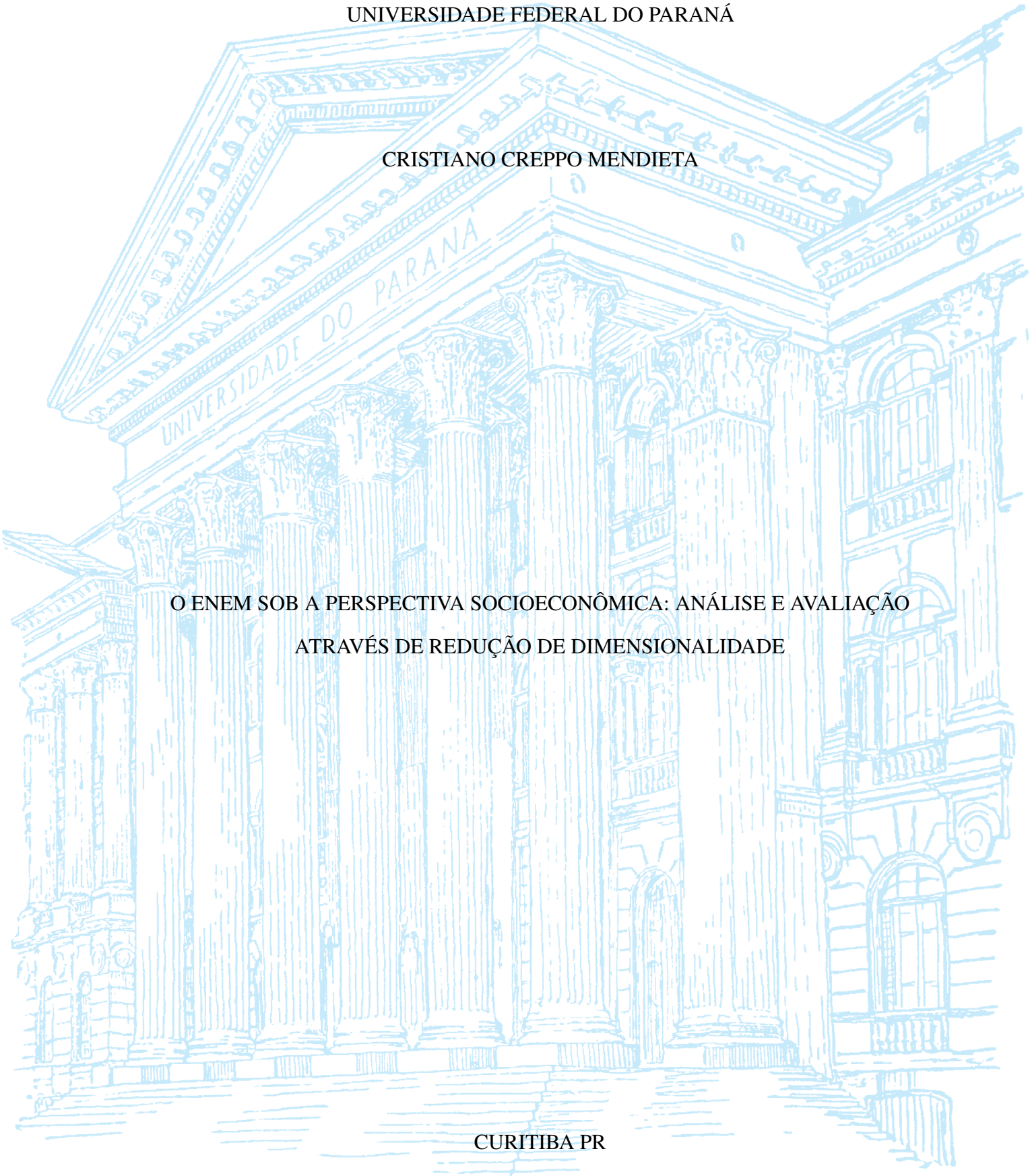
UNIVERSIDADE FEDERAL DO PARANÁ

CRISTIANO CREPPO MENDIETA



O ENEM SOB A PERSPECTIVA SOCIOECONÔMICA: ANÁLISE E AVALIAÇÃO ATRAVÉS DE REDUÇÃO DE DIMENSIONALIDADE

CURITIBA PR

2024

CRISTIANO CREPPO MENDIETA

O ENEM SOB A PERSPECTIVA SOCIOECONÔMICA: ANÁLISE E AVALIAÇÃO ATRAVÉS DE REDUÇÃO DE DIMENSIONALIDADE

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: André Vignatti.

