APLICAÇÃO DE ANÁLISE PREDITIVA NAS BASES DE DADOS DE UMA INSTITUIÇÃO DE ENSINO SUPERIOR, VISANDO AUMENTAR A EFICIÊNCIA NA ANÁLISE DE GRANDES QUANTIDADES DE LEADS

Jackson de Araújo Belloli[1], Marlon de Matos de Oliveira [2]

Resumo: O presente trabalho tem como objetivo aplicar técnicas de análise preditiva em bases de dados institucionais da Universidade do Extremo Sul Catarinense (UNESC), com o intuito de estimar a probabilidade de matrícula de estudantes a partir de informações coletadas durante o processo de inscrição. O estudo buscou otimizar a tomada de decisão e aumentar a eficiência na análise de grandes volumes de *leads* oriundos de campanhas de marketing digital. Para isso, foram utilizados dois algoritmos de aprendizado de máquina amplamente empregados em classificação supervisionada: Regressão Logística e *Random Forest*. A metodologia incluiu o tratamento e a codificação das variáveis, criação de atributos derivados e aplicação de métricas de avaliação como acurácia, precisão, *recall*, *F1-score* e AUC-ROC. Os resultados evidenciaram que o modelo de Regressão Logística apresentou desempenho mais equilibrado e interpretável, alcançando AUC-ROC de 0,709 e acurácia de 84,5%, enquanto o *Random Forest* obteve maior acurácia 87,5% porém menor capacidade discriminativa, AUC-ROC de 0,604. Conclui-se que o modelo logístico é o mais adequado ao contexto institucional, permitindo identificar com maior precisão os alunos com maior propensão à matrícula. Este estudo contribui tanto para o meio acadêmico, ao demonstrar o potencial da análise preditiva na educação, quanto para o contexto prático, oferecendo subsídios para o aprimoramento das estratégias de captação e permanência estudantil.

Palavras-chave: análise preditiva; aprendizado de máquina; regressão logística; random forest; previsão de matrículas.

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. jakson.belloli@unesc.net

[2]Orientador, Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. marlon.oliveira@unesc.net

ABSTRACT: This study aims to apply predictive analysis techniques to the institutional databases of the Universidade do Extremo Sul Catarinense (UNESC) in order to estimate students' enrollment probability based on data collected during the application process. The research seeks to enhance decision-making efficiency and improve the analysis of large volumes of leads generated through digital marketing campaigns. Two machine learning algorithms commonly used for supervised classification were employed: Logistic Regression and Random Forest. The methodology involved data preprocessing, categorical encoding, feature engineering, and performance evaluation using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Results indicated that the Logistic Regression model provided a more balanced and interpretable performance, achieving an AUC-ROC of 0.709 and accuracy of 84.5%, while the Random Forest achieved higher accuracy 87.5% but lower discriminative power, AUC-ROC of 0.604. It was concluded that the logistic model is more suitable for the institutional context, enabling better identification of students most likely to enroll. This study contributes both academically, by demonstrating the potential of predictive analytics in education, and practically, by providing insights for improving student recruitment and retention strategies.

Keywords: predictive analysis; machine learning; logistic regression; random forest; enrollment forecasting.

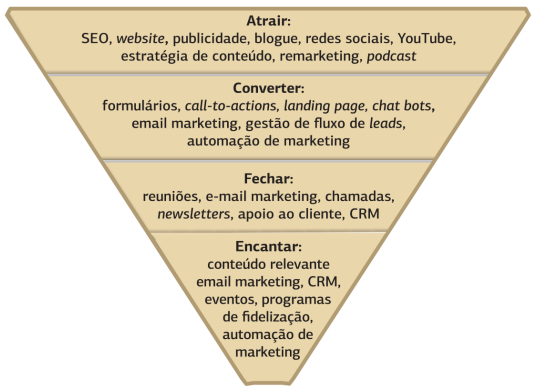
1 INTRODUÇÃO

A captação de clientes sempre foi considerada um dos fatores determinantes para o sucesso de qualquer organização, independentemente do setor de atuação. No contexto educacional, isso não é diferente: tanto instituições públicas quanto privadas dependem de estudantes para manter suas atividades. Enquanto na rede pública os alunos garantem repasses governamentais, nas instituições privadas as mensalidades financiam custos operacionais, infraestrutura e a própria sustentabilidade do negócio (Souza; Arantes; Dias, 2015). Assim, a capacidade de atrair e reter estudantes torna-se estratégica para a sobrevivência e expansão das Instituições de Ensino Superior (IES).

Nesse cenário, as práticas de marketing digital têm ganhado destaque. O avanço da internet, das mídias sociais e de dispositivos móveis transformou a relação entre organizações e consumidores (Grewal et al., 2020). Entre as principais abordagens, destaca-se o inbound marketing,

que busca atrair potenciais alunos por meio da produção de conteúdos relevantes, aproximando instituição e estudante de maneira menos invasiva e mais personalizada (Schuchmann; Figueira, 2020; Lima; Silva, 2021). Essa estratégia, demonstrada na figura 1 como funil de vendas, tem se mostrado eficaz para ampliar a visibilidade institucional e fortalecer a marca universitária, além de estabelecer vínculos duradouros com os alunos (Nadanyiova; Gajanova; Kliestikova, 2020; Belostecinic; Jomir, 2023).

