UNIVERSIDADE FEDERAL DO ABC

HACKATHON DATAS

ANA CAROLINA SIQUEIRA FERNANDES PARRA

HEITOR HIROYUKI SHIRAI

ISAC DO NASCIMENTO VIEIRA

JOÃO MARCO JORGE FRANCISCON

ANÁLISE DOS FATORES SOCIOECONÔMICOS NO INGRESSO À UFABC

Santo André

2025

ANA CAROLINA SIQUEIRA FERNANDES PARRA

HEITOR HIROYUKI SHIRAI

ISAC DO NASCIMENTO VIEIRA

JOÃO MARCO JORGE FRANCISCON

ANÁLISE DOS FATORES SOCIOECONÔMICOS NO INGRESSO À UFABC

Projeto de análise apresentado ao Hackathon do curso de Bacharelado em Ciência De Dados da Universidade Federal do ABC.

Santo André

2025

SUMÁRIO

1 INTRODUÇÃO

No Brasil, o acesso à educação superior pública sempre esteve marcado por profundas desigualdades sociais. Ainda que o país tenha avançado significativamente nas últimas décadas com políticas de expansão universitária e ações afirmativas, o Exame Nacional do Ensino Médio (ENEM), principal instrumento de seleção para universidades públicas por meio do Sistema de Seleção Unificada (SISU), continua refletindo, e muitas vezes reproduzindo, as disparidades históricas que afetam milhões de estudantes.

Enquanto o ENEM se propõe como uma porta aberta a todos, a realidade mostra que nem todos os candidatos chegam à prova em condições iguais. Estudantes de famílias com baixa renda, oriundos de escolas públicas, moradores de regiões periféricas e pertencentes a grupos étnico-raciais historicamente marginalizados enfrentam obstáculos estruturais ao longo de toda sua trajetória educacional. Esses fatores impactam diretamente o desempenho no exame e, por consequência, as chances de ingresso nas universidades públicas, especialmente nas instituições mais concorridas, como a Universidade Federal do ABC (UFABC).

A UFABC, por sua proposta pedagógica inovadora e seu compromisso com a inclusão, tem se tornado cada vez mais um destino buscado por estudantes de diferentes partes do país. No entanto, é fundamental investigar quem de fato está conseguindo entrar na universidade e quais são os fatores socioeconômicos que contribuem para esse sucesso ou para a exclusão.

A análise crítica dessas desigualdades permite compreender até que ponto políticas como o SISU, as cotas e as bonificações regionais estão sendo eficazes na promoção da equidade.

2 OBJETIVO

O objetivo principal deste projeto é analisar o perfil socioeconômico dos candidatos que realizam o ENEM e utilizam sua nota para concorrer a vagas na UFABC por meio do SISU. A proposta busca identificar e compreender a influência de diferentes variáveis socioeconômicas, tais como: renda familiar, tipo de escola cursada (pública ou privada), cor ou raça, localização geográfica e outros fatores. Assim, analisar a probabilidade de aprovação dos candidatos, nas diversas modalidades de ingresso oferecidas pela universidade, por meio de análise e modelagem de dados.

A motivação central da análise é revelar padrões e possíveis desigualdades no vestibular, evidenciando como determinados grupos sociais têm mais ou menos chances de acesso ao ensino superior público, mesmo dentro de um sistema que visa promover equidade. Para isso, serão utilizados métodos estatísticos e de modelagem preditiva, com o intuito de calcular a probabilidade de aprovação de candidatos a partir de seu perfil socioeconômico e educacional.

Além de levantar dados e propor interpretações quantitativas, o projeto também visa fomentar uma reflexão crítica sobre a efetividade das políticas de inclusão, como as cotas raciais e sociais, e sobre o papel das universidades federais na promoção da justiça social. Os resultados poderão subsidiar futuras decisões institucionais e políticas públicas voltadas à ampliação do acesso e à redução das desigualdades educacionais.

O objetivo geral é, portanto, não apenas diagnosticar realidades, mas também contribuir com propostas de transformação, reafirmando o compromisso da universidade com a diversidade, a inclusão e a excelência acadêmica.