CURITIBA PR

2024

Ficha catalográfica

Substituir o arquivo `0-iniciais/catalografica.pdf` pela ficha catalográfica fornecida pela Biblioteca da UFPR (PDF em formato A4).

Instruções para obter a ficha catalográfica e fazer o depósito legal da tese/dissertação (contribuição de André Hochuli, abril 2019. Links atualizados Wellton Costa, Nov 2022):

1. Estas instruções se aplicam a dissertações de mestrado e teses de doutorado. Trabalhos de conclusão de curso de graduação e textos de qualificação não precisam segui-las.
2. Verificar se está usando a versão mais recente do modelo do PPGInf e atualizar, se for necessário (`https://gitlab.c3sl.ufpr.br/maziero/tese`).
3. conferir o *checklist* de formato do Sistema de Bibliotecas da UFPR, em `https://bibliotecas.ufpr.br/servicos/normalizacao/`
4. Enviar e-mail para "`referencia.bct@ufpr.br`" com o arquivo PDF da dissertação/tese, solicitando a respectiva ficha catalográfica.
5. Ao receber a ficha, inseri-la em seu documento (substituir o arquivo `0-iniciais/catalografica.pdf` do diretório do modelo).
6. Emitir a Certidão Negativa (CND) de débito junto a biblioteca, em `https://bibliotecas.ufpr.br/servicos/certidao-negativa/`
7. Avisar a secretaria do PPGInf que você está pronto para o depósito. Eles irão mudar sua titulação no SIGA, o que irá liberar uma opção no SIGA pra você fazer o depósito legal.
8. Acesse o SIGA (`http://www.prppg.ufpr.br/siga`) e preencha com cuidado os dados solicitados para o depósito da tese.
9. Aguarde a confirmação da Biblioteca.
10. Após a aprovação do pedido, informe a secretaria do PPGInf que a dissertação/tese foi depositada pela biblioteca. Será então liberado no SIGA um link para a confirmação dos dados para a emissão do diploma.

Ficha de aprovação

Substituir o arquivo 0-iniciais/aprovacao.pdf pela ficha de aprovação fornecida pela secretaria do programa, em formato PDF A4.

À minha família e aos amigos que deram todo o suporte necessário para que eu pudesse percorrer este caminho.

AGRADECIMENTOS

Agradeço à minha mãe por todo o suporte e apoio indispensáveis em todas as etapas da minha vida. Sou profundamente grato a toda a minha família, que sempre acreditou em mim e me incentivou a seguir em frente.

Aos amigos, pela companhia e pelo papel essencial que desempenharam ao longo dessa jornada, deixo meu sincero agradecimento.

Expresso minha gratidão ao meu orientador, André Vignatti, pelo apoio, pelas orientações valiosas e por guiar-me durante a realização deste trabalho.

À UFPR e ao DInf, pela experiência transformadora e enriquecedora que moldou a pessoa e o profissional que sou hoje, o meu mais profundo reconhecimento.

