Statistician**, v. 48, n. 1, p. 2-11, 1994.

KOLB, David Allen. **Experimental learning:** experience as the source of learning and development. Upper Saddle River: Prentice-Hall, 1984.

MCKNIGHT, Patrick E.; NAJAB, Julius. Mann-Whitney U test. **The Corsini Encyclopedia of Psychology**, p. 1-1, 2010.

MENARD, S. **Applied logistic regression analysis.** 2. ed. Thousand Oaks: Sage Publications, 2002.

MENOLLI, André; NETO, João Coelho. Uma análise do perfil dos cursos de licenciatura em computação no Brasil. **Revista Brasileira de Informática na Educação**, v. 29, p. 01-24, 2021.

MEURER, Alison Martins *et al.* Estilos de aprendizagem e rendimento acadêmico na universidade. **REICE: Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación,** v. 16, n. 4, p. 23-43, 2018. Disponível em: https://revistas.uam.es/reice/article/view/9955. Acesso em: 18 maio 2023.

NORVIG, Peter; RUSSELL, Stuart. **Inteligência artificial.** Rio de Janeiro: Grupo GEN, 2013.

OSBORNE, Jason W. **Regression & linear modeling: best practices and modern methods.** Thousand Oaks: Sage Publications, 2016.

PIANEZZER, Guilherme Augusto. **Modelagem estatística.** Curitiba: Ed. Contentus, 2020.

PRESS, S. J.; WILSON, S. Choosing between logistic regression and discriminant analysis. **Journal of the American Statistical Association**, v. 73, n. 364, p. 699-705, 1978.

RODRIGUES, Francisco Scheffel. **Estudo sobre a evasão no curso de ciência da computação da UFRGS**. 2013.

SILVA, Rodrigo Feitosa da. **Fatores que influenciam o desempenho acadêmico**, 2013. Disponível em: https://repositorio.insper.edu.br/handle/11224/793. Acesso em: 20 maio 2023.

SOUZA MIRANDA, Cláudio de; LIMA, João Paulo Resende de; MARINELLO, Matheus Canuto. Análise do rendimento acadêmico dos alunos de Ciências Contábeis da FEARP-USP beneficiados pelo INCLUSP/PASUSP. **Revista de Educação e Pesquisa em Contabilidade (REPeC)**, v. 12, n. 2, 2018. Disponível em: https://www.repec.org.br/repec/article/view/1629. Acesso em: 1 jun. 2023.

SOUZA, Vanessa Faria de. Mineração de dados educacionais com aprendizagem de

máquina. **Revista Educar Mais**, v. 5, n. 4, p. 766-787, 2021.

ZIMMERMAN, Donald W. Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. **The Journal of Experimental Education**, v. 55, n. 3, p. 171-174, 1987.