Figura 1 - Funil de vendas



Fonte: Adaptada de Lima (2021)

O crescimento do número de *leads*, potenciais consumidores, gerados por canais digitais, entretanto, gera um desafio: identificar rapidamente quais contatos apresentam maior probabilidade de conversão em matrícula. Nesse ponto, a utilização de tecnologias analíticas se torna fundamental. O conceito de *Big Data* está relacionado à capacidade de processar grandes volumes de dados e extrair informações úteis para apoiar decisões estratégicas (Scaico; Queiroz; Scaico, 2014). No setor educacional, isso possibilita compreender o comportamento dos estudantes, prever tendências de matrícula e aprimorar processos de captação e retenção.

Entre as técnicas de análise de dados, destaca-se a Análise Preditiva, que utiliza métodos estatísticos e algoritmos de aprendizado de máquina para prever eventos futuros com base em padrões históricos (Pereira, 2014). Essa abordagem é frequentemente associada à Mineração de Dados (*Data Mining*), responsável por identificar padrões ocultos e relações entre variáveis (Peng, 2021). Complementarmente, os algoritmos de *Machine Learning* (ML) permitem identificar relações complexas e não lineares nos dados, aumentando a precisão das previsões (Santos, 2018; Herhausen et al., 2024).

Diversas técnicas estatísticas dão suporte a esse processo. A

Regressão Linear, por exemplo, é um método tradicional para modelar relações entre variáveis e realizar previsões (Bazdaric et al., 2021). O método dos mínimos quadrados amplia essa abordagem ao minimizar erros preditivos em contextos mais complexos, inclusive por meio de variantes robustas como *Least-squares support vector machine*, LS-SVM (Kumar et al., 2018). Já o método de Monte Carlo via Cadeias de Markov (MCMC) possibilita a simulação de distribuições complexas e incertezas, sendo amplamente aplicado em cenários de previsão e inferência estatística (Krüger et al., 2021; Al-Duais; Al-Sharpi, 2023).

No contexto educacional, a utilização desses modelos busca apoiar decisões estratégicas sobre a previsão do número de matrículas, recurso essencial para otimizar a alocação de docentes, turmas, infraestrutura e recursos financeiros. Estudos recentes demonstram que o uso de métodos preditivos, como regressão linear, *Random Forest* e MCMC, tem se mostrado eficaz na estimativa de ingressos em diferentes países e instituições (Williamson, 2022; Mohamed; Bukhatwa, 2024; Nurdin; Fauziah, 2024; Silva et al., 2025).

Embora diversos estudos utilizem técnicas de aprendizado de máquina na educação, muitos concentram-se em bases de dados limitadas, abordagens genéricas ou análises teóricas, sem considerar as particularidades institucionais e operacionais das universidades. Além disso, parte significativa das pesquisas foca apenas na acurácia dos modelos, negligenciando aspectos fundamentais como interpretabilidade, aplicabilidade prática e equilíbrio entre sensibilidade e especificidade (Guevara-Reyes et al., 2025).

O presente trabalho se diferencia por aplicar a análise preditiva diretamente sobre dados reais da UNESC, integrando variáveis acadêmicas, financeiras e demográficas em um modelo voltado à previsão de matrícula. Ao utilizar a estatística de Youden para otimizar o limiar de decisão, a pesquisa alcançou um equilíbrio superior entre precisão e sensibilidade, oferecendo um modelo com maior utilidade prática para o processo de captação e gestão de *leads* educacionais.

Portanto, compreender e comparar modelos de análise preditiva aplicados ao ensino superior é de grande relevância prática e científica. Este trabalho tem como objetivo geral avaliar e comparar diferentes modelos preditivos na previsão de matrículas em uma Instituição de Ensino Superior, utilizando dados históricos para mensurar sua acurácia. Os objetivos específicos incluem: compreender os conceitos fundamentais de marketing

digital, *data mining*, *machine learning* e análise preditiva; organizar e normalizar a base de dados da instituição; treinar diferentes modelos estatísticos e computacionais; aplicar métricas de avaliação; e, por fim, comparar os resultados obtidos.

A crescente complexidade do cenário educacional brasileiro, marcado por instabilidade econômica, concorrência acirrada e transformação digital. Como apontam Ribeiro e Reis (2020) e Costa e Almeida (2019), prever adequadamente a demanda por cursos permite às IES otimizar recursos, reduzir custos e elaborar estratégias de captação e retenção mais alinhadas às necessidades do público-alvo. Dessa forma, os resultados aqui obtidos poderão contribuir tanto para o avanço acadêmico quanto para a gestão estratégica institucional, oferecendo subsídios para decisões mais assertivas em relação ao planejamento e sustentabilidade das universidades.