RESUMO

Este estudo investiga a relação entre fatores socioeconômicos e o desempenho acadêmico dos estudantes no ENEM 2022, aplicando técnicas de redução de dimensionalidade ao conjunto de microdados disponibilizado pelo INEP. Esses dados são compostos por informações coletadas pelo exame, incluindo provas, gabaritos, itens avaliados, notas dos participantes e respostas ao questionário socioeconômico. A pesquisa compara métodos lineares, como *Principal Component Analysis* (PCA), *Singular Value Decomposition* (SVD) e *Independent Component Analysis* (ICA), e não-lineares, como *Autoencoders* e *Pairwise Controlled Manifold Approximation Projection* (PaCMAP), em cenários de classificação binária e multiclasse. Os resultados indicam que métodos lineares apresentam um bom equilíbrio entre precisão e eficiência computacional, especialmente em cenários de classificação binária. No entanto, métodos não-lineares são mais adequados para capturar estruturas complexas em classificações multiclasse, embora apresentem maior custo computacional. A técnica de Seleção de Características, utilizando *XGBoost*, mostrou-se eficaz na identificação de variáveis-chave que diferenciam os estudantes com base em características socioeconômicas e desempenho escolar. Este estudo contribui para a análise de grandes volumes de dados educacionais, proporcionando resultados que podem orientar a formulação de políticas públicas voltadas à promoção da equidade no sistema educacional brasileiro.

Palavras-chave: Redução de Dimensionalidade; Dados Educacionais; ENEM; Mineração de Dados.

ABSTRACT

This study investigates the relationship between socioeconomic factors and student academic performance in the 2022 ENEM, applying dimensionality reduction techniques to the microdata set provided by INEP. This dataset includes information collected from the exam, such as test scores, answer keys, evaluated items, participant scores, and responses to the socioeconomic questionnaire. The research compares linear methods, such as *Principal Component Analysis* (PCA), *Singular Value Decomposition* (SVD), and *Independent Component Analysis* (ICA), with non-linear methods, such as *Autoencoders* and *Pairwise Controlled Manifold Approximation Projection* (PaCMAP), in binary and multiclass classification scenarios. The results indicate that linear methods provide a good balance between accuracy and computational efficiency, especially in binary classification scenarios. However, non-linear methods are more suitable for capturing complex structures in multiclass classifications, despite their higher computational cost. The Feature Selection technique using *XGBoost* proved effective in identifying key variables that differentiate students based on socioeconomic characteristics and academic performance. This study provides a comprehensive analysis of large educational datasets, generating results that can guide the formulation of public policies aimed at promoting equity within the Brazilian educational system.

Keywords: Dimensionality Reduction, Educational Data, ENEM, Data Mining

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

ENEM	Exame Nacional do Ensino Médio
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
PCA	Principal Component Analysis (Análise de Componentes Principais)
SVD	Singular Value Decomposition (Decomposição em Valores Singulares)
ICA	Independent Component Analysis (Análise de Componentes Independentes)
PaCMAP	Pairwise Controlled Manifold Approximation Projection
PLN	Processamento de Linguagem Natural
UMAP	Uniform Manifold Approximation and Projection
XGBoost	Extreme Gradient Boosting
IBGE	Instituto Brasileiro de Geografia e Estatística
POF	Pesquisa de Orçamentos Familiares

SUMÁRIO

1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) é uma das principais avaliações de larga escala no Brasil, destinada a medir o desempenho acadêmico dos estudantes do ensino médio. Através da aplicação do exame, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) coleta não apenas informações de desempenho dos alunos, mas também dados socioeconômicos e demográficos detalhados obtidos por meio de questionários respondidos pelos participantes. Esses microdados, como são denominados pelo INEP, permitem uma análise aprofundada dos fatores que influenciam o desempenho dos estudantes. No entanto, o grande volume e a complexidade desses dados representam desafios significativos, como a necessidade de identificar características relevantes e minimizar redundâncias, especialmente em análises voltadas para a compreensão da relação entre fatores socioeconômicos e desempenho escolar.

Esse conjunto de dados destaca-se pela riqueza e complexidade das informações, o que o torna ideal para a aplicação de técnicas de redução de dimensionalidade. O elevado número de variáveis aumenta a complexidade das análises, tornando os modelos de aprendizado de máquina mais suscetíveis a enviesamentos e ineficiências computacionais. Para lidar com esses desafios, é essencial aplicar métodos que simplifiquem a representação dos dados, reduzindo o número de variáveis sem comprometer a qualidade das informações.

Entre as abordagens mais eficazes para mitigar a complexidade dos dados, destacam-se os métodos de redução de dimensionalidade, amplamente aplicados em diversas áreas. No processamento de linguagem natural (PLN), por exemplo, técnicas de *embeddings* como *Word2Vec* e BERT facilitam a análise textual ao gerar representações mais compactas e eficientes (Mikolov et al., 2013; Devlin et al., 2019). Na biologia, técnicas como t-SNE e UMAP são usadas para visualizar dados de células individuais, revelando padrões genéticos e agrupamentos celulares (Becht et al., 2019). Em processamento de imagens, métodos como PCA e *Autoencoders* auxiliam na compressão e segmentação de imagens, otimizando tanto a análise quanto o armazenamento (Hinton e Salakhutdinov, 2006). Esses exemplos demonstram a versatilidade e eficácia das técnicas de redução de dimensionalidade em uma ampla gama de domínios.

No contexto educacional, a redução de dimensionalidade não apenas otimiza os recursos computacionais, mas também aprimora a compreensão da relação entre fatores socioeconômicos e desempenho acadêmico, possibilitando uma análise mais eficiente e precisa. Esse conhecimento é fundamental para avaliar o impacto do contexto social na educação e apoiar a formulação de políticas públicas que promovam a equidade e a qualidade no ensino.

Diante desse cenário, este trabalho busca investigar a relação entre fatores socioeconômicos e o desempenho acadêmico dos estudantes no ENEM 2022. Para isso, são aplicadas técnicas de redução de dimensionalidade aos microdados disponibilizados pelo INEP, com o objetivo de simplificar a análise dos dados educacionais sem comprometer a relevância das informações essenciais para a compreensão dessa relação.

1.1 TRABALHOS RELACIONADOS

No campo educacional, o uso de grandes volumes de dados para análise e modelagem vem se expandindo significativamente, especialmente em avaliações em larga escala como o ENEM. Os microdados do ENEM, fornecidos pelo INEP, abrangem informações detalhadas sobre o desempenho acadêmico dos alunos, bem como dados socioeconômicos e demográficos, permitindo estudos aprofundados dos fatores associados ao desempenho escolar. No entanto, o

grande volume e a alta complexidade desses dados impõem desafios, como o aumento dos custos computacionais e a necessidade de selecionar variáveis de maior relevância, especialmente em análises voltadas para compreender a influência de fatores externos, como aspectos demográficos e socioeconômicos, no desempenho dos estudantes (Oliveira et al., 2024; Santos et al., 2023).

Esse crescimento exponencial no volume e na complexidade dos dados reflete uma tendência global da geração acelerada de dados. A partir disso, a aplicação de técnicas de aprendizado de máquina tem avançado em diversos campos de pesquisa e setores da indústria, promovendo análises mais profundas e informativas (Jia et al., 2022; Binois e Wycoff, 2022). Entretanto, o uso de grandes quantidades de dados apresenta desafios, como a necessidade de mais recursos computacionais e o fenômeno conhecido como "maldição da dimensionalidade", que afeta negativamente a eficiência dos modelos (Binois e Wycoff, 2022; Köppen, 2000).

Para enfrentar esses problemas, técnicas de redução de dimensionalidade tornaram-se essenciais na análise de dados. Métodos como *Principal Component Analysis* (PCA), *t-distributed Stochastic Neighbor Embedding* (t-SNE), *Uniform Manifold Approximation and Projection* (UMAP) e *Pairwise Controlled Manifold Approximation Projection* (PaCMAP) destacam-se por sua capacidade de eliminar atributos redundantes e irrelevantes, melhorando o desempenho computacional e otimizando o uso dos dados (Wang et al., 2021; McInnes et al., 2018).

Estudos recentes têm explorado o contexto educacional para investigar questões socioeconômicas em dados complexos, como os microdados do ENEM. Oliveira et al. (2024), por exemplo, analisaram esses dados para explorar desigualdades educacionais em escolas públicas do Ceará, aplicando mineração de dados para identificar o impacto de fatores socioeconômicos sobre o desempenho estudantil (Oliveira et al., 2024). De forma semelhante, Santos et al. (2023) examinaram o efeito da pandemia nas desigualdades sociais nos resultados do ENEM de 2019 e 2020, utilizando métodos de redução de dimensionalidade para interpretar as variações observadas (Santos et al., 2023). Esses trabalhos destacam o valor das técnicas de redução de dimensionalidade na otimização da análise de grandes volumes de dados educacionais.

Além disso, estudos como o de Queiroga et al. (2024) destacam a importância de abordagens orientadas por dados para promover a equidade escolar no Brasil. Nesse cenário, a combinação entre mineração de dados e redução de dimensionalidade tem se mostrado eficaz na identificação de fatores que contribuem para as desigualdades educacionais, especialmente no contexto do ENEM. Essa análise contribui diretamente para o desenvolvimento de políticas públicas mais inclusivas e eficazes, visando promover a qualidade do sistema educacional. Este trabalho, portanto, busca avançar a discussão ao avaliar comparativamente diferentes métodos de redução de dimensionalidade aplicados aos microdados do ENEM, analisando como esses métodos podem melhorar a eficiência e a precisão na análise dos fatores socioeconômicos e seu impacto no desempenho acadêmico.

1.2 NOSSOS RESULTADOS

Neste trabalho, analisamos a aplicação de diferentes técnicas de redução de dimensionalidade aos microdados do ENEM 2022, com o objetivo de avaliar a influência dos fatores socioeconômicos no desempenho escolar dos estudantes. A investigação concentrou-se em métodos de redução de dimensionalidade lineares e não-lineares, comparando sua eficácia e eficiência em cenários de classificação binária e multiclasse. Em nosso estudo, os métodos lineares, como o PCA, SVD e ICA, mostraram-se eficazes ao preservar a acurácia e reduzir significativamente o tempo de treinamento, destacando-se como alternativas adequadas para tarefas de classificação com grandes volumes de dados.

Em contrapartida, os métodos não-lineares, como *Autoencoders* e PaCMAP, revelaram-se mais robustos na captura de estruturas complexas, embora apresentassem maior custo computacional. A técnica de Seleção de Características com *XGBoost* mostrou-se particularmente eficaz na identificação de variáveis relevantes para diferenciar estudantes com base em características socioeconômicas. Esses resultados sugerem que a escolha da técnica de redução de dimensionalidade deve considerar tanto a complexidade dos dados quanto os recursos computacionais disponíveis, sendo essencial equilibrar a precisão dos modelos com a eficiência de processamento. Nossa contribuição consiste em uma análise comparativa detalhada dessas técnicas no contexto educacional, com o objetivo de apoiar a compreensão dos diversos fatores que influenciam, direta ou indiretamente, o desempenho e o desenvolvimento educacional no Brasil.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Este trabalho está estruturado da seguinte forma. No Capítulo 2, apresentamos os fundamentos teóricos das técnicas de redução de dimensionalidade exploradas, incluindo métodos lineares e não-lineares, e a relevância da Seleção de Características para a simplificação dos dados. O Capítulo 3 descreve em detalhes a metodologia empregada, abordando desde a preparação dos dados até os experimentos realizados. O Capítulo 4, discutimos os resultados dos experimentos, comparando o desempenho dos diferentes métodos em termos de acurácia, *F1-score* e tempo de processamento. Concluímos no Capítulo 5, onde discutimos os resultados obtidos e o impacto das técnicas de redução de dimensionalidade no contexto educacional.

2 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo discute os fundamentos teóricos sobre a redução de dimensionalidade. A redução de dimensionalidade é essencial para mitigar os desafios impostos pela alta quantidade de variáveis e pela redundância informacional, fatores que podem comprometer tanto a eficiência computacional quanto a interpretabilidade dos modelos. Neste contexto, abordaremos as técnicas redução de dimensionalidade utilizadas neste trabalho.

2.1 REDUÇÃO DE DIMENSIONALIDADE DE DADOS

Com o aumento exponencial dos conjuntos de dados em diversas áreas, a redução de dimensionalidade tornou-se uma técnica essencial. Esse processo visa reduzir o número de dimensões de um conjunto de dados, mapeando o espaço original para um espaço de menor dimensão, preservando ao máximo