ANÁLISE DA PARTICIPAÇÃO FEMININA NOS CURSOS DA ÁREA DE COMPUTAÇÃO DA REDE FEDERAL.

ANALYSIS OF FEMALE PARTICIPATION IN COMPUTER SCIENCE COURSES IN THE FEDERAL NETWORK.

Ellen Patricia Lopes de Santana
epls@discente.ifpe.edu.br
Viviane Cristina Oliveira Aureliano
viviane.aureliano@jaboatao.ifpe.edu.br

RESUMO

Este estudo analisa a participação feminina nos cursos de Computação da Rede Federal de Ensino do Brasil, a partir dos microdados disponibilizados na Plataforma Nilo Peçanha no período de 2017 a 2023. A análise foi conduzida em duas etapas. Inicialmente, adotou-se uma abordagem quantitativa descritiva, com o uso do Power BI, para examinar os dados de todos os cursos de Computação da Rede Federal. Em seguida, foram considerados exclusivamente os dados do Instituto Federal de Pernambuco (IFPE), a fim de realizar uma análise preditiva por meio de técnicas de *Machine Learning*, com o objetivo de identificar variáveis capazes de prever a evasão de estudantes nos cursos de Computação dessa instituição. Os resultados indicam crescimento contínuo da participação feminina, alcançando 37,64% no IFPE em 2023, percentual superior à média nacional. Contudo, a evasão mostra-se crítica entre estudantes de baixa renda e pardas. O algoritmo *Random Forest* apresentou o melhor desempenho na predição do risco de evasão, com acurácia de 70,32% e recall de 84,67%. Conclui-se que fatores socioeconômicos e raciais exercem maior peso preditivo sobre a evasão do que o gênero isoladamente. Dessa forma, políticas de retenção devem priorizar a vulnerabilidade social, visando garantir a permanência feminina na área da Computação.

Palavras-chave: Participação Feminina; Gênero; Computação; Aprendizado de Máquina; Evasão Escolar; Rede Federal; IFPE.

ABSTRACT

This study analyzes female participation in Computing courses at the Federal Education Network of Brazil, based on the microdata provided on the Nilo Peçanha Platform from 2017 to 2023. The analysis was conducted in two stages. Initially, a descriptive quantitative approach was adopted, using Power BI to examine the data from all Computer Science courses in the Federal Network. Next, only the data from

Instituto Federal de Pernambuco. Campus Jaboatão dos Guararapes. Curso de Análise e Desenvolvimento de Sistemas. 05 de janeiro de 2026.

the Federal Institute of Pernambuco (IFPE) was considered, in order to conduct a predictive analysis using Machine Learning techniques, with the aim of identifying variables capable of predicting student dropout in the Computing courses of this institution. The results indicate a continuous growth in female participation, reaching 37.64% at IFPE in 2023, a percentage higher than the national average. However, dropout rates are critical among low-income and brown students. The Random Forest algorithm showed the best performance in predicting dropout risk, with an accuracy of 70.32% and a recall of 84.67%. It is concluded that socioeconomic and racial factors have a greater predictive weight on dropout rates than gender alone. Thus, retention policies should prioritize social vulnerability, aiming to ensure the continued presence of women in the field of Computing.

Keywords: Female Participation; Gender; Computing; Machine Learning; School Dropout; Federal Network; IFPE.

1 INTRODUÇÃO

Compreender detalhes sobre o ensino é uma ferramenta valiosa para reduzir a evasão escolar e promover a integração em escolas e universidades. A análise dos registros disponíveis permite identificar padrões e comportamentos que auxiliam na previsão de desistências, além de contribuir para a elaboração de estratégias e políticas oficiais voltadas à permanência dos estudantes. No Brasil, organizações como o Ministério da Educação (MEC) e o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) trabalham para coletar e fornecer informações relevantes para o aprendizado em diferentes fases e contextos. Outro ponto relevante decorrente da análise desses dados é a oportunidade de identificar a desigualdade entre os gêneros (considerando apenas as categorias “feminino” e “masculino”, conforme a variável “Sexo” adotada pelo MEC) nos diferentes estágios de ensino.

Essa diferença de gênero destaca a baixa presença feminina no campo da computação. De acordo com uma pesquisa realizada, houve um aumento significativo de mais de 150% no número total de novos estudantes que começaram o ensino superior em áreas como Processamento de Dados e Tecnologias da Informação entre 2014 e 2022 (Fapesp, 2023). No entanto, ao considerar o fator relacionado ao sexo, nota-se uma significativa falta de representação feminina. Apenas 15% das pessoas que concluíram cursos relacionados à Ciência da Computação e TIC eram do gênero feminino (INEP, 2022).

Essa falta de representatividade também é refletida historicamente, onde muitas mulheres participaram da construção da história da computação, porém os nomes dos homens são sempre os mais citados e, dificilmente, uma pessoa que não é especialista na área tem conhecimento dos nomes e dos papéis que exerceram essas mulheres (Oliveira et al. 2014). Nesse cenário, é notório que a participação das mulheres em ciência e tecnologia tem sido historicamente subestimada e desvalorizada, um reflexo de preconceitos e estigmas de gênero que criam barreiras significativas à sua plena inserção no setor (Maia, 2016). Esse padrão também pode ser observado nas instituições voltadas à formação técnica. O ingresso em cursos técnicos que demandam maior domínio de tecnologias e áreas da engenharia ainda representa um fator de exclusão para mulheres. De forma semelhante ao que ocorre

no Ensino Superior e na Pós-Graduação, a presença feminina em profissões de nível técnico tende a se afastar da área da computação e se concentrar em outras áreas, como serviços pessoais, saúde e educação (Alves, 2016).

Fatores pessoais e comportamentais são apontados como os mais decisivos na escolha de alunas de um curso na área da Computação (Ribeiro et al., 2020). Barreiras como a falta de apoio familiar, os estereótipos e estigmas machistas em torno da área, bem como as dificuldades enfrentadas pelas mulheres nos cursos de ensino superior e no mercado de trabalho, também são mencionadas como elementos que contribuem para o afastamento das jovens da Tecnologia da Informação (TI) (Souza, 2017). Além disso, aspectos sociais e culturais influenciam 'diretamente o interesse feminino pela computação, já que, desde cedo, meninas são menos incentivadas a se envolverem com atividades relacionadas à tecnologia e à lógica, sendo frequentemente direcionadas a áreas vistas como mais “adequadas” ou “femininas” (Castro, 2013).

A presença feminina na produção tecnológica é crucial para garantir um avanço científico e tecnológico equitativo e bem-sucedido, além de promover uma transformação cultural na dinâmica entre gênero e tecnologia (Motogna et al., 2022). Diante disso, é fundamental criar e implementar projetos que conectem as áreas de tecnologia com os interesses das alunas de forma divertida e acessível. Tais iniciativas devem também visar o aumento da autoestima dessas estudantes e a promoção da igualdade de gênero, em especial no ambiente acadêmico, onde um grande potencial foi identificado, mas também um alto grau de insegurança (Medeiros et al., 2022).

Diante desse cenário, o presente estudo busca responder às seguintes perguntas de pesquisa: “Como é caracterizada a participação feminina nos cursos de Computação da Rede Federal de ensino no Brasil e quais são os principais fatores preditivos para a conclusão ou evasão do curso por mulheres no IFPE?” Para responder essas questões, o objetivo deste artigo é analisar a participação feminina nos cursos de computação da rede federal de ensino, além de usar algoritmos de Machine Learning para predizer sucesso/evasão no IFPE. Para atingir esses objetivos, foram analisados os microdados abertos da Plataforma Nilo Peçanha disponibilizados pelo MEC, no período compreendido entre 2017 e 2023.

Com isso, entende-se a relevância deste trabalho no meio científico e acadêmico: garantir um melhor entendimento sobre a participação de mulheres nos cursos da área de tecnologia do IFPE. A análise desse cenário permite não apenas identificar a presença feminina nesses espaços, mas também compreender os fatores que influenciam sua entrada, permanência e conclusão dos cursos. A partir dessas informações, torna-se possível desenvolver ações afirmativas e políticas institucionais mais eficazes, voltadas ao incentivo da participação de mulheres na área da computação.

O restante do artigo foi organizado da seguinte maneira: na seção 2 são apresentados os conceitos básicos, que abordam conceitos pertinentes ao entendimento da pesquisa, na seção 3 é apresentada uma revisão da literatura sobre a participação feminina em cursos da área de Computação; na seção 4 é descrita a metodologia adotada no presente trabalho; na seção 5 são apresentados os resultados obtidos; e, por último, na seção 6 são apresentadas as considerações finais.

2 CONCEITOS BÁSICOS

A compreensão dos conceitos apresentados nesta seção é fundamental para o entendimento da pesquisa, uma vez que este estudo envolve a articulação entre educação, desigualdade de gênero, análise estatística e técnicas de Machine Learning. Cada um desses campos possui especificidades que, quando não devidamente contextualizadas, podem dificultar a interpretação dos resultados. Assim, busca-se aqui apresentar, os fundamentos teóricos e técnicos que servem de base para a investigação realizada.

2.1 Dados Abertos da Plataforma Nilo Peçanha

A Plataforma Nilo Peçanha (PNP) foi instituída com o objetivo de disseminar dados estatísticos de toda a Rede Federal de Educação. Ela surgiu em 2018 pela Secretaria de Educação Profissional e Tecnológica do Ministério da Educação (SETEC/MEC), tendo em vista a necessidade de se melhorar a gestão pública com base em indicadores de desempenho (Moraes et al., 2020). A plataforma é online e de livre acesso, possui um sistema de Business Intelligence (BI) que serve para prover suas informações, contando com painéis interativos que fornecem diversas maneiras de visualizar os dados.

É possível acessar um conjunto de microdados da PNP através do Portal de Dados Abertos do Ministério da Educação. Esses microdados possuem indicadores sobre Eficiência Acadêmica, Matrículas, Servidores e Financeiro, sendo separados por ano, onde podem ser baixados em arquivos no formato Comma-Separated Values (CSV), que é um formato em que os dados são salvos em forma de tabela. Nesses microdados, a variável “Sexo” aparece rigidamente categorizada como “masculino” ou “feminino”, seguindo a estrutura burocrática tradicional dos sistemas educacionais brasileiros. Embora o debate contemporâneo sobre gênero reconheça uma multiplicidade de identidades e vivências, esta pesquisa se alinha à classificação oficial por motivos metodológicos, já que análises quantitativas exigem padronização e consistência entre os anos avaliados.

Para este trabalho, foram utilizados os microdados de matrículas dos anos de 2017 a 2023, que contém referências como unidade de ensino, renda, faixa etária, raça e ciclo de matrícula dos alunos, que nesse caso, foram utilizados microdados relativos aos alunos das instituições que ofereciam cursos na área da Computação e dos diferentes níveis de curso na mesma área.