2 TRABALHOS CORRELATOS

Os estudos recentes sobre previsão de matrículas em instituições de ensino superior evidenciam a relevância de modelos preditivos para apoiar o planejamento estratégico e a gestão acadêmica. Williamson (2022), ao investigar uma universidade no sudeste do Tennessee, aplicou regressão linear múltipla e simulações via Cadeias de Markov (MCMC) a partir de dados de seis semestres. Os modelos de regressão identificaram diferentes conjuntos de variáveis significativas por curso sendo a taxa de retenção institucional o único preditor comum a todos os modelos. Nas simulações MCMC, executadas 1.000 vezes com intervalo de confiança de 95% as previsões mostraram boa precisão em alguns semestres: por exemplo, o modelo institucional estimou 10.176 alunos em 2017 e 10.239 em 2018, próximos aos valores reais (10.111 e 10.158, respectivamente). Em programas específicos, as previsões também se mostraram consistentes, como em Biologia (712 alunos previstos vs. 710 reais em 2018) e Enfermagem (746 previstos vs. 746 reais em 2020). Ainda que a precisão tenha diminuído em 2021, atribuída a fatores externos como a pandemia de COVID-19 , os resultados demonstraram o potencial dos modelos preditivos baseados em MCMC para apoiar o planejamento estratégico e a alocação de recursos institucionais.

Em contexto distinto, Mohamed e Bukhatwa (2024) analisaram dados da Faculdade de Educação da Universidade de Benghazi, entre 2015 e 2023, testando três modelos matemáticos: mudança exponencial,

logístico e mínimos quadrados. A avaliação, conduzida por métricas como MAD, MAPE e RMSE, apontou o método dos mínimos quadrados como o mais eficaz, com 92% de precisão. Os autores destacam sua aplicabilidade como ferramenta estratégica para aprimorar o processo decisório e o planejamento institucional.

De forma semelhante, Nurdin e Fauziah (2024) investigaram a previsão de novas matrículas na Universitas Nasional, na Indonésia, utilizando os algoritmos *Random Forest* (RF) e Regressão Linear (LR) sobre um conjunto de 37.078 registros históricos de formulários de inscrição entre 2017 e 2023. Os autores compararam o desempenho dos modelos com base em métricas de erro, obtendo resultados superiores para o *Random Forest*, que apresentou MAE = 508,21, MSE = 514.729,78, RMSE = 717,45, MAPE = 8,22% e MAD = 274,71, enquanto a Regressão Linear obteve MAE = 1.210,00, MSE = 1.670.489,92, RMSE = 1.292,47, MAPE = 20,83% e MAD = 1.253,00. Além disso, o modelo *Random Forest* projetou uma tendência de crescimento estável nas matrículas de 2024 a 2028 (de 5.306 para 5.321 estudantes), em contraste com a regressão linear, que previu uma queda contínua (de 4.791 para 4.107). Diante desses resultados, os autores concluíram que o *Random Forest* é mais eficaz e consistente para a previsão de admissões estudantis, por lidar melhor com os valores discrepantes e apresentar menores desvios médios absolutos, reforçando a aplicabilidade de modelos de aprendizado de máquina no planejamento estratégico de instituições de ensino superior.

3 MATERIAIS E MÉTODOS

Esta pesquisa é de natureza aplicada, com abordagem quantitativa e caráter descritivo e comparativo. O estudo tem como propósito desenvolver e avaliar modelos de análise preditiva aplicados às bases de dados de uma Instituição de Ensino Superior (IES), com foco na previsão de matrículas a partir de leads captados por campanhas de marketing digital.

O desenvolvimento do projeto envolveu quatro etapas principais: (i) coleta e tratamento de dados institucionais; (ii) pré-processamento e preparação da base de dados; (iii) modelagem e avaliação de algoritmos preditivos; e (iv) análise comparativa dos resultados. A metodologia foi estruturada de forma a garantir a reprodutibilidade e a confiabilidade dos resultados obtidos.

3.1 FONTE DOS DADOSS E CONTEXTO

Os dados utilizados neste estudo foram fornecidos pela Universidade do Extremo Sul Catarinense (UNESC), provenientes do sistema interno de gestão acadêmica e financeira da instituição. O conjunto contempla informações referentes a inscrições em cursos da modalidade de ensino a distância e presencial, abrangendo variáveis relacionadas ao perfil do aluno, modalidade e categoria do curso, e dados financeiros da matrícula, sendo eles anonimizados e utilizados apenas para treinamento e teste dos modelos.

As principais variáveis consideradas no estudo incluem: curso, polo, dados do pagamento, status, desconto, reembolso, data de inscrição, categoria do curso, modalidade do curso, data de nascimento, sexo, escolaridade, educação, nome da cidade onde mora, estado onde mora, Raça/Cor e Escola de conclusão do Ensino Médio.

O objetivo da modelagem foi prever a probabilidade de um aluno efetivar a matrícula a partir de seu registro inicial de inscrição, auxiliando a instituição na identificação de potenciais matriculados e na formulação de estratégias de engajamento.

3.2 COLETA E PREPARAÇÃO DOS DADOS

Os dados foram exportados do sistema acadêmico da UNESC em formato tabular, planilha eletrônica, sendo posteriormente tratados e analisados em ambiente Python (*Jupyter Notebook*).