3 DADOS

Neste projeto, optamos por utilizar algumas bases de dados entre as várias fornecidas pela organização do Hackathon, todas referentes ao ano de 2023. As fontes escolhidas foram: os microdados do INEP, que reúnem um conjunto de informações detalhadas relacionadas às pesquisas, aos exames e avaliações do Instituto; os dados abertos do SISU, que contêm registros sobre inscrições e resultados do vestibular; e o repositório de dados da UFABC, que fornece a relação de candidatos inscritos na UFABC pelo SISU na chamada regular e na lista de espera, classificados por campus, curso, turno e modalidade de concorrência. A escolha dessas fontes se deu pelo potencial de cruzamento entre elas, o que nos permitiu analisar o perfil dos candidatos que tentam uma vaga na UFABC via SISU e compreender como fatores sociais e econômicos influenciam suas chances de aprovação.

4 FERRAMENTAS

Para a realização da análise dos dados, utilizamos duas principais ferramentas. Inicialmente, o Microsoft Excel foi empregado para a extração e organização preliminar dos dados, facilitando a visualização e a filtragem de informações relevantes. Em seguida, para uma análise mais aprofundada e detalhada, utilizamos a linguagem de programação Python, que permite maior flexibilidade no tratamento dos dados e na criação de visualizações. As bibliotecas utilizadas incluíram o pandas, para manipulação e limpeza das tabelas; numpy, para operações numéricas; matplotlib e seaborn, para a geração de gráficos e visualizações que ajudaram a identificar padrões, correlações e comparações entre diferentes variáveis socioeconômicas.

Na etapa de modelagem, foram testadas diferentes abordagens, incluindo regressão (Random Forest Regressor, Regressão Linear e XGBRegressor) e classificação (XGBClassifier). Foram utilizadas bibliotecas como XGBoost e Scikit-learn. Cada modelo apresentou resultados distintos na tarefa de prever a nota do ENEM, com foco na aplicação de técnicas preditivas. Com o uso do XGBClassifier, foi possível classificar, a partir da nota prevista, a probabilidade de um candidato ser aprovado ou não.

5 METODOLOGIA

Este estudo adota uma abordagem quantitativa e exploratória, com ênfase na análise estatística e modelagem preditiva. O objetivo é compreender como variáveis socioeconômicas influenciam o desempenho dos candidatos no Enem e, por consequência, sua aprovação na UFABC por meio do Sisu. A metodologia está estruturada em cinco etapas principais: coleta e integração de dados, tratamento e limpeza, análise descritiva, modelagem preditiva e visualização dos resultados.

5.1 Coleta e Integração de Dados

Foram utilizadas bases de dados públicas de 2023 disponibilizadas pela organização do Hackathon, incluindo:

- **Microdados do INEP**, contendo informações detalhadas sobre o ENEM;
- **Dados abertos do SISU**, com registros de inscrições e resultados;
- **Repositório da UFABC**, com a lista de candidatos classificados por curso, turno, campus e modalidade.

Não foi possível realizar o cruzamento entre essas fontes o que dificultou identificar candidatos que se inscreveram na UFABC via SISU e associar suas informações socioeconômicas e educacionais aos resultados obtidos no vestibular.

5.2 Tratamento e Qualidade dos Dados

Inicialmente, as bases foram importadas para o ambiente Python, onde foram tratadas utilizando as bibliotecas Pandas e NumPy. As etapas incluíram:

- Limpeza de dados nulos ou inválidos;
- Padronização de variáveis;
- Criação de categorias para análise por faixas de renda, tipo de escola, cor ou raça e localização geográfica.

5.3 Análise Descritiva

Com o uso das bibliotecas pandas, numpy, matplotlib.pyplot e seaborn, foram gerados gráficos e visualizações para analisar a distribuição e a correlação entre variáveis. Antes de iniciar a análise, foram elaboradas perguntas-chave para guiar o estudo, como: Qual é a relação entre concorrência (número de candidatos por vaga) e nota de corte em cada modalidade? Há diferença significativa entre candidatos de escolas públicas e privadas? Qual é o perfil predominante dos aprovados na UFABC por meio do SISU? A definição dessas questões permitiu direcionar o uso das ferramentas de forma mais eficaz. Diversas técnicas foram empregadas, como histogramas, gráficos de dispersão e boxplots, com o objetivo de destacar os grupos com maior influência no desempenho dos candidatos.