2.3 Machine Learning

Machine Learning (Aprendizado de Máquina) é uma subárea da IA que se concentra na utilização de algoritmos para desenvolver modelos autônomos capazes de aprender a partir de dados históricos (Homem, 2020). Estes modelos podem identificar padrões, relações e características nos dados, permitindo-lhes realizar tarefas específicas e melhorar seu desempenho com base nos dados históricos analisados.

No Machine Learning os algoritmos podem ter o aprendizado classificado de maneiras diferentes. O presente estudo utiliza o aprendizado supervisionado, no qual os modelos aprendem a partir de exemplos cuja resposta final é conhecida, onde o

algoritmo irá encontrar padrões existentes entre os dados de entrada e a saída esperada que, posteriormente, o tornará capaz de classificar com exatidão um valor ainda não conhecido pelo mesmo (Petersson, 2021).

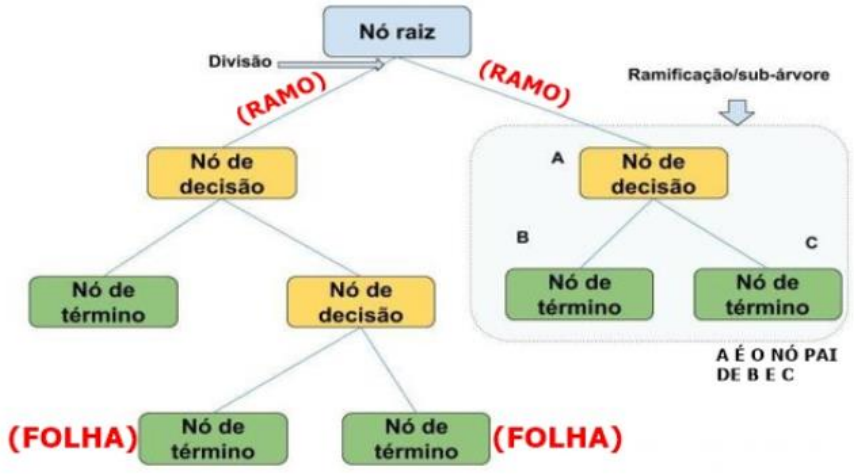
Os modelos preditivos frequentemente aplicam funções de aprendizado supervisionado para estimar valores desconhecidos ou futuros de variáveis dependentes em função das características das variáveis independentes relacionadas. Modelos preditivos têm o objetivo específico que nos permite predizer os valores desconhecidos de variáveis de interesse a partir de valores conhecidos de outras variáveis. O formato da previsão pode ser pensado como um mapeamento de aprendizagem a partir de um conjunto de entradas como um vetor de medições e uma saída como um escalar (Han et al. 2011).

O Aprendizado Supervisionado traz com si diversos algoritmos capazes de realizar tarefas de classificação em conjuntos de dados. Existem diversos algoritmos para este propósito, e nesta seção serão abordados os algoritmos utilizados no trabalho.

2.3.1 Decision Tree

O algoritmo de Decision Tree (Árvore de Decisão) é um modelo que organiza os dados em uma estrutura hierárquica fundamentada em regras lógicas extraídas de uma base de dados rotulada (Alpaydin, 2020). Essa arquitetura é composta por nós que, partindo de uma raiz comum, testam condições específicas ou operações matemáticas predefinidas. O desfecho da predição é encontrado nas extremidades da estrutura, denominadas folhas. Uma das principais virtudes desse método reside em sua transparência e interpretabilidade, permitindo que a lógica por trás da classificação seja facilmente compreendida e analisada por seres humanos (Ramezankhani et al., 2014). A Figura 1 ilustra detalhadamente essa organização estrutural.

Figura 1. Ilustração do funcionamento da Decision Tree.



Fonte: Vieira et al. (2020).

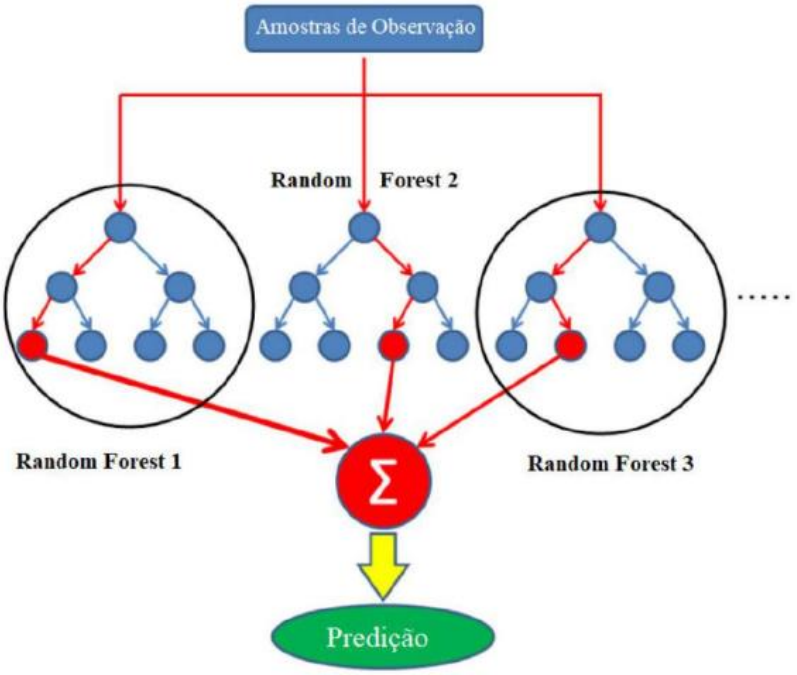
A estrutura de uma Decision Tree está ilustrada na Figura 1, sendo composta por três componentes essenciais. O ponto de partida é a “Raiz”, que representa o atributo com maior poder de distinção para a divisão inicial dos dados. A partir dela, desenvolvem-se os “Nós de Decisão”, que funcionam como ramificações onde são

aplicadas regras lógicas baseadas nas características dos atributos. Finalmente, a estrutura chega ao “Nó de Término” (ou “Folha”), que indica o desfecho da classificação ou a decisão final obtida pelo modelo.

2.3.2 Random Forest

O algoritmo Random Forest (Floresta Aleatória) representa uma evolução técnica da Decision Tree, fundamentando-se na construção de múltiplos modelos de decisão simultâneos. Sua estratégia principal consiste em combinar os resultados dessas diversas árvores para elevar o poder de generalização e a precisão do sistema (Quinlan, 1986). Utilizando métodos estatísticos para realizar inferências indutivas, este modelo de aprendizagem supervisionada é amplamente reconhecido por sua eficácia tanto em tarefas de classificação quanto em análises preditivas (Honda, 2021). A organização estrutural do modelo Random Forest pode ser visualizada na Figura 2.

Figura 2. Ilustração do funcionamento da Random Forest.



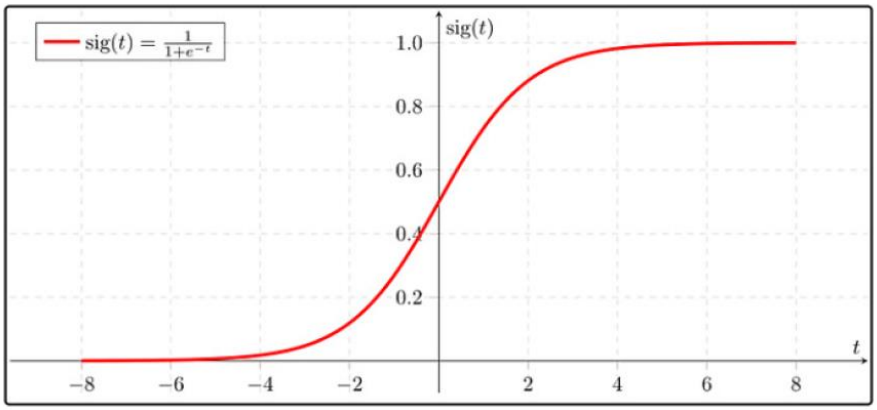
Fonte: James et al. (2013).

Conforme ilustrado na Figura 2, o funcionamento do Random Forest baseia-se na integração de múltiplas árvores de decisão, cada uma desenvolvida a partir de subconjuntos distintos dos dados originais. Nesse processo, cada unidade gera uma predição isolada que, posteriormente, é consolidada por meio de mecanismos de agregação, como a votação majoritária ou o cálculo de médias. Essa convergência de resultados resulta em uma predição final coletiva, conferindo ao modelo uma robustez e uma precisão superiores às de uma árvore de decisão sozinha.

2.3.3 Logistic Regression

O modelo de Logistic Regression (Regressão Logística) é amplamente aplicado em problemas de classificação binária. Conforme aponta Hilbe (2011), quando a variável dependente assume apenas dois estados, ela é comumente codificada como 0 ou 1, onde 1 geralmente representa a ocorrência do evento de interesse e 0 a sua ausência. No contexto deste estudo, a classificação é binária, definindo-se "concluinte" como 0 e "evadido" como 1. Batista (2015) complementa essa definição ao descrever a Logistic Regression como um método estatístico voltado ao cálculo da probabilidade de ocorrência de um evento específico, baseando-se em um conjunto de variáveis independentes. A organização lógica desse modelo está representada na Figura 3.

Figura 3. Ilustração do funcionamento da Logistic Regression.



Fonte: Swaminathan (2018)

A Logistic Regression (LR) se assemelha com o modelo de Regressão Linear, utilizado para prever valores numéricos contínuos. Contudo, na LR, os resultados são transformados em probabilidades através da função Sigmoide, ilustrada na Figura 3. Esta função é caracterizada por seu formato em "S" e por garantir que a saída gerada esteja invariavelmente compreendida no intervalo entre 0 e 1.

3 TRABALHOS RELACIONADOS

A pesquisa sobre a participação feminina na área de tecnologia da informação e ciência da computação tem sido amplamente discutida, com diversos estudos investigando as causas e os resultados da baixa representatividade de mulheres nesse campo.

O estudo de Marinho et al. (2019), analisou a participação de mulheres em cursos de Informática na Rede Federal de Educação Profissional, Científica e Tecnológica. O estudo, que utilizou dados da Plataforma Nilo Peçanha, no ano de 2017, constatou que as mulheres são minoria tanto em cursos superiores de bacharelado e tecnologia quanto em cursos técnicos, embora sua participação seja maior nestes últimos. No Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (Cefet/RJ), campus Nova Friburgo, foi analisado o período referente ao segundo semestre de 2018, e notou-se que a proporção de mulheres no curso de Sistemas de Informação

foi de apenas 5%, enquanto a de formandas no curso técnico integrado foi superior à de homens.