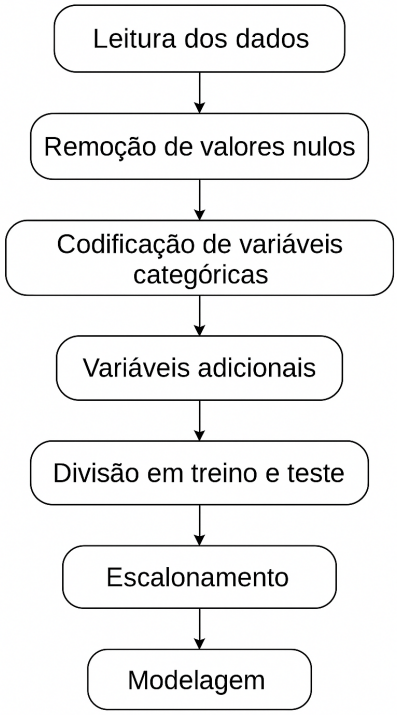
O pré-processamento envolveu múltiplas etapas para garantir a consistência e a qualidade das informações utilizadas na modelagem. Primeiramente, foi realizada a remoção de valores nulos e inconsistentes, como dados fora do formato esperado, especialmente em campos financeiros e de identificação de curso. Em seguida, foram aplicadas transformações para normalização de formatos de data e padronização de tipos de dados.

As variáveis categóricas, como sexo, modalidade do curso, categoria do curso e status, foram codificadas numericamente por meio de técnicas de *Label Encoding*, atribui números inteiros às categorias, e *One-Hot Encoding*, cria colunas binárias para cada categoria, conforme a natureza de cada atributo.

Também foram derivados atributos adicionais relevantes para o processo preditivo, como a faixa etária do aluno (obtida a partir da data de nascimento) e o tempo entre a inscrição e o vencimento do boleto. Esses

atributos auxiliaram na captura de relações comportamentais e financeiras entre os perfis de alunos e sua probabilidade de matrícula, assim finalizando o processo de preparação elencado na Figura 2.

Figura 2 - Fluxograma preparação dos dados



Fonte: Elaborado pelo autor.

3.3 AMOSTRAGEM E DIVISÃO DOS DADOS

Após o tratamento, o conjunto de dados foi dividido em duas amostras distintas, sendo uma destinada ao treinamento dos modelos e outra reservada para os testes de validação. A proporção adotada foi de 70% para o conjunto de treinamento e 30% para o conjunto de teste. Essa divisão foi realizada de forma estratificada, garantindo que ambas as amostras mantivessem a mesma proporção entre as classes “matriculado” e “não matriculado”. Esse cuidado assegurou a representatividade da variável-alvo e evitou que o modelo apresentasse viés decorrente de distribuições desbalanceadas.

3.4 MODELAGEM PREDITIVA

Na fase inicial do projeto, foram considerados algoritmos voltados para regressão contínua, com o intuito de estimar a probabilidade de matrícula em escala numérica. Contudo, a análise exploratória demonstrou que a variável dependente apresentava caráter binário, representando apenas dois possíveis resultados: matrícula efetivada ou não efetivada. Di-

ante dessa constatação, o problema foi reformulado como uma tarefa de classificação supervisionada.

A partir dessa reformulação, optou-se pela utilização de dois algoritmos amplamente empregados em problemas de classificação binária: Regressão Logística e *Random Forest*. A escolha da Regressão Logística foi motivada pela sua capacidade de fornecer uma interpretação direta dos coeficientes das variáveis independentes, permitindo compreender o peso de cada uma na probabilidade de matrícula. Já o algoritmo *Random Forest* foi selecionado como modelo comparativo, devido à sua habilidade de lidar com relações não lineares e de capturar interações complexas entre as variáveis preditoras.

Os modelos foram treinados utilizando o conjunto de dados tratado e dividido previamente. Durante o treinamento, seus hiperparâmetros foram ajustados com base em validação cruzada, a fim de equilibrar desempenho e capacidade de generalização. Essa etapa foi fundamental para evitar sobreajuste (*overfitting*), garantindo que os modelos mantivessem um bom desempenho quando aplicados a novos dados.

3.5 MÉTRICAS DE AVALIAÇÃO

O desempenho dos modelos foi avaliado a partir de métricas amplamente utilizadas em problemas de classificação supervisionada. A acurácia foi empregada para medir a proporção de previsões corretas sobre o total de observações, enquanto a precisão avaliou a proporção de verdadeiros positivos entre todas as previsões positivas realizadas pelo modelo. O *recall* (ou sensibilidade) mensurou a capacidade do modelo em identificar corretamente os casos positivos, e o *F1-score* foi calculado como a média harmônica entre precisão e *recall*, oferecendo uma visão equilibrada entre ambos. Além dessas métricas, utilizou-se a AUC-ROC (Área sob a Curva ROC) como indicador da capacidade discriminativa dos modelos, ou seja, a habilidade de distinguir corretamente entre alunos que se matricularam e os que não se matricularam.