5.4 Modelagem Preditiva

Na etapa de modelagem, foram testadas abordagens de regressão e classificação, com foco na previsão da nota do ENEM e na probabilidade de aprovação na UFABC. Os modelos aplicados incluíram:

- **Regressão Linear**;
- **Random Forest Regressor**;
- **XGBoost Regressor e Classifier;**

Após validação dos modelos, o XGBRegressor apresentou maior desempenho para estimar a chance de aprovação, a partir da nota prevista e de variáveis socioeconômicas.

A modelagem preditiva foi dividida em duas etapas complementares:

Na primeira etapa, utilizamos os macrodados do ENEM. Apesar da riqueza de informações, a grande quantidade de variáveis e a alta heterogeneidade dos dados dificultaram a obtenção de resultados precisos. Isso evidenciou as limitações das abordagens utilizadas (como regressão e XGBoost) nesse contexto, sugerindo a possibilidade de explorar modelos mais robustos, como os baseados em deep learning, para lidar com a complexidade dos dados. Os resultados obtidos com os modelos para os macrodados do INEP foram:

- **Regressão Linear:** (R2: 0.3222, MSE: 6526.13);
- **Random Forest Regressor:** (R2: 0.3430, MSE: 6326.07);
- **XGBoost Regressor:** (R2: 0.4098, MSE: 5726.13).

Na segunda etapa, realizamos a modelagem com base nos dados da Prograd/UFABC, focando nos candidatos efetivamente matriculados. Os dados foram agrupados por curso, permitindo a criação de clusters mais homogêneos. Com isso, os modelos alcançaram métricas significativamente melhores em termos de erro médio (MSE) e coeficiente de determinação (R2), indicando maior capacidade de previsão em contextos com menos variabilidade. Os resultados para os grupos clusterizados por opção de curso da Prograd foram:

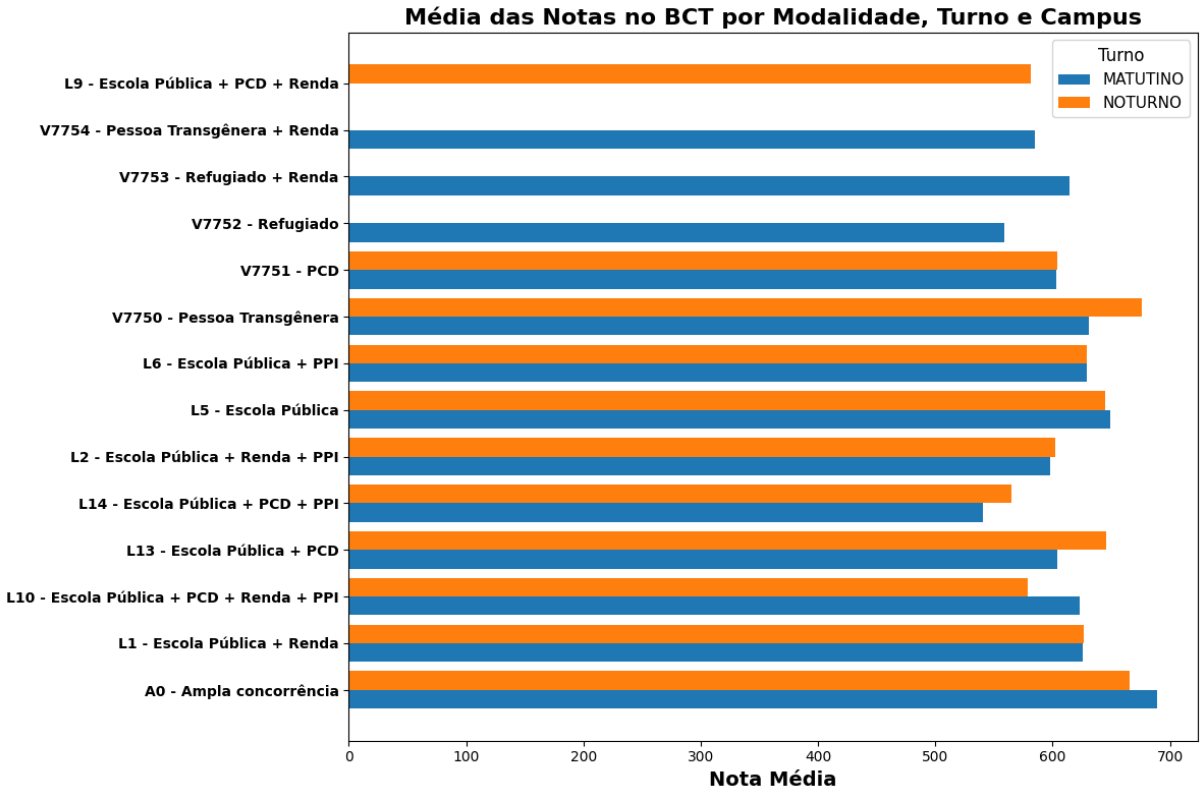
- **Grupo:** ('BACHARELADO EM CIÊNCIA E TECNOLOGIA ',) MSE: 42.30 | R2: 0.99
- **Grupo:** ('BACHARELADO EM CIÊNCIAS E HUMANIDADES ',) MSE: 35.67 | R2: 0.99
- **Grupo:** ('LICENCIATURA EM CIÊNCIAS HUMANAS ',) MSE: 655.76 | R2: 0.80
- **Grupo:** ('LICENCIATURA EM CIÊNCIAS NATURAIS E EXATAS ',) MSE: 282.72 | R2: 0.88