Uma análise estatística dos dados do Censo da Educação Superior de 2009 a 2018 foi realizada para verificar a inserção de mulheres em cursos de TI no Brasil. O estudo revelou que a participação feminina em cursos de tecnologia caiu de 20,14% em 2009 para 15,07% em 2018. A análise de regressão mostrou uma forte correlação negativa entre os anos e a porcentagem de mulheres ingressantes, enquanto a porcentagem de homens seguiu em crescimento. As autoras reforçam que, sem intervenções, essa tendência de queda na presença feminina deve continuar (Cursino e Martinez, 2021).

Já a pesquisa de Santos et al. (2025) analisou a participação feminina nos cursos superiores de Computação do IFPE entre 2009 e 2022, utilizando os microdados educacionais do Censo da Educação Superior. Os resultados indicam que o número de mulheres ingressantes, matriculadas e concluintes é inferior ao de homens. No entanto, a participação feminina aumentou após a expansão dos cursos para novos campi em 2019. Apesar disso, as taxas de conclusão são baixas para ambos os gêneros, sendo de 13% para mulheres e 17,7% para homens no Campus Recife.

A representação feminina em cursos técnicos integrados ao ensino médio do Instituto Federal de Mato Grosso (IFMT) entre 2010 e 2019 foi investigada. Os resultados revelaram que a representatividade feminina foi inferior à masculina nos cursos analisados, como Agrimensura, Edificações, Eletroeletrônica e Informática. No entanto, as mulheres apresentaram taxas de evasão menores e taxas de conclusão maiores do que os homens (Paiva e Silva, 2021). Ainda no contexto de institutos federais, a participação de mulheres nos cursos técnico em informática e superior em Ciência da Computação no Instituto Federal do Sudeste de Minas Gerais (IF Sudeste MG) foi analisada. O estudo identificou que poucas alunas de cursos técnicos optavam por seguir para o ensino superior na área. No curso de Ciência da Computação, apenas 14,25% dos ingressantes entre 2007 e 2019 eram mulheres, e menos da metade delas concluíram o curso (Pereira et al., 2021).

Por último, Santos et al. (2021), analisaram a participação de mulheres em cursos superiores de TI no Brasil, utilizando dados do Censo da Educação Superior de 2014 a 2019. A pesquisa confirmou a baixa representatividade feminina na área, com as mulheres compondo apenas 13,8% dos estudantes e 15,2% dos formados no período analisado. A maior concentração de mulheres foi encontrada na região Sudeste. O estudo também notou uma mudança de preferência, com o curso de Análise e Desenvolvimento de Sistemas superando o de Sistemas de Informação em número de matrículas femininas a partir de 2018.

Diante deste cenário, o presente artigo busca preencher a lacuna existente na literatura ao investigar a evolução da participação feminina nos cursos de diferentes níveis de Computação do IFPE, uma temática ainda pouco explorada em estudos anteriores. Embora haja pesquisas sobre a presença feminina na área de Computação em diferentes contextos educacionais, não há análises detalhadas sobre como esse fenômeno se manifesta no IFPE ao longo dos anos.

4 METODOLOGIA

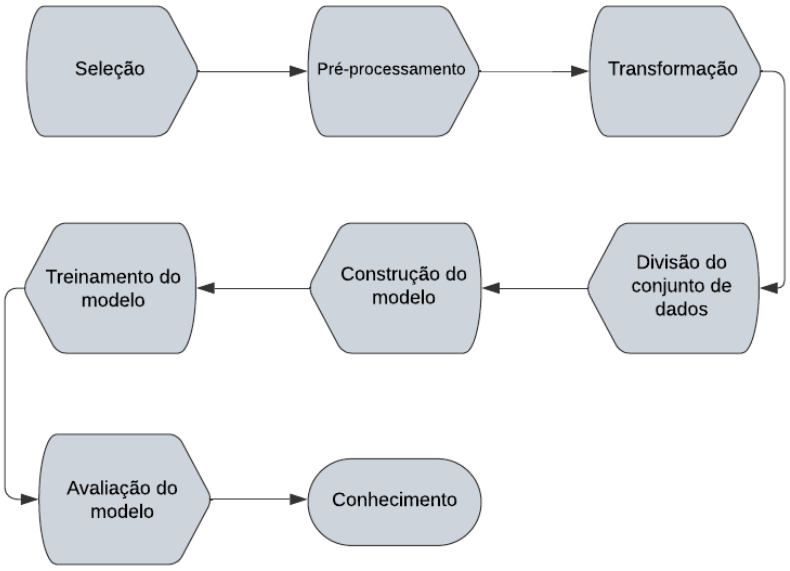
Este artigo científico, está baseado na metodologia de Nascimento (2016), dada as classificações da pesquisa. Quanto à abordagem, foi utilizado o método quantitativo, que emprega medidas padronizadas e sistemáticas, reunindo respostas pré-determinadas, facilitando a comparação e a análise de medidas estatísticas de dados. Quanto à natureza, foi utilizada a pesquisa aplicada, que é definida por dirigir à busca da verdade para determinada aplicação prática em situação particular. O estudo se baseia também no método descritivo, pois busca a descrição de características de populações ou fenômenos e de correlação entre variáveis. Dado o cenário apresentado anteriormente, foi realizado um estudo de caso, que é um estudo de certo caso singular visando a descoberta de fenômenos em determinado contexto. Enfatiza a interpretação de fenômeno específico e busca retratar a realidade de maneira complexa e profunda (Nascimento, 2016).

A metodologia adotada para a análise de dados das matrículas dos alunos envolveu o uso do Power BI[1], uma ferramenta de visualização de dados que facilitou a construção de relatórios visuais detalhados. Foram analisados os microdados de matrículas dos anos de 2017 à 2023, os anos existentes no site disponibilizado pela Plataforma Nilo Peçanha. Além disso, foi conduzida uma análise preditiva (para prever tendências e padrões) utilizando Machine Learning, por meio da linguagem de programação Python, aplicando bibliotecas especializadas para manipulação, tratamento e visualização de dados, como Pandas, Matplotlib e Seaborn. Foram treinados três modelos de predição: Decision Tree, Random Forest e Logistic Regression. Também foi utilizada a plataforma do Google Colab[2], para executar os códigos necessários. O processo de tratamento dos dados extraídos foi dividido em algumas etapas, ilustrado na Figura 4. Cada uma dessas etapas desempenhou um papel fundamental na criação dos dashboards (painel de controle visual) e previsões que sintetizam as informações de maneira clara e compreensível.

Figura 4 - Desenho metodológico da pesquisa

[1] https://www.microsoft.com/pt-br/power-platform/products/power-bi

[2] https://colab.research.google.com/



Fonte: elaborado pela autora.

A análise de dados teve início com a seleção das bases citadas anteriormente, abrangendo informações detalhadas sobre alunos de diversas instituições de ensino (disponibilizado o link com os datasets[3]). Especificamente, focou-se na análise das colunas que apresentavam dados de sexo, renda familiar, faixa etária, cor/raça e a situação de matrícula dos estudantes. Essa seleção visava compreender como essas variáveis se relacionam entre si e impactam a situação acadêmica dos alunos. Ao correlacionar sexo com renda familiar, faixa etária, cor/raça e situação de matrícula, buscou-se identificar padrões e possíveis desigualdades presentes nas matrículas dos cursos analisados.

A etapa de pré-processamento dos dados envolveu uma filtragem cuidadosa das informações para garantir a relevância e a precisão da análise. Inicialmente, selecionaram-se apenas as colunas que continham os dados mencionados anteriormente: sexo, renda familiar, faixa etária, cor/raça e situação de matrícula. Em seguida, os dados foram filtrados especificamente para os cursos de computação, de modo a focar a análise em uma área específica do ensino. Além disso, foram incluídos todos os campi existentes na base que ofertavam cursos na área de computação, para proporcionar uma visão abrangente das instituições. Durante essa fase, tratou-se também de valores ausentes e inconsistências, garantindo que o conjunto de dados estivesse limpo e pronto para a análise subsequente. Esse processo de filtragem e limpeza é essencial para assegurar que os resultados obtidos sejam confiáveis e significativos.

A transformação dos dados foi uma etapa crucial para facilitar a análise e a visualização das informações. Uma das transformações principais foi a padronização da coluna de sexo, onde todos os registros foram uniformizados para "masculino" ou "feminino", pois os dados estavam escritos de formas diferentes entre os anos. Essa padronização visou simplificar a categorização e análise dos dados, evitando

[3] https://bit.ly/drive-datasets

discrepâncias causadas por variações nos registros. Além disso, foram excluídas colunas que não tinham relevância para a análise, como por exemplo, informações sobre o código da unidade, carga horária do curso, código de município e quantidade de vagas, com o objetivo de reduzir a complexidade do conjunto de dados e focar nas informações mais pertinentes. Essa eliminação de dados desnecessários ajudou a melhorar a clareza e a eficiência da análise, permitindo uma compreensão mais direta e eficaz dos resultados apresentados.

Após as 3 primeiras etapas explicadas anteriormente, que englobam as duas análises feitas no artigo, vieram as etapas necessárias para análise preditiva com Machine Learning em específico:

A etapa seguinte, a Divisão do Conjunto de Dados, foi essencial para a preparação dos dados para o Machine Learning. O conjunto total de dados utilizados para a análise foi classificado pela categoria da situação (concluintes ou evadidos) e por gênero, os números de cada um podem ser observados no Quadro 1. Esses dados foram processados e transformados, sendo segmentados em duas partes distintas: dados de treinamento (80%) e dados de teste (20%). Essa separação rigorosa é crucial para garantir que os modelos de Machine Learning sejam avaliados de forma justa, utilizando dados que nunca viram antes. O subconjunto de treinamento foi utilizado para o ajuste dos modelos, enquanto o subconjunto de teste serviu exclusivamente para medir o desempenho final.

Quadro 1 - Quantidade total por gênero e situação de matrícula.

Situação do aluno	**Feminino**	**Masculino**
Concluintes	1.279	1.853
Evadidos	1.722	3.385

Fonte: Elaborado pela autora.

Em seguida, deu-se a Construção do Modelo, onde foram escolhidas e definidas as arquiteturas de Machine Learning a serem empregadas para a classificação da situação de matrícula dos estudantes (Concluinte ou Evadido). Três algoritmos distintos foram selecionados para comparação: Decision Tree, Random Forest e Logistic Regression. Essa variedade de modelos — abrangendo desde técnicas mais simples e lineares até métodos baseados em árvores e conjuntos (ensembles) — permitiu uma avaliação robusta sobre qual abordagem se adaptaria melhor à complexidade e aos padrões inerentes aos dados demográficos e acadêmicos.

A etapa de Treinamento do Modelo envolveu o ajuste de cada arquitetura selecionada (Decision Tree, Random Forest e Logistic Regression) utilizando-se, exclusivamente, o conjunto de dados de treinamento. Durante este processo, os modelos aprenderam as relações entre as variáveis de entrada (gênero, renda, cor/raça, etc.) e a variável de saída (situação da matrícula). Além disso, foi realizada a otimização dos hiperparâmetros de cada modelo para maximizar seu potencial preditivo, garantindo que o aprendizado fosse o mais eficiente possível.

Posteriormente, na Avaliação do Modelo, o desempenho de cada modelo treinado foi minuciosamente testado utilizando o conjunto de dados de teste, anteriormente reservado. Para quantificar a eficácia de cada modelo na previsão da situação de matrícula, foram obtidas e analisadas métricas de avaliação padrão: acurácia (proporção de previsões corretas), precisão (capacidade de evitar falsos positivos), F1-score (média harmônica entre precisão e recall) e recall (capacidade de evitar falsos negativos). A análise comparativa dessas métricas permitiu identificar qual modelo obteve o melhor desempenho na tarefa de classificação.

Por fim, a fase de Conhecimento consistiu na análise aprofundada dos resultados obtidos na avaliação do modelo de melhor desempenho. Esta etapa não se limitou apenas à verificação das métricas, mas envolveu a interpretação das características do modelo, como a importância das variáveis (gênero, renda, cor/raça, etc.) na tomada de decisão do algoritmo. A partir dessa interpretação, foi possível extrair insights valiosos sobre os fatores demográficos e socioeconômicos que mais influenciam a permanência ou a evasão das estudantes na área de Computação, resultando nas discussões e conclusões do artigo.

5 RESULTADOS E DISCUSSÕES

Esta seção do artigo mostra os resultados e a discussão das duas abordagens metodológicas realizadas. Primeiramente, a análise descritiva utilizando o Power BI, em seguida, a seção incorpora a análise preditiva, apresentando os resultados da aplicação de modelos de Machine Learning para quantificar o risco de insucesso e a probabilidade de conclusão dos alunos, validando a influência das variáveis de perfil sobre a trajetória acadêmica.

5.1 Análise Com a Ferramenta Power BI

Esta seção do trabalho apresenta a análise detalhada utilizando o Power BI, dos dados quantitativos de matrícula, conclusão e evasão das mulheres em todos os cursos de diferentes níveis da área de Computação da Rede Federal de ensino do Brasil, trazendo também os números do IFPE de forma comparativa, utilizando como base os microdados da PNP. Além de expor as séries históricas e as medidas estatísticas de desempenho (taxas das matrículas em curso, conclusão e evasão) por gênero, a análise revela os perfis sociodemográficos (raça/cor, renda familiar, faixa etária) das discentes. O dashboard mais detalhado está disponível no link[4].

Vinculada ao Ministério da Educação e integrante do sistema federal de ensino, a Rede Federal é composta por diferentes instituições: os 38 Institutos Federais de Educação, Ciência e Tecnologia (IFs); a Universidade Tecnológica Federal do Paraná (UTFPR); os Centros Federais de Educação Tecnológica Celso Suckow da Fonseca, no Rio de Janeiro (Cefet-RJ), e de Minas Gerais (Cefet-MG); as 22 escolas técnicas associadas às universidades federais; e o Colégio Pedro II (CPII). No total, estas instituições contabilizam 686 unidades de ensino público federal distribuídas nos 26 estados do Brasil e no Distrito Federal. Elas são responsáveis pelo ensino em diferentes modalidades, desde o ensino técnico integrado ao ensino médio até o doutorado. Uma lista dos cursos analisados é apresentada no Quadro 2.

[4] https://bit.ly/dash-tcc

Quadro 2 - Cursos da Rede Federal analisados.

Nível	Lista de cursos
Técnico	Computação Gráfica; Desenvolvimento de Sistemas; Informática; Informática para Internet; Manutenção e Suporte em Informática; Programação de Jogos Digitais; Redes de Computadores.
Tecnológico	Agrocomputação; Análise e Desenvolvimento de Sistemas; Gestão da Tecnologia da Informação; Jogos Digitais; Redes de Computadores; Sistemas para Internet.
Pós-graduação	Administrador de Banco de Dados; Agente de Inclusão Digital em Centros Públicos de Acesso à Internet; Desenvolvedor de Aplicativos para Mídias Digitais; Desenvolvedor de Jogos Eletrônicos; Especialização - Informação e Comunicação; Informática.
Mestrado	Mestrado e Mestrado Profissional em Informação e Comunicação.
Doutorado	Doutorado em Informação e Comunicação.

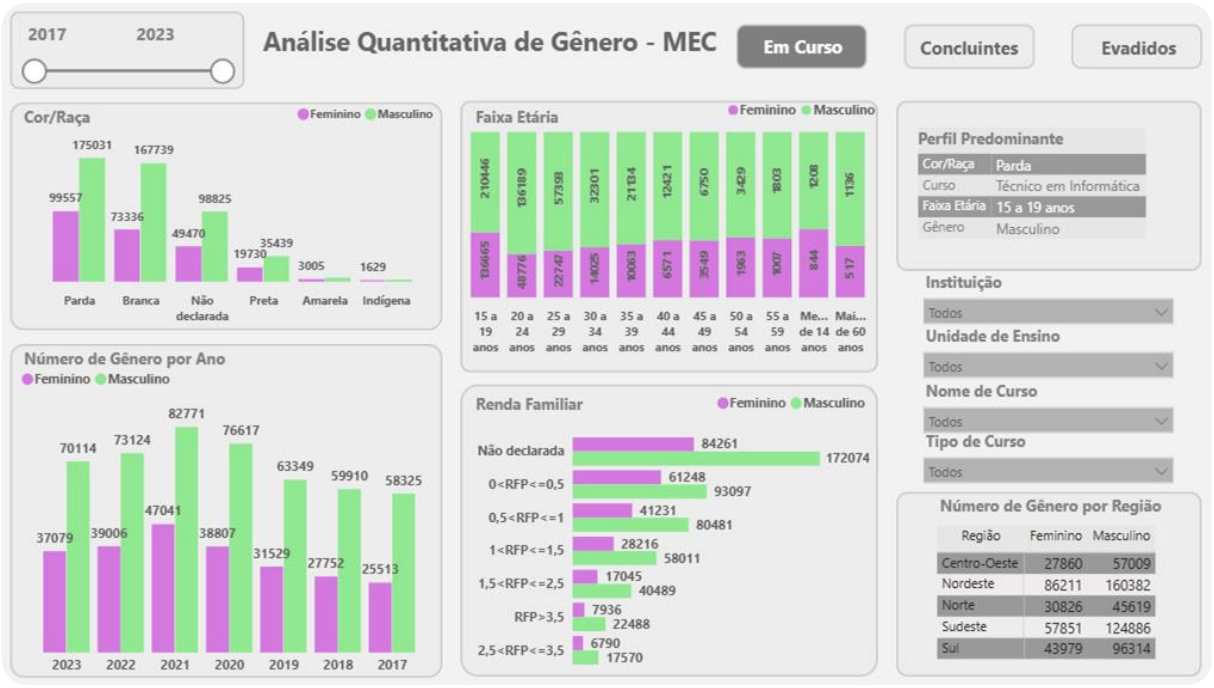
Fonte: Elaborado pela autora.

Os dados obtidos foram categorizados em três grupos: estudantes Em Curso, Concluintes e Evadidos. Para cada categoria, foram analisadas as variáveis de gênero, faixa etária, cor/raça, renda familiar e região.

5.1.1 Em Curso

A Figura 5 apresenta os resultados referentes ao número de estudantes “Em Curso”, de forma geral. Os dados desse painel revelam uma discrepância de gênero, com um número significativamente maior de homens “Em Curso” em comparação com as mulheres ao longo do período analisado. Em 2017, as mulheres representavam 30,43% do total de matriculados ativos (25.513 mulheres de 83.838 estudantes). Essa porcentagem cresceu de forma consistente nos anos seguintes, alcançando seu ponto máximo de 36,24% em 2021 (47.041 mulheres de 129.812 estudantes). Após esse pico, houve uma ligeira redução na participação, com a porcentagem estabilizando em 34,59% em 2023. Ao todo, a média de mulheres matriculadas no período de 2017 a 2023 é de aproximadamente 37,23%, o que ainda representa uma grande lacuna em relação aos 62,77% de homens.

Figura 5. Painel com dados para os estudantes “Em Curso”.

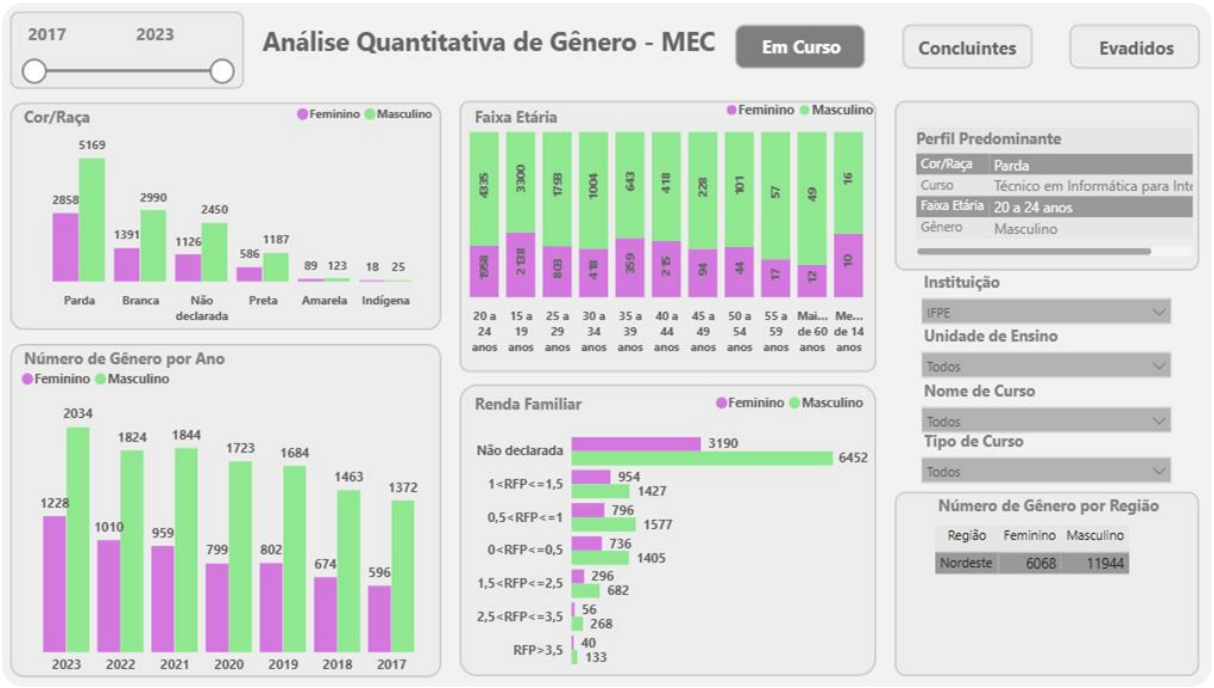


Fonte: Elaborado pela autora.

A maior concentração de matrículas femininas ocorre na faixa etária de 15 a 19 anos com 136.665 alunas (51,23% ao longo dos anos analisados, indicando que a principal via de ingresso para as mulheres na tecnologia é o ensino técnico subsequente ou superior logo após a educação básica. Em relação à cor/raça, as mulheres autodeclaradas pardas (99.557 representando 37,32% do total de mulheres) são o grupo mais numeroso. Este dado é essencial, pois sugere que as ações de inclusão e o interesse pela área foram particularmente eficazes na atração de mulheres Pardas. Finalmente, o perfil socioeconômico é marcado por uma forte presença de estudantes de baixa renda, com a maioria das matrículas femininas concentradas nas faixas de renda familiar per capita de até 1 salário mínimo, demonstrando que o crescimento da participação feminina está intrinsecamente ligado à ampliação do acesso e das oportunidades para classes menos favorecidas.

A trajetória das matrículas femininas nos cursos de tecnologia do IFPE seguiu um padrão de expansão constante, com um ponto de inflexão nos últimos anos, conforme mostrado na Figura 6. Em 2017, o IFPE registrava 596 alunas "Em Curso", representando 30,28% do total. Assim como na tendência nacional, a participação feminina aumentou continuamente, atingindo o pico de 1.228 matrículas em 2023 (contra 596 em 2017). No entanto, em termos percentuais, a maior representatividade feminina ocorreu em 2023, com 37,64% (1.228 mulheres de 3.262 estudantes), superando o pico percentual nacional (36,24% em 2021). Esse resultado indica que o IFPE tem conseguido, em anos recentes, atrair e manter uma proporção de mulheres superior à média nacional. O Perfil Predominante das matrículas no IFPE, ao longo dos anos, manteve-se consistentemente masculino, concentrado em estudantes Pardos e nas faixas etárias de 15 a 24 anos (variando entre 15-19 e 20-24, dependendo do ano), o que alinha a instituição ao perfil demográfico geral.

Figura 6. Painel com dados para os estudantes “Em Curso” no IFPE.



Fonte: Elaborado pela autora.

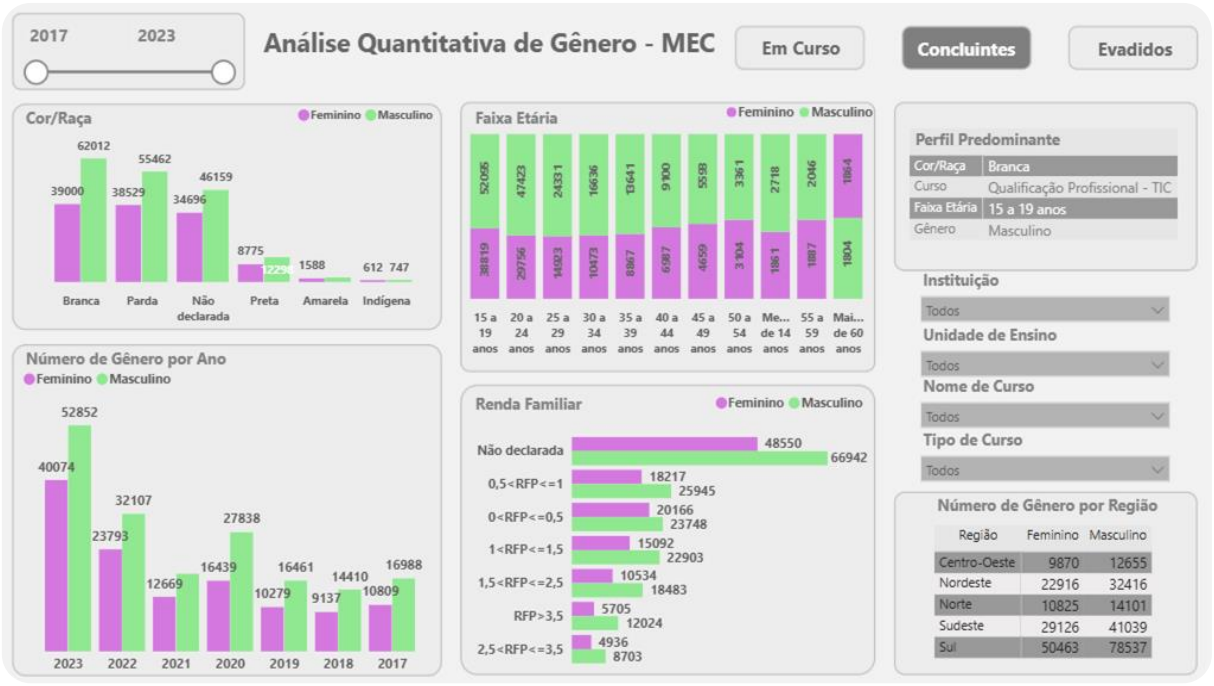
Ao confrontar o cenário do IFPE com o panorama nacional, notam-se importantes contrastes na dinâmica de crescimento e no perfil demográfico. Enquanto o Brasil atingiu o pico de matrículas femininas em 2021 (47.041) e enfrentou uma contração nos anos subsequentes (queda de 21,18% até 2023), o IFPE demonstrou uma expansão contínua, atingindo seu pico percentual mais tarde, em 2023 (1.228 matrículas e 37,64% de participação). Essa persistência no crescimento no IFPE sugere uma maior resiliência da instituição ou o sucesso de políticas internas que conseguiram contrariar a desaceleração observada no cenário geral após 2021.

Em termos demográficos, o IFPE também se destaca na inclusão de raça, embora o perfil nacional mostra uma forte presença de mulheres Pardas e Brancas, o IFPE, em anos como 2023, registrou um volume significativo de mulheres Pretas (239) e Não Declaradas (257), superando em proporção às tendências nacionais. A Renda Familiar no IFPE reforça o padrão de inclusão socioeconômica, com grande concentração na categoria Não Declarada e nas faixas de baixa renda.

5.1.2 Concluintes

No cenário geral, o número de concluintes masculinos é superior, mas a performance de conclusão das mulheres é um ponto de destaque. A porcentagem de mulheres concluintes teve seu ponto mínimo em 2021, com 30,73%, mas demonstrou uma recuperação nos anos seguintes, atingindo seu pico histórico de 43,12% em 2023 (40.074 mulheres em um total de 92.926 concluintes). Esse crescimento elevado, que levou a porcentagem feminina de conclusão a superar a porcentagem de matrícula ativa, sugere que as mulheres que conseguem superar as barreiras do curso tendem a ter uma taxa de sucesso acadêmico superior. A média acumulada de mulheres concluintes no período é de aproximadamente 38,89% (Figura 7).

Figura 7. Painel com dados para os estudantes “Concluintes”.

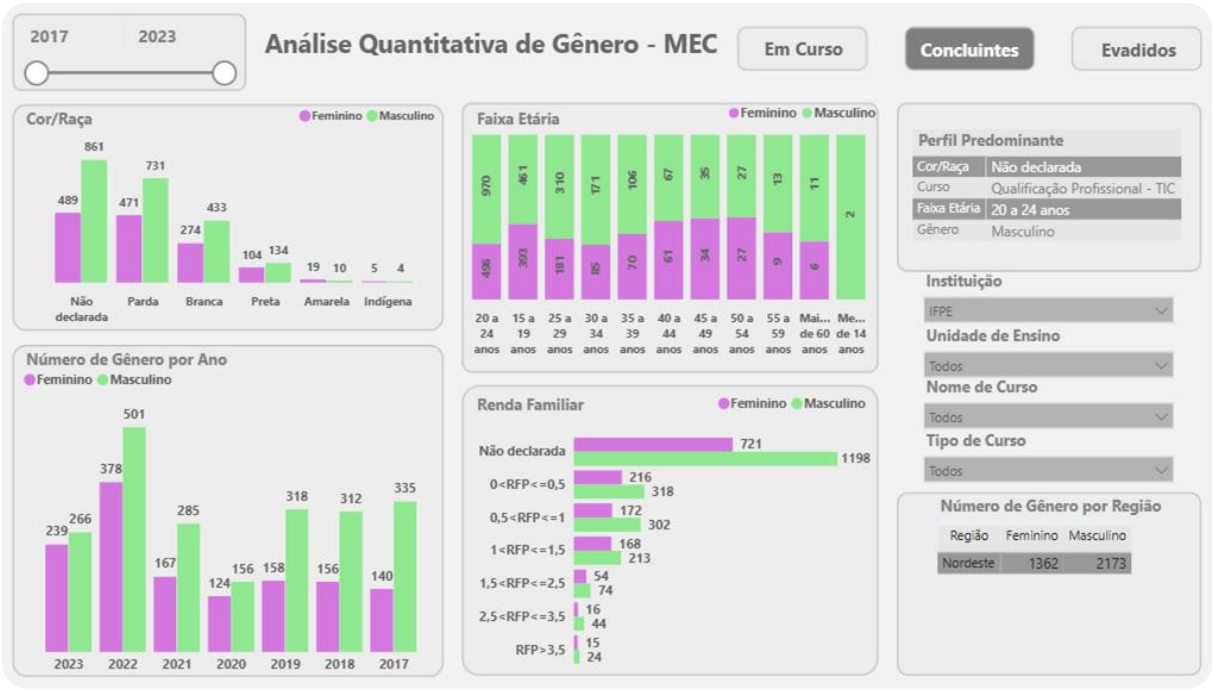


Fonte: Elaborado pela autora.

Em relação ao perfil demográfico das formandas, observa-se uma diferença importante em relação ao grupo "Em Curso". Enquanto as mulheres pardas lideram a matrícula ativa, as mulheres brancas são o grupo mais numeroso entre os concluintes (29,36% do total de formandas), seguidas pelas pardas (29,01%). Essa inversão sugere que, apesar do sucesso no ingresso das mulheres pardas, elas podem enfrentar mais desafios para a conclusão, indicando uma disparidade nas taxas de permanência entre os grupos raciais. O perfil etário é concentrado, com as formandas majoritariamente nas faixas de 15 a 19 anos (29,23%) e 20 a 24 anos (22,40%). Do ponto de vista socioeconômico, a alta incidência na categoria de faixa de renda Não Declarada representa 26,12% das concluintes (34.696).

A trajetória das conclusões femininas no IFPE, mostrado na Figura 8, embora com números menores que os de “Em Curso”, revela uma eficiência notável na formação. Em 2017, 140 mulheres concluíram os cursos, representando 29,47% do total. O IFPE registrou variações anuais, com uma queda no número de 2022 para 2023 (de 378 para 239) e uma baixa percentual em 2020 (44,28% do total). No entanto, o ponto mais alto de conclusão se deu em 2022, com 378 formandas. Em termos percentuais, a participação feminina de concluintes superou a barreira dos 40% em três dos sete anos analisados, demonstrando que, em média, o IFPE forma uma proporção de mulheres superior à tendência de matrícula nacional. O Perfil Predominante dos concluintes no IFPE variou, mas frequentemente se concentrou no gênero masculino, com Cor/Raça Parda ou Não Declarada, e na faixa etária de 20 a 24 anos, em cursos como Técnico em Informática ou Qualificação Profissional (um tipo de educação profissional de curta duração que prepara o indivíduo para o exercício de uma função específica no mercado de trabalho).

Figura 8. Painel com dados para os estudantes “Concluintes” no IFPE.



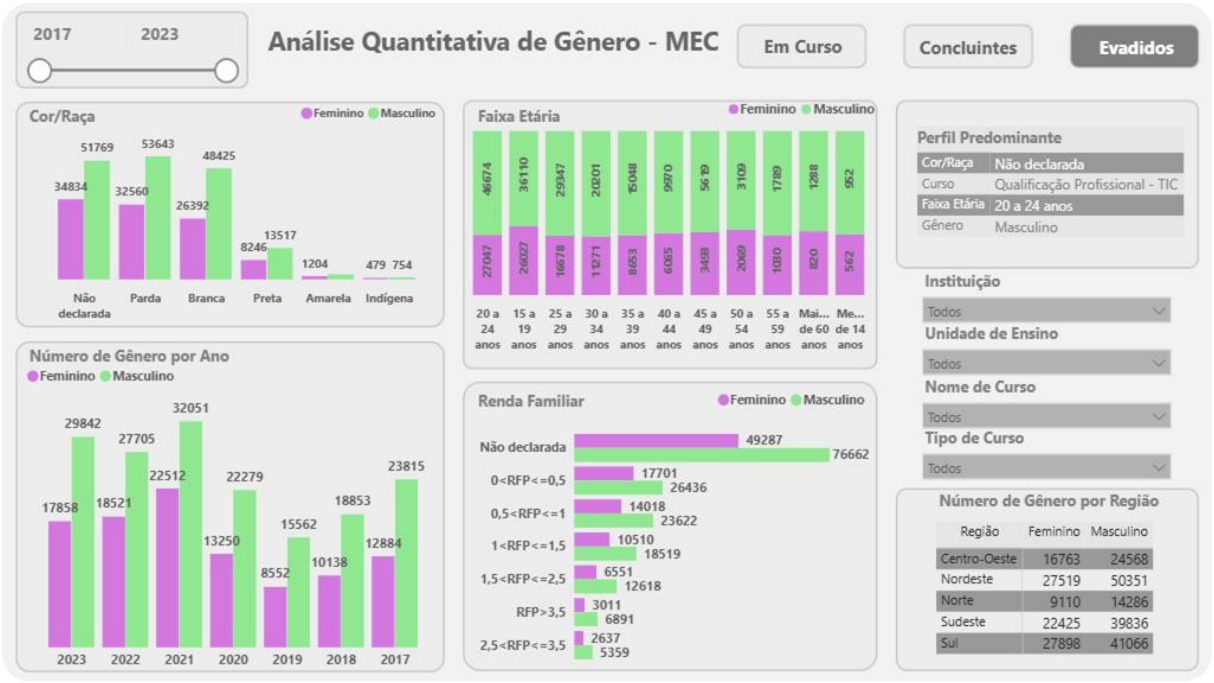
Fonte: Elaborado pela autora.

Ao comparar a conclusão no IFPE com o panorama nacional, observa-se que a instituição se destaca por alcançar taxas de participação feminina percentualmente mais elevadas do que a média geral em diversos anos, atingindo picos como 44,28% em 2020, enquanto a média nacional teve seu maior percentual em 43,12% apenas em 2023. Essa alta proporção de mulheres formadas no IFPE, apesar do volume menor, indica uma maior eficiência na retenção e sucesso acadêmico das alunas. Em termos de perfil, o IFPE apresenta um cenário racial distinto: enquanto a tendência nacional mostrou que as mulheres Brancas lideravam as conclusões, no IFPE, a liderança se alternou entre mulheres Pardas e Não Declaradas, que juntas somam a maioria das formandas. Este dado sugere que o IFPE tem sido mais bem-sucedido em converter o ingresso de mulheres de grupos raciais vulneráveis em diplomas.

5.1.3 Evadidos

O número total de mulheres evadidas no cenário nacional, no período de 2017 a 2023 foi de 37,87% (Figura 9). Diferentemente dos estudantes Em curso e Concluintes, a evasão feminina apresentou um comportamento mais volátil. O ponto mais crítico da série histórica foi atingido em 2021, com 22.512 desistências (representando 41,26%), um crescimento de 69,90% em relação ao ano anterior.

Figura 9. Painel com dados para os estudantes “Evadidos”.

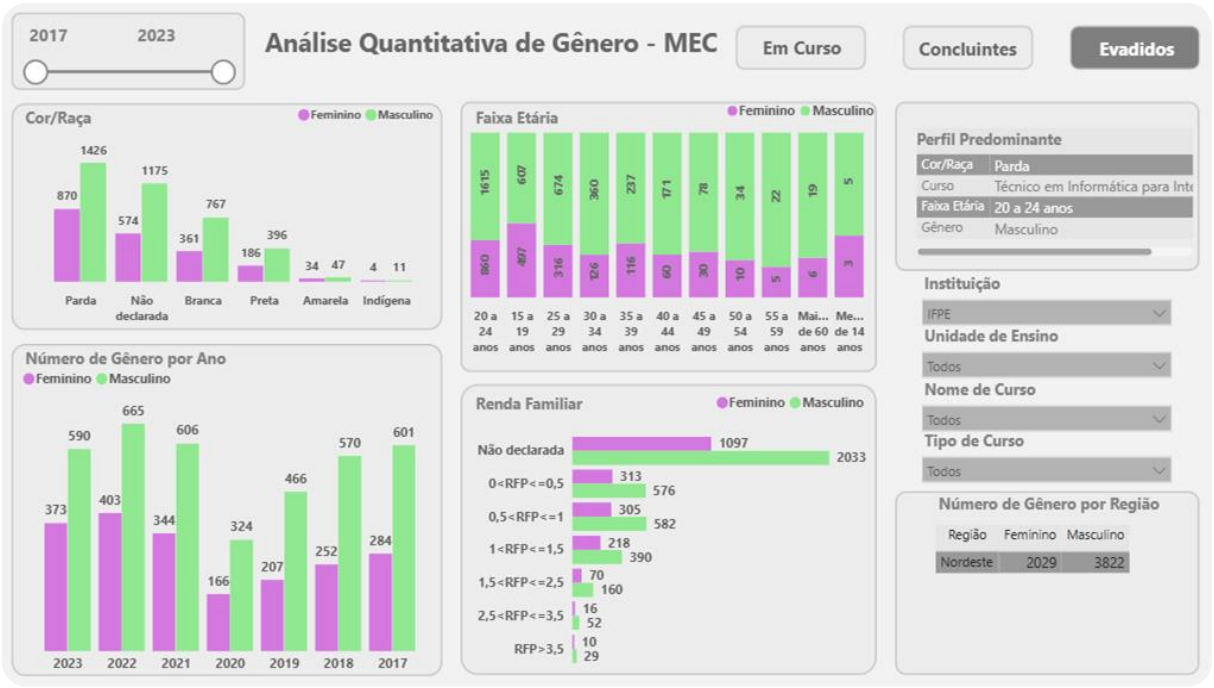


Fonte: Elaborado pela autora.

A evasão é liderada pela categoria de Renda Familiar Não Declarada (que acumula 47,52%), seguida pela faixa de 0 < RFP ≤ 0,5 salário mínimo (com 17,07%). Essa concentração nas faixas de baixa renda sugere que as dificuldades financeiras ou a falta de apoio institucional adequado são possíveis fatores para a interrupção da trajetória acadêmica feminina. Em relação à Cor/Raça, as categorias Não Declarada (33,59%) e Parda (31,39%) lideram as desistências, indicando que esses grupos raciais, embora importantes no ingresso, são os mais vulneráveis à evasão. O abandono concentra-se nas faixas etárias de 15 a 19 anos (25,09%) e 20 a 24 anos (26,08%), sendo necessário aplicação de políticas de acolhimento e retenção especialmente eficazes nos primeiros anos de curso para garantir que o ingresso de mulheres, notavelmente as de baixa renda e grupos raciais minoritários, se converta em conclusão.

A evasão feminina nos cursos de tecnologia do IFPE, demonstrada na Figura 10, seguiu de perto a tendência nacional, com o número de desistências em correlação com os picos de ingresso. O ponto de maior evasão feminina no IFPE ocorreu em 2022, com 403 evadidas, o que representou 37,79% do total. Assim como no cenário geral, o IFPE também registrou um alto volume de evasão em 2021 (344 evadidas), confirmando que a retenção é o maior desafio institucional logo após a atração e matrícula. O Perfil Predominante das evadidas no IFPE está concentrado em estudantes Pardos ou Não Declarados, na faixa etária de 20 a 24 anos, em cursos como Técnico em Informática ou Qualificação Profissional. A análise acumulada reforça a vulnerabilidade, com o número de mulheres Pardas (870 acumuladas) sendo o mais alto entre as evadidas, seguido pelas Não Declaradas (574).

Figura 10. Painel com dados para os estudantes “Evadidos” no IFPE.



Fonte: Elaborado pela autora.

A análise da evasão feminina no IFPE (2017–2023) espelha as características observadas no cenário nacional, confirmando que os desafios de permanência são sistêmicos. Embora o país tenha atingido o pico de evasão absoluta em 2021, o IFPE viu seu pico ocorrer no ano seguinte, em 2022 (403 evadidas), mantendo a taxa percentual de evasão feminina próxima à média nacional (cerca de 37%). A principal semelhança reside no perfil de vulnerabilidade, onde em ambos os cenários a evasão é fortemente concentrada em mulheres Pardas e Não Declaradas, grupos que, apesar de ingressarem em grande volume, são os mais propensos à desistência. Em síntese, os dados do IFPE são um reflexo da realidade brasileira, ressaltando que o sucesso na atração de mulheres para a tecnologia deve ser urgentemente complementado por políticas de retenção socioeconômica eficazes para evitar a perda de talentos.

5.2 Análise Preditiva

A avaliação dos modelos de classificação treinados demonstrou que as arquiteturas baseadas em árvores de decisão apresentaram um desempenho geral ligeiramente superior à Logistic Regression na tarefa de predição da situação final do aluno. Os modelos convergem para uma Acurácia próxima de 70%.

O modelo Random Forest obteve a maior Acurácia (70,32%) e o maior Recall (84,67%) para a classe Não Concluiu no conjunto de testes. O alto Recall é um indicador crucial de que o modelo é mais eficaz em identificar corretamente os casos de insucesso acadêmico (evasão), minimizando os falsos negativos. Por sua vez, a Logistic Regression e o Decision Tree ficaram com acurácias ligeiramente inferiores (69,67% e 69,43%, respectivamente). O desempenho geral das três arquiteturas (com F1-score na faixa de 77-78% para a classe Não Concluiu) sugere que, embora as variáveis demográficas sejam relevantes, elas estabelecem um limite preditivo, indicando que a inclusão de features adicionais (como desempenho acadêmico, notas

ou fatores socioeconômicos mais detalhados) poderia ser necessária para aumentar significativamente a capacidade de previsão. Ilustrado no Quadro 3.

Quadro 3 - Avaliação dos modelos treinados.

	Acurácia	F1-score	Recall	Precisão
Random Forest	70,32%	77,96%	84,67%	72,23%
Decision Tree	69,43%	77,15%	83,24%	71,89%
Logistic Regression	69,67%	77,25%	83,04%	72,21%

Fonte: Elaborado pela autora.

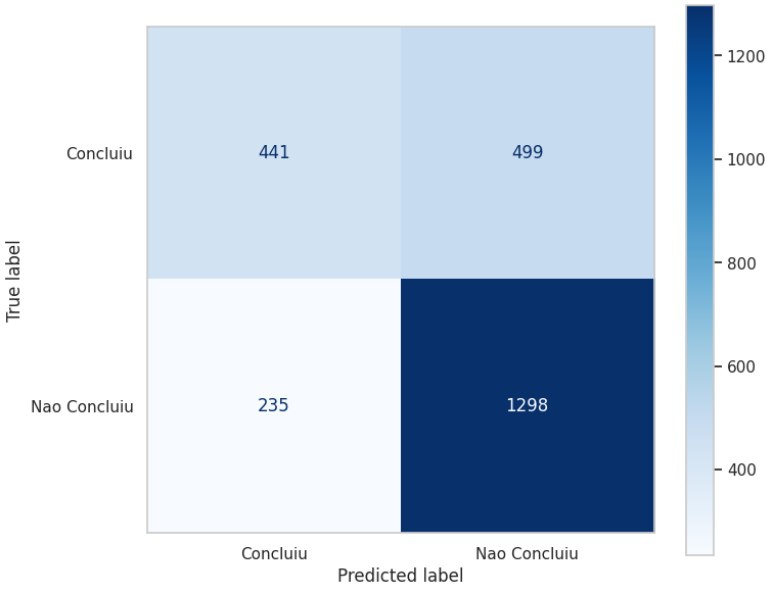
A aplicação dos modelos em um cenário simulado de alta vulnerabilidade, definido pelo perfil (Cor/Raça: Parda; Renda Familiar: 1 < RFP ≤ 1,5; Curso: Técnico; Turno: Matutino; Faixa Etária: 20-24 anos), resultou em uma convergência preditiva: todos os modelos indicaram 'Não Concluiu' como a situação final mais provável. As probabilidades de insucesso variaram entre 62,54% (Decision Tree e Logistic Regression) e 65,73% (Random Forest). Este resultado reitera que a combinação de baixa renda e raça Parda constitui um fator de risco potente, confirmando a necessidade de intervenções focadas neste grupo.

Na comparação de gênero para este perfil, a diferença nas probabilidades de evasão foi marginal em todos os modelos. O Random Forest, por exemplo, previu uma probabilidade de insucesso de 65,73% para o perfil feminino e 61,11% para o masculino, quando a renda familiar era mantida constante em 1 < RFP ≤1,5. A Logistic Regression também mostrou pouca variação, com 65,25% de risco para mulheres e 65,31% para homens. Embora a análise descritiva anterior pudesse sugerir uma maior resiliência feminina, a predição da modelagem indica que as características de vulnerabilidade socioeconômica e curricular (Cor/Raça, Renda Familiar e Tipo de Curso) são os preditores dominantes de evasão neste perfil, com o Sexo tendo um impacto menos diferenciado.

5.2.1 Matriz de Confusão

O modelo Random Forest demonstrou a melhor performance na classificação da evasão. Na classe de interesse ('Não Concluiu'), o modelo identificou corretamente 1.298 alunos (Verdadeiros Positivos - TP), enquanto errou ao prever apenas 235 alunos que não concluíram como se tivessem concluído (Falsos Negativos - FN). Esta baixa taxa de Falsos Negativos é o motivo pelo qual o Random Forest alcançou o maior Recall (84,67%). No entanto, o modelo apresentou 499 Falsos Positivos (FP), ou seja, alunos que de fato concluíram, mas foram erroneamente sinalizados como em risco de evasão. Ilustrado na Figura 11.

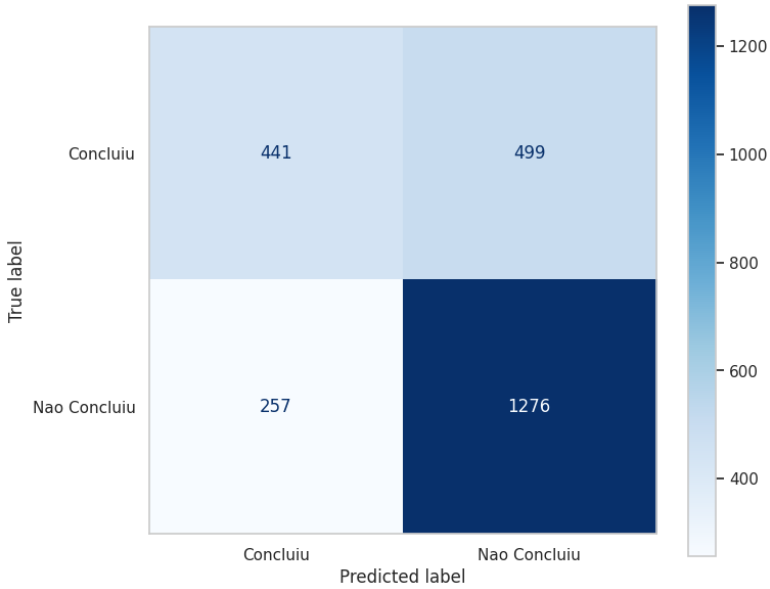
Figura 11. Matriz de confusão modelo Random Forest.



Fonte: Elaborado pela autora.

O Decision Tree apresentou o desempenho mais fraco na identificação de alunos que não concluíram. Embora tenha acertado a previsão de 1.276 alunos que não concluíram (TP), ele registrou 257 Falsos Negativos (FN), ficando atrás do Random Forest e com um Recall de 83,24%. Seu número de Falsos Positivos (FP) com 499 casos, o que contribuiu para sua menor Precisão (71,89%) e F1-score (77,15%). Em um contexto de intervenção, uma alta taxa de Falsos Positivos pode levar a recursos desnecessários investidos em alunos que não precisariam de suporte. Apresentado na Figura 12 abaixo.

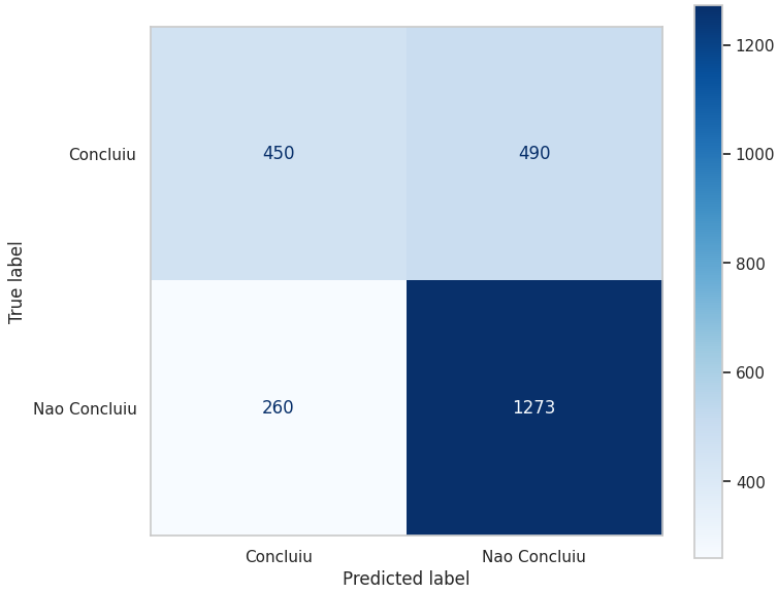
Figura 12. Matriz de confusão modelo Decision Tree.



Fonte: Elaborado pela autora.

O modelo de Logistic Regression teve um desempenho ligeiramente inferior ao Random Forest na identificação da evasão. O modelo acertou a previsão de 1.273 alunos que não concluíram (TP), mas cometeu 260 Falsos Negativos (FN), um número um pouco maior do que o Random Forest, resultando em um Recall de 83,04%. O Logistic Regression foi o modelo com a menor quantidade de erros ao classificar um aluno como "em risco" (menor FP), registrando 490 Falsos Positivos (FP). Isso resultou na maior Precisão para a classe 'Não Concluiu' (72,21%), indicando que, quando o modelo prevê "Não Concluiu", ele está ligeiramente mais propenso a estar certo. Ilustrado a seguir na Figura 13.

Figura 13. Matriz de confusão modelo Logistic Regression.



Fonte: Elaborado pela autora.

O modelo Random Forest é o mais adequado para o contexto de intervenção acadêmica, pois minimiza a taxa de Falsos Negativos. Na prática, o FN representa um aluno em risco que o modelo não conseguiu sinalizar, e que, consequentemente, não receberia suporte, resultando na evasão. A capacidade do Random Forest de identificar o maior número de casos reais de evasão (TP) com o menor número de FNs o torna a escolha mais segura para mitigar o problema.

6 CONSIDERAÇÕES FINAIS

O presente estudo buscou analisar a participação feminina em cursos de computação da rede federal de ensino no Brasil, e identificar os fatores preditivos para o sucesso/evasão nos cursos do IFPE, conforme estabelecido nos objetivos. Os resultados obtidos por meio da análise descritiva (Power BI) e da modelagem preditiva (Machine Learning) oferecem uma caracterização da trajetória da mulher na área de Computação.

A análise descritiva demonstrou que a presença feminina é marcada por um crescimento percentual constante, embora ainda minoritário, atingindo o pico de 36,24% em 2021 no cenário nacional e superando essa marca no IFPE (37,64% em 2023). Este avanço está fortemente concentrado em alunas jovens (15 a 24 anos) e,

crucialmente, em grupos de baixa renda e mulheres Pardas e Não Declaradas, sugerindo que os cursos de tecnologia, notadamente nas Universidades e Institutos Federais, atuam como um importante vetor de inclusão social. Além disso, as mulheres demonstram maior resiliência na conclusão, com a taxa percentual de formandas (pico de 43,12% em 2023 no cenário nacional) sendo consistentemente superior à sua proporção de matrícula, indicando maior foco no desfecho acadêmico. Os dados encontrados para os cursos de diferentes níveis de Computação do IFPE para os ingressantes estão alinhados com os trabalhos de Marinho e co-autores (2019) e Pereira e co-autores (2021).

A modelagem preditiva estabeleceu o algoritmo Random Forest (RF) como o modelo mais eficaz, obtendo a maior Acurácia (70,32%) e, crucialmente, o maior Recall (84,67%) para a classe 'Não Concluiu'. A principal vantagem do RF reside na minimização dos Falsos Negativos (FN=235), o que o torna a escolha ideal para intervenção, pois mitiga o risco de omissão e garante a máxima cobertura na identificação de alunos em risco de evasão. O risco de insucesso para mulheres (65,73%) é próximo ao dos homens (61,11%). Isso indica que, no contexto do IFPE, as barreiras socioeconômicas e o tipo de formação exercem uma influência preditiva mais forte do que a variável Sexo isoladamente. Portanto, para aumentar a conclusão do curso por mulheres, as intervenções devem focar principalmente em mitigar os riscos associados ao Tipo de Curso e à situação socioeconômica (Renda e Raça/Cor), e não em programas baseados unicamente no gênero.

Apesar dos avanços alcançados neste trabalho, existem algumas limitações a serem consideradas. Como limitação tem-se a utilização de base de dados públicas, fazendo com que a mesma viesse com algumas informações inconsistentes, dificultando o entendimento. Outra limitação é a falta de um dicionário indicando o que cada coluna significa, para facilitar no momento de filtragem das colunas necessárias.

Com base nos achados deste estudo, sugere-se que trabalhos futuros aprofundem a investigação sobre o ingresso e permanência das alunas trazendo os números detalhados por nível de curso, especialmente nos grupos de alta vulnerabilidade identificados. Recomenda-se a realização de uma análise qualitativa aprofundada sobre as causas de evasão no perfil pardo de baixa renda, complementando a estatística com a experiência vivida pelas estudantes. Alguns resultados desse público foram obtidos através do estudo de Santos (2025), mas é preciso um trabalho que aborda detalhadamente o perfil das egressas nos cursos. Além disso, para refinar a precisão e o Recall dos modelos preditivos, é essencial explorar a incorporação de variáveis comportamentais e acadêmicas nos modelos, como o desempenho em disciplinas de cálculo e programação, fornecendo uma visão mais completa dos fatores que determinam o sucesso ou o insucesso do aluno no contexto da computação.

REFERÊNCIAS

Alpaydin, E. **Introduction to Machine Learning Cambridge.** [S.l.]: MIT Press, 2004.

Alves, H. V. S. (2016). **Educação profissional e percepção de gênero:** uma investigação entre alunas e alunos do Serviço Nacional de Aprendizagem Comercial SENAC de Porto Velho - RO. Revista Formação (Online), 4(23):31–56.

Batista, A.S. (2015). **Regressão Logística:** Uma introdução ao modelo estatístico -Exemplo de aplicação ao Revolving Credit. Vida Economica Editorial. Disponível em: https://books.google.com.br/books?id=EtAsCgAAQBAJ. Acesso em: 03 dez. 2025.

Castro, B. (2013). **Os gargalos para o ingresso e a permanência das mulheres no mercado de TI, no Brasil.** In Conferencia Regional sobre la Mujer de América Latina y Caribe. CEPAL, Santo Domingo (Vol. 15, pp. 2018-2019).

Cursino, A. R.; Martinez, J. F. P. **Análise Estatística Descritiva e Regressão da Inserção das Mulheres nos Cursos de TI nos Anos de 2009 a 2018.** In: Anais XV Women in Information Technology (WIT 2021), XLI Congresso da Sociedade Brasileira de Computação (CSBC 2021), 2021.

Fapesp (2023). **Ingressos em programas de engenharia e de computação.** https://revistapesquisa.fapesp.br/ingressos-em-programas-de-engenharia-e-de-computacao-2/. Acesso em: 02 out. 2025.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. Elsevier, 2011. ISBN 0123814804.

HILBE, J. M. **Logistic regression.** International encyclopedia of statistical science, v. 1, p. 15–32, 2011.

HOMEM, W.L. **Apostila de Machine Learning**, 2020. Apostila do Minicurso de Machine Learning. Disponível em: https://petmecanica.ufes.br/sites/petengenhariamecanica.ufes.br/files/field/anexo/apostila_do_minicurso_de_machine_learning.pdf. Acesso em: 03 dez. 2025.

HONDA, F.P. **Estudo dos condicionantes espaciais para avaliação imobiliária utilizando técnicas de inteligência artificial** – São Paulo/SP. Dissertação (Mestrado). Universidade Federal de São Carlos, 2021. Disponível em https://repositorio.ufscar.br/handle/ufscar/14587 Acesso em: 02 dez. 2025.

INEP (2022). **Resumo técnico:** Censo da educação superior 2022. https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_educacao_superior_2022.pdf. Acesso em: 02 out. 2025.

JAMES, G. et al. **An introduction to statistical learning.** [S.l.]: Springer, 2013. v. 112.

MAIA, M. M. (2016). **Limites de gênero e presença feminina nos cursos superiores brasileiros do campo da computação.** Cadernos Pagu, n. 46, p. 223–244.

MARINHO, Gisele; FAGUNDES, Simone; AGUILAR, Carolina. **Análise da participação feminina nos cursos técnicos e de graduação da área de Informática da Rede Federal de Educação Tecnológica e do Cefet/RJ campus Nova Friburgo.** In: WOMEN IN INFORMATION TECHNOLOGY (WIT), 13., 2019, Belém. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 21-30. ISSN 2763-8626. DOI: https://doi.org/10.5753/wit.2019.6709.

MEDEIROS, A.; FERREIRA, B. M. C. I.; FONSECA, L.; ROLIM, C. (2022) **Percepções sobre a tecnologia da informação por alunas de ensino médio:** um estudo sobre gênero e escolhas profissionais. In: WIT, p. 122-132.

MORAES, G. H.; JÚNIOR, W. Tavares da S.; KENCHIAN, G. **Guia de referência metodológica PNP**, 2020. Brasília/DF, Editora Evobiz, p. 1–181, 2020.

Motogna, S., Alboaie, L., Todericiu, I. A., and Zaharia, C. (2022). **Retaining women in computer science:** The good, the bad and the ugly sides. In Proceedings of the Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering, GE@ICSE '22, page 35–42, New York, NY, USA. Association for Computing Machinery.

NASCIMENTO, Francisco Paulo Do. **Metodologia da pesquisa científica:** teoria e prática - como elaborar TCC. Brasília: Thesauros, 2016.

Oliveira, A., Moro, M., & Prates, R. (2014). **Perfil feminino em computação:** Análise inicial. In Anais do XXII Workshop sobre Educação em Computação (pp. 179-188). SBC.

PAIVA, T. S. Z. N.; SILVA, J. S. **A Participação Feminina nos Cursos Técnicos Integrados ao Ensino Médio da Educação Profissional e Tecnológica.** Revista Brasileira de Informática na Educação, [S. l.], v. 29, p. 993–1006, 2021. DOI: 10.5753/rbie.2021.29.0.993. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/3509. Acesso em: 22 set. 2025.

PEREIRA, J. S. et al. **Uma análise da participação das mulheres nos cursos técnico em informática e ciência da computação do instituto federal do sudeste de minas gerais.** In: Anais do XIV Women in Information Technology (WIT 2020). XL Congresso da Sociedade Brasileira de Computação (CSBC 2021), 2021.

PETERSSON, D. **What is Supervised Learning?** — techtarget.com. 2021. <https://www.techtarget.com/searchenterpriseai/definition/supervised-learning>. Acesso em: 01 dez. 2025.

QUINLAN, J.R. **Induction of decision trees.** Machine learning, 1(1), p.81-106,1986.

RAMEZANKHANI, A. et al. **Applying decision tree for identification of a low risk population for type 2 diabetes.** tehran lipid and glucose study. Diabetes research and clinical practice, Elsevier, v. 105, n. 3, p. 391–398, 2014.

RIBEIRO, K. S. F. M.; MACIEL, C. (2020) **Fatores de Influência na Escolha pela Continuidade da Carreira em Computação pelas Estudantes de Ensino Médio Técnico em Informática.** In: WIT, p. 40-49.

SANTOS, Ana Carolina Barbosa dos. **Mulheres na computação:** uma análise dos fatores de ingresso, permanência e sucesso profissional. Jaboatão dos Guararapes: IFPE, 2025. Trabalho de conclusão de curso (Análise e Desenvolvimento de Sistemas) - campus Jaboatão dos Guararapes, IFPE, 2025.

SANTOS, Ana Carolina Barbosa dos; SANTANA, Ellen Patrícia Lopes de; AURELIANO, Viviane Cristina Oliveira. **Análise da Participação Feminina nos Cursos de Nível Superior da área de Computação do IFPE.** In: WOMEN IN INFORMATION TECHNOLOGY (WIT), 19. , 2025, Maceió/AL. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 287-297. ISSN 2763-8626. DOI: https://doi.org/10.5753/wit.2025.9315.

SANTOS, Vívian Ludimila Aguiar; CARVALHO, Thales Francisco Mota; BARRETO, Maria do Socorro Vieira. **Mulheres na Tecnologia da Informação:** Histórico e Cenário Atual nos Cursos Superiores. In: WOMEN IN INFORMATION

TECHNOLOGY (WIT), 15. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 111-120. ISSN 2763-8626. DOI: https://doi.org/10.5753/wit.2021.15847.

SOUZA, T. P. (2017). **A desigualdade de gênero no campo da tecnologia da informação**. In: SEMINÁRIO INTERNACIONAL FAZENDO GÊNERO, 11., 2017. Anais [...]. Florianópolis: UFSC.

SWAMINATHAN, S. **Logistic Regression** - Detailed Overview. 2018. Disponível em: https://medium.com/data-science/logistic-regression-detailed-overview-46c4da4303bc. Acesso em: 03 dez. 2025.

VIEIRA, V. et al. **Análise de algoritmos de árvores de decisão e floresta randômica.** 2020.