Os resultados demonstraram que o algoritmo *Random Forest* obteve maior acurácia global, enquanto a Regressão Logística apresentou um valor superior de AUC-ROC, indicando maior capacidade de diferenciação entre as classes. Para otimizar o desempenho da Regressão Logística, aplicou-se a estatística de Youden (J = TPR - FPR), com o objetivo de determinar o limiar de decisão mais apropriado para o problema. A utilização desse método permitiu identificar o ponto que maximiza a diferença entre

a taxa de verdadeiros positivos e a taxa de falsos positivos, resultando em um equilíbrio mais adequado entre sensibilidade e especificidade. Esse ajuste melhorou o desempenho do modelo em identificar corretamente alunos com maior probabilidade de matrícula, sem comprometer a acurácia geral.

3.6 FERRAMENTAS E AMBIENTE COMPUTACIONAL

As análises foram realizadas na linguagem Python (versão 3.11), utilizando as bibliotecas pandas para manipulação de dados, numpy para operações numéricas, scikit-learn para modelagem estatística e aprendizado de máquina e matplotlib para geração de gráficos e visualização dos resultados. O ambiente de desenvolvimento adotado foi o *Jupyter Notebook*, o que possibilitou a execução modular do código, a documentação detalhada de cada etapa do processo e a visualização interativa das curvas ROC e demais métricas de desempenho.

3.7 SÍNTESE METODOLÓGICA

De forma resumida, o processo metodológico envolveu a coleta dos dados acadêmicos e financeiros da UNESC, o tratamento e a limpeza de inconsistências, a codificação das variáveis categóricas, a criação de atributos derivados e a divisão amostral em conjuntos de treino e teste de maneira estratificada. Em seguida, foram treinados e avaliados os modelos de Regressão Logística e *Random Forest*, cujos desempenhos foram comparados a partir de métricas de avaliação como acurácia, precisão, *recall*, *F1-score* e AUC-ROC. Posteriormente, aplicou-se a estatística de Youden para otimização do limiar de decisão da Regressão Logística, resultando em um modelo mais equilibrado e adequado ao contexto preditivo proposto. Por fim, a análise comparativa das métricas permitiu selecionar o modelo com melhor desempenho discriminativo e maior aplicabilidade prática para o cenário institucional analisado.

4 RESULTADOS E DISCUSSÃO

Após a etapa de pré-processamento dos dados e definição dos modelos, foram conduzidos testes comparativos entre os algoritmos de Regressão Logística e *Random Forest* para avaliar sua capacidade preditiva em relação à probabilidade de um estudante efetivar sua matrícula. A escolha desses modelos foi motivada tanto por sua ampla aplicação em problemas de classificação binária quanto pela necessidade de balancear in-

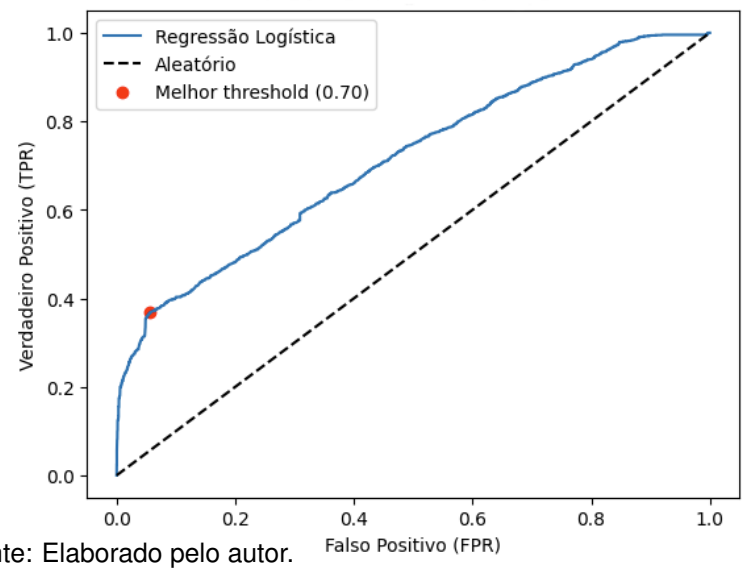
terpretabilidade e desempenho computacional.

Inicialmente, é importante ressaltar que, durante o desenvolvimento do projeto, houve a necessidade de substituição dos algoritmos preditivos originalmente propostos. A decisão se deu após análise exploratória dos dados, que revelou características mais compatíveis com modelos de classificação supervisionada, e não com os de previsão contínua originalmente considerados. Assim, optou-se pela aplicação da Regressão Logística, devido à sua natureza explicativa e interpretabilidade dos coeficientes, e da *Random Forest*, por sua robustez e capacidade de capturar interações não lineares entre as variáveis.

A etapa experimental foi conduzida utilizando os dados históricos da UNESC, contendo informações acadêmicas e demográficas dos estudantes, tais como curso, polo, valor total da matrícula, descontos aplicados, método de pagamento, cidade, sexo, escolaridade e outros atributos. Os dados foram divididos em conjuntos de treino e teste, garantindo a validação cruzada dos resultados e a minimização de viés amostral.