Após essa etapa, seguimos para o deploy utilizando a biblioteca Streamlit, criando uma interface interativa em que o usuário pode inserir informações como modalidade de matrícula, opção de curso, campus e turno. A partir desses dados, o sistema realiza a previsão da nota necessária e apresenta uma análise da probabilidade de aprovação no curso desejado.

6 RESULTADOS OBTIDOS

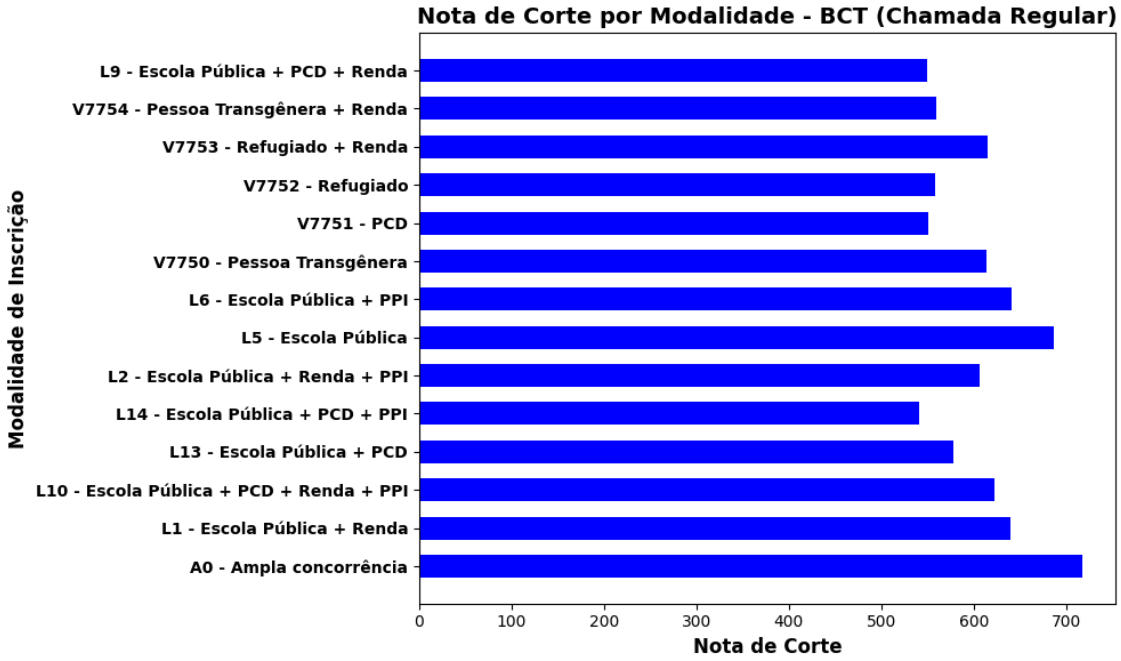
Os resultados das análises foram interpretados com foco na identificação de padrões de exclusão ou inclusão de determinados grupos sociais. Os modelos geraram probabilidades individuais de aprovação que, agregadas, revelaram tendências estruturais no acesso à UFABC. Além da análise quantitativa, os dados subsidiaram reflexões sobre a efetividade de políticas como o SISU, as cotas e bonificações regionais. A partir das análises dos gráficos, foi possível notar alguns resultados importantes que evidenciam as desigualdades no acesso à UFABC:

Figura 1: Gráfico de média das notas dos candidatos ao curso de BCT, por modalidade de ingresso, turno e campus



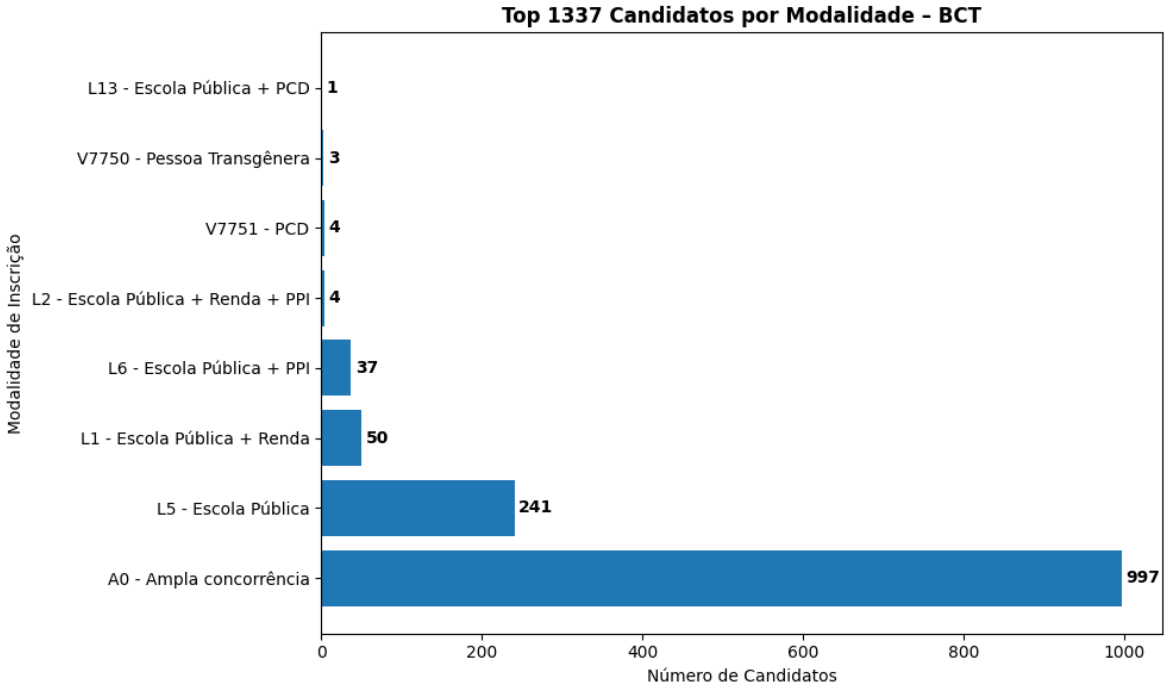
Fonte: Elaboração própria com dados da UFABC (2023).

Figura 2: Gráfico de Nota de Corte por Modalidade - BCT (Chamada Regular)



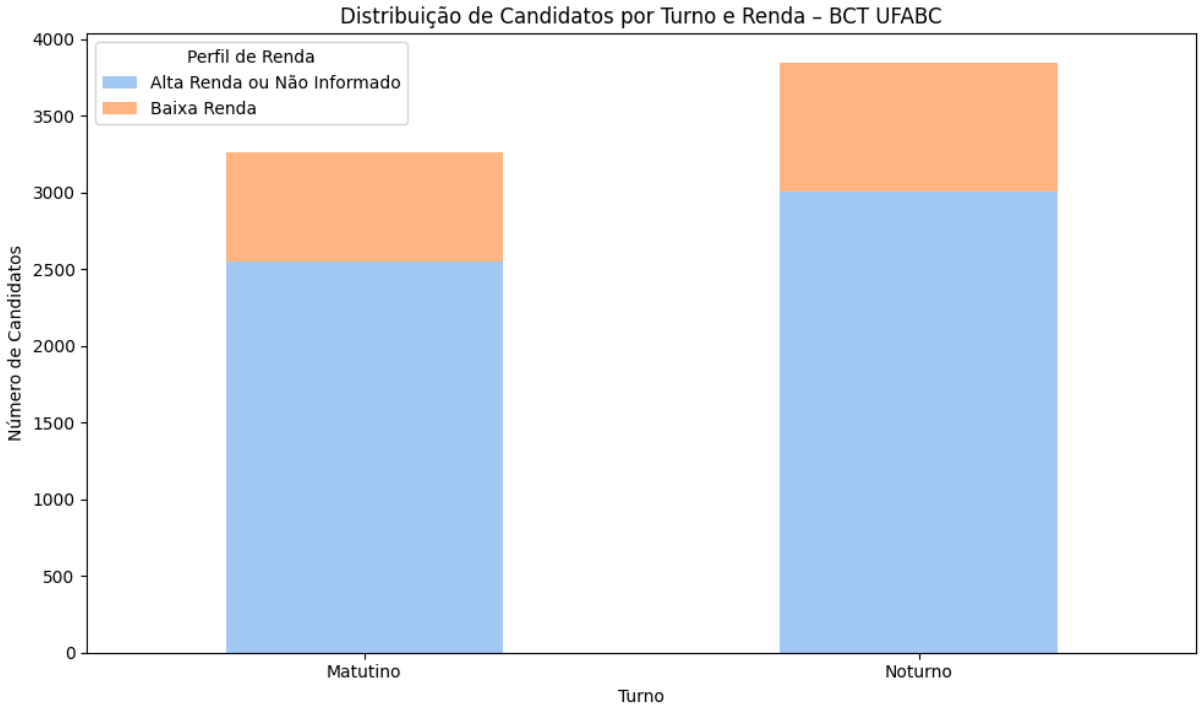
Fonte: Elaboração própria com dados do INEP e da UFABC (2023).

Figura 3: Gráfico Comparativo de Classificação por Modalidade no Cenário Hipotético Sem Políticas Afirmativas



Fonte: Elaboração própria com dados da UFABC (2023).

Figura 4: Distribuição de Candidatos por Turno e Renda do Bacharelado em Ciência e Tecnologia



Fonte: Elaboração própria com dados do INEP, SISU e UFABC (2023).

A análise dos dados revelou padrões claros de desigualdade no acesso à UFABC por meio do SISU, especialmente quando se observam as diferentes modalidades de ingresso. A partir da comparação entre notas médias e notas de corte nas diferentes modalidades de inscrição (ampla concorrência, cotas raciais, escola pública, renda, entre outras), foi possível identificar que candidatos da ampla concorrência tendem a ter notas mais altas. Isso reflete uma vantagem significativa na disputa por vagas, especialmente nos cursos mais concorridos, como o Bacharelado em Ciência e Tecnologia.

A análise mostrou que as políticas de cotas têm um papel importante em equilibrar o ingresso na UFABC, ampliando o acesso para grupos historicamente desfavorecidos. No entanto, os dados também evidenciam que ainda há uma grande diferença de desempenho entre os candidatos dessas modalidades e os da ampla concorrência, especialmente em relação às notas médias e às notas de corte. Isso indica que, embora as cotas ajudem a garantir o acesso, às desigualdades educacionais e socioeconômicas continuam influenciando diretamente as chances de

aprovação, mostrando que o problema vai além do sistema de seleção e está profundamente ligado às condições de formação dos estudantes antes do ENEM.

A modelagem preditiva aplicada ao longo deste estudo revelou distinções importantes na performance dos modelos em diferentes conjuntos de dados. Apesar da vasta riqueza e da alta heterogeneidade dos macrodados do INEP, a aplicação de modelos de regressão tradicionais (Regressão Linear, Random Forest Regressor e XGBoost Regressor) para prever a nota do ENEM não se mostrou tão eficaz quanto o esperado. O desempenho desses modelos, embora demonstrasse alguma capacidade preditiva, apontou para a complexidade intrínseca dessa base de dados, sugerindo que abordagens mais robustas, como as baseadas em Deep Learning, poderiam ser necessárias para capturar as nuances e interações entre a multitude de variáveis presentes.

Adicionalmente, foram testadas técnicas de pré-processamento, como a clusterização não supervisionada via KMeans e a redução de dimensionalidade com PCA. Observou-se que o uso de KMeans não impactou o desempenho do XGBRegressor, indicando que o modelo já era capaz de identificar padrões latentes nos dados sem necessidade de agrupamentos explícitos. Da mesma forma, a aplicação do PCA que reduziu o número de variáveis de 45 para 33 não resultou em melhorias significativas nas métricas de avaliação, o que reforça a eficiência do XGBoost em lidar com alta dimensionalidade e selecionar automaticamente as variáveis mais relevantes.

Em contraste, a segunda etapa da modelagem, que utilizou a base de dados da Prograd/UFABC, focando nos candidatos efetivamente matriculados e agrupando-os por curso (clusterização), demonstrou uma precisão significativamente maior. A menor quantidade de variáveis e a maior homogeneidade dos dados dentro de cada cluster permitiram que os modelos alcançassem métricas substancialmente melhores de erro médio (MSE) e coeficiente de determinação (R^2), evidenciando a capacidade preditiva em contextos mais específicos e com menor variabilidade. Esta abordagem facilitou uma compreensão mais acurada dos fatores que influenciam a aprovação em cursos específicos da UFABC.

7 CONCLUSÃO

Este estudo evidenciou como fatores socioeconômicos impactam significativamente as chances de aprovação de candidatos no ensino superior público, em especial na UFABC. Ao utilizar diferentes bases de dados e aplicar técnicas de análise estatística e modelagem preditiva, foi possível observar que a complexidade e heterogeneidade dos macrodados do Enem dificultam previsões precisas, mesmo com o uso de modelos avançados como o XGBoost. A aplicação de técnicas como KMeans e PCA também demonstrou limitações nesse contexto, indicando que o modelo XGBRegressor já é eficiente em identificar padrões sem necessidade de agrupamentos ou redução de variáveis.

Com base nos resultados obtidos da análise, é possível concluir que o acesso à UFABC via Sisu ainda reflete desigualdades socioeconômicas significativas entre os candidatos. Embora as políticas de cotas desempenhem um papel importante ao ampliar as oportunidades de ingresso para estudantes de baixa renda, de escolas públicas e autodeclarados pretos, pardos ou indígenas, os dados mostram que o desempenho médio desses grupos continua inferior ao dos candidatos da ampla concorrência. Essa diferença revela não apenas as limitações do sistema de cotas em compensar totalmente as desigualdades estruturais, mas também reforça a necessidade de políticas educacionais mais amplas, que atuem na base do ensino e garantam maior equidade no preparo dos estudantes antes mesmo da realização do Enem. A análise também evidencia que fatores como tipo de escola cursada, renda familiar e localização geográfica seguem sendo determinantes nas chances de aprovação, o que aponta para um cenário onde o mérito individual ainda está profundamente condicionado pelas oportunidades oferecidas ao longo da trajetória escolar.

Por outro lado, a modelagem com dados da Prograd/UFABC, mais específicos e homogêneos, mostrou desempenho substancialmente superior. O agrupamento dos dados por curso reduziu a variabilidade e permitiu análises mais precisas, com métricas preditivas significativamente melhores (R^2 próximos de 0,99 em alguns casos). Essa diferença de performance ressalta a importância de contextualizar os dados utilizados em estudos preditivos, bem como a necessidade de abordagens mais direcionadas quando se trabalha com conjuntos amplos e diversos como os do Enem.

Além dos resultados quantitativos, este trabalho reforça a necessidade de revisão constante das políticas públicas de acesso ao ensino superior. Apesar dos avanços com ações afirmativas e sistemas como o Sisu, ainda persistem desigualdades estruturais que afetam o desempenho de estudantes em situação de vulnerabilidade. Os achados aqui apresentados podem subsidiar novas estratégias institucionais voltadas à ampliação da inclusão e ao fortalecimento do papel social das universidades públicas no Brasil.

8 REFERÊNCIAS

UNIVERSIDADE FEDERAL DO ABC – PROGRAD/DSSI. *bd_prograd01_2023: Classificação Geral do Ingresso por modalidade – Candidatos inscritos no SISU/Ufabc (2023)*. Santo André: UFABC, 2024. 1 arquivo (formato CSV). Disponível em: https://dados.ufabc.edu.br/bases-dados/44-bd-prograd01. Acesso em: mai. 2025.

BRASIL. Ministério da Educação. **Sistema de Seleção Unificada – SISU: dados abertos**. Brasília: MEC, [2023]. Disponível em: https://dadosabertos.mec.gov.br/sisu. Acesso em: mai. 2025.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP. *Microdados do Enem 2023*. Brasília: INEP, 2024. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: mai. 2025.