No caso da Regressão Logística, a análise da curva ROC demonstrou uma AUC-ROC de 0,703, quanto mais próximo de 1 melhor, indicando uma boa capacidade discriminativa do modelo, conforme ilustrado na Figura 3. O ponto de corte (*threshold*) otimizado foi identificado em aproximadamente 0,70, representando o equilíbrio entre taxa de verdadeiros positivos e falsos positivos. Com esse valor, o modelo atingiu acurácia de 84,5% precisão de 57,8% *recall* de 36,8% e um *F1-Score* de 0,450, conforme apresentado na Tabela 1.

Figura 3 - Curva ROC Regressão Logística



Fonte: Elaborado pelo autor.

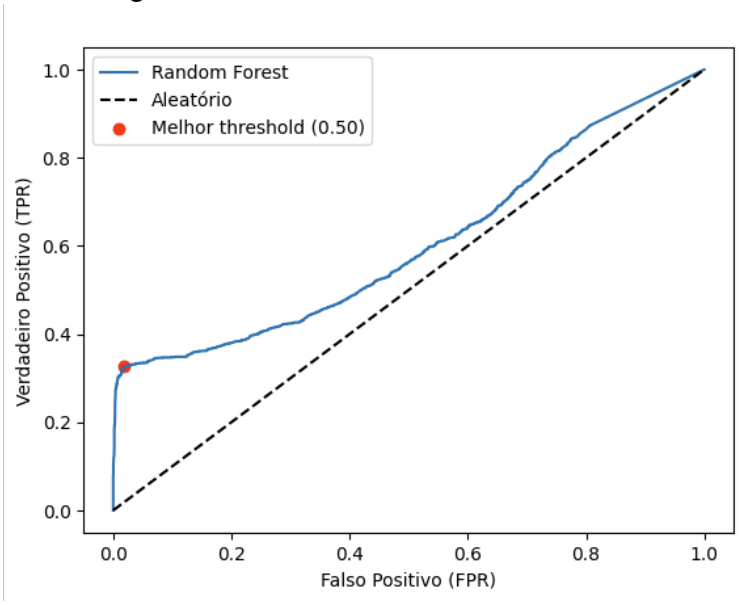
Tabela 1 - Matriz de confusão do modelo de Regressão Logística

	Predito Negativo (0)	**Predito Positivo (1)**
Real Negativo (0)	6050	358
Real Positivo (1)	844	491

Fonte: Elaborado pelo autor.

Já o modelo de *Random Forest*, apresentado na Figura 4, obteve uma AUC-ROC de 0,604, inferior à da Regressão Logística, indicando menor capacidade de separação entre classes. Contudo, sua acurácia global foi ligeiramente superior, alcançando 86,8% com precisão de 78,1% *recall* de 32,6% e *F1-Score* de 0,460, conforme demonstrado na Tabela 2. Esse comportamento demonstra que, embora a *Random Forest* apresente maior precisão geral, ela tende a ser mais conservadora na identificação de casos positivos (estudantes com alta probabilidade de matrícula), reduzindo o *recall* e, consequentemente, sua sensibilidade.

Figura 4 - Curva ROC *Random Forest*



Fonte: Elaborado pelo autor.

Tabela 2 - Matriz de confusão do modelo de *Random Forest*

	Predito Negativo (0)	**Predito Positivo (1)**
Real Negativo (0)	6286	122
Real Positivo (1)	900	435

Fonte: Elaborado pelo autor.

Ao comparar os dois modelos, como demostrado na Tabela 3, observa-se um *trade-off* entre acurácia e capacidade discriminativa. A Re-

gressão Logística apresentou maior área sob a curva ROC, evidenciando melhor equilíbrio entre sensibilidade e especificidade, o que a torna mais confiável para cenários em que é importante identificar potenciais alunos que ainda não confirmaram matrícula. Já o *Random Forest* se mostrou mais preciso para o conjunto majoritário, sendo útil em contextos operacionais que priorizam acertos absolutos, mesmo que à custa de uma menor detecção de casos positivos.

Tabela 3 - Comparativo de desempenho entre os modelos de Regressão Logística e *Random Forest*.

Métrica	**Regressão Logística**	***Random Forest***
Acurácia (%)	84,5	**86,8**
Precisão (%)	57,8	**78,1**
Recall (Sensibilidade) (%)	**36,8**	32,6
F1-Score	0,450	**0,460**
AUC-ROC	**0,713**	0,604

Fonte: Elaborado pelo autor.

A aplicação da métrica de Youden (J = TPR - FPR) para definição do ponto ótimo de corte foi um diferencial metodológico do estudo, pois permitiu otimizar o desempenho preditivo dos modelos em relação ao contexto institucional, como indica a Tabela 4. Essa abordagem é relevante, uma vez que, em problemas de classificação educacional, a simples maximização da acurácia pode mascarar a importância de prever corretamente casos minoritários, como estudantes em risco de evasão ou indecisos quanto à matrícula.

Tabela 4 - Matriz de confusão do modelo Regressão Logística sem a métrica.

	Predito Negativo (0)	**Predito Positivo (1)**
Real Negativo (0)	4568	1840
Real Positivo (1)	590	745

Fonte: Elaborado pelo autor.

De modo geral, os resultados indicam que a aplicação de técnicas de aprendizado de máquina a dados institucionais pode gerar informações valiosas para o processo decisório da universidade. O modelo de regressão logística, mesmo com uma acurácia ligeiramente inferior, apresentou um desempenho mais equilibrado e interpretável, sendo mais adequado para a análise preditiva de matrículas em um contexto de gestão educacional.

Em comparação com trabalhos correlatos, o presente estudo diferencia-se principalmente pela aplicação prática de modelos de aprendizado de máquina em dados reais e institucionais da Universidade do Extremo Sul Catarinense (UNESC). Enquanto pesquisas como a de Nurdin e Fauziah (2024) se concentraram em análises preditivas de séries históricas para estimar tendências de matrícula a partir de dados agregados, este trabalho aplicou os algoritmos de Regressão Logística e *Random Forest* diretamente sobre registros individuais de candidatos, incluindo variáveis demográficas, acadêmicas e financeiras. Essa abordagem permitiu avaliar o comportamento dos *leads* educacionais de maneira detalhada, identificando fatores associados à decisão de matrícula. Além disso, a adoção da estatística de Youden como critério de otimização do limiar de decisão representa um diferencial metodológico em relação às pesquisas anteriores, pois melhora a capacidade discriminatória dos modelos e amplia sua aplicabilidade prática.

Por fim, a contribuição deste trabalho se dá em dois eixos principais. No âmbito acadêmico, o estudo reforça a importância da utilização de métodos estatísticos e de aprendizado de máquina na área da educação, demonstrando empiricamente como modelos supervisionados podem ser empregados para prever comportamentos estudantis e apoiar políticas de permanência. No âmbito institucional, o projeto oferece uma ferramenta prática que pode auxiliar a UNESC a antecipar demandas de matrícula, identificar perfis de estudantes propensos à evasão e otimizar estratégias de retenção. A combinação de análise preditiva com dados reais institucionais evidencia o potencial transformador da integração entre ciência de dados e gestão educacional, abrindo caminhos para novas pesquisas e aplicações futuras.

5 CONCLUSÃO

O desenvolvimento deste trabalho permitiu validar a aplicação de técnicas de análise preditiva em bases de dados institucionais, confirmando sua viabilidade técnica e operacional para a previsão de matrículas na UNESC. A utilização dos algoritmos de Regressão Logística e *Random Forest* possibilitou a comparação de abordagens distintas de classificação. Constatou-se que o primeiro apresentou melhor equilíbrio entre sensibilidade e especificidade, enquanto o segundo alcançou maior acurácia geral. Essa análise comprovou que os modelos preditivos podem auxiliar de maneira efetiva o processo decisório em contextos educacionais.

O projeto atendeu integralmente aos objetivos propostos, englobando o tratamento e a preparação dos dados, o treinamento e a avaliação dos modelos, bem como a comparação de suas métricas de desempenho. A aplicação da estatística de Youden se destacou como um diferencial metodológico, pois otimizou o limiar de decisão e aprimorou a capacidade discriminativa da Regressão Logística. A pesquisa também reforçou a importância da interpretabilidade dos modelos, aspecto essencial para que as previsões possam ser compreendidas e utilizadas por gestores institucionais.

Em síntese, a pesquisa demonstra que o uso de aprendizado de máquina é uma ferramenta promissora para o planejamento e gestão educacional. Além de contribuir para o avanço acadêmico na área de ciência de dados aplicada à educação, o estudo oferece um recurso prático que pode auxiliar instituições como a UNESC a aprimorar suas estratégias de marketing, prever demandas futuras e otimizar recursos.

Como continuação, recomenda-se expandir o conjunto de dados e explorar novos algoritmos, como *Gradient Boosting* e *XGBoost*, a fim de aperfeiçoar a precisão e a robustez das previsões em cenários institucionais mais complexos. Também é sugerido o desenvolvimento de um painel interativo que integre os modelos preditivos e permita aos gestores de marketing e captação acompanhar, em tempo real, as previsões de conversão de *leads*, tornando o processo decisório mais ágil e orientado por dados. Tais aprimoramentos poderão fortalecer a aplicabilidade prática dos resultados e ampliar o impacto da análise preditiva na gestão educacional.

REFERÊNCIAS

AL-DUAIS, F. S.; AL-SHARPI, R. S. **A unique markov chain monte carlo method for forecasting wind power utilizing time series model**. Alexandria Engineering Journal, Elsevier B.V., v. 74, p. 51–63. ISSN 11100168. 2023.

BAZDARIC, K. et al. **The abc of linear regression analysis: What every author and editor should know**. [S.l.]: Pensoft Publishers, 2021. v. 47. ISSN 02583127.

BELOSTECINIC, G.; JOMIR, E. **Online Educational Marketing as a Means to Increase the Attractiveness of the University and Its Image**. Economica, v. 3, n. 3(125), p. 7–27. ISSN 18109136. 2023.

COSTA, F.; ALMEIDA, J. **Gestão baseada em dados no ensino superior: desafios e oportunidades**. Revista Brasileira de Administração Educacional, Belo Horizonte, v. 21, n. 2, p. 58–74. 2019.

GREWAL, D. et al. **The future of technology and marketing: a multidisciplinary perspective**. Journal of the Academy of Marketing Science, Springer, v. 48. ISSN 15527824. 2020.

GUEVARA-REYES, R. et al. **Machine learning models for academic performance prediction: interpretability and application in educational decision-making**. Frontiers in Education, Frontiers Media SA, v. 10. ISSN 2504284X. 2025.

HERHAUSEN, D. et al. **Machine learning in marketing: Recent progress and future research directions**. Journal of Business Research, v. 170, p. 114254. ISSN 0148-2963. Acesso em: 8 out. 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0148296323006136>. 2024.

KRüGER, F. et al. **Predictive inference based on markov chain monte carlo output**. International Statistical Review, International Statistical Institute, v. 89, p. 274–301. ISSN 17515823. 2021.

KUMAR, L. et al. **Effective fault prediction model developed using least square support vector machine (lssvm)**. Journal of Systems and Software, v. 137, p. 686–712. ISSN 0164-1212. Acesso em: 23 maio 2025. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0164121217300717>. 2018.

LIMA, A.; SILVA, R. **Modelos preditivos aplicados à educação: uma revisão sistemática**. Revista de Computação Aplicada, São Paulo, v. 13, n. 1, p. 45–63. 2021.

LIMA, A. P. D. **Inbound Marketing**. [S.l.]: Porto. CEOS Publicações, 2021. 207-225 p.

MOHAMED, F. N. A.; BUKHATWA, B. **Predicting student enrollment at the education college of benghazi via mathematical modeling approaches**. Journal of Al-Qadisiyah for Computer Science and Mathematics, v. 16. ISSN 2521-3504. Acesso em: 5 abr. 2025. Disponível em: <https://jqcsm.qu.edu.iq/index.php/journalcm/article/view/1803>. 2024.

NADANYIOVA, M.; GAJANOVA, L.; KLIESTIKOVA, J. **Building the brand of university through progressive forms of marketing communication**. IATED, 2020. 6769-6776 p. (14th International Technology, Education and Development Conference). Acesso em: 28 ago. 2024. Disponível em: <https://doi.org/10.21125/inted.2020.1802>.

NURDIN, M.; FAUZIAH, F. **Analytical study forecasting students using random forest and linear regression algorithms**. sinkron, v. 8, p. 2369–2378. ISSN 2541-2019. Acesso em: 21 mar. 2025. Disponível em: <https://jurnal.polgan.ac.id/index.php/sinkron/article/view/13886>. 2024.

PENG, Z. **Application of Mining Algorithm in Personalized Internet Marketing Strategy in Massive Data Environment**. 2021. Acesso em: 28 nov. 2024. Disponível em: <https://www.researchsquare.com/article/rs-181097/v1>.

PEREIRA, J. L. **Análise preditiva em sistemas de informação no contexto do big data**. Trabalho de conclusão de cursos Sistemas de Informação - UNIVEM. 2014.

RIBEIRO, T.; REIS, J. L. **Artificial Intelligence Applied to Digital Marketing**. Cham: Springer International Publishing, 2020. 158–169 p.

SANTOS, H. G. dos. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina**. p. 207. Acesso em: 10 nov. 2024. Disponível em: <http://www.teses.usp.br/teses/disponiveis/6/6141/tde-09102018-132826/>. 2018.

SCAICO, P. D.; QUEIROZ, R. J. G. B.; SCAICO, A. **O Conceito Big Data Na Educação**. 3º Congresso Brasileiro de Informática na Educação (CBIE 2014), v. 3, p. 328–336. 2014.

SCHUCHMANN, B. M.; FIGUEIRA, A. A. **Do marketing tradicional ao marketing digital uma análise a partir dos programas de marketing digital online**. [S.l.]: Companhia Brasileira de Producao Cientifica, 2020. v. 2. 1–12 p.

SILVA, L. C. e et al. **Forecasting student enrollments in brazilian schools for equitable and efficient education resource allocation**. Conference on Digital Government Research, v. 1. ISSN 3050-8681. Acesso em: 19 maio 2025. Disponível em: <https://proceedings.open.tudelft.nl/DGO2025/article/view/939>. 2025.

SOUZA, B. C. C.; ARANTES, J. C. d. S.; DIAS, S. A. A. **Captação de alunos**. [s.n.], 2015. v. 15. 87–105 p. Acesso em: 22 nov. 2021. ISSN 2178-6909. Disponível em: <http://pgsskroton.com.br/seer/index.php/rcger/article/view/2061>.

WILLIAMSON, C. T. **Predicting enrollment in a metropolitan university in Southeast Tennessee**. Tese (PhD dissertation) — University of Tennessee at Chattanooga, Chattanooga, Tennessee, 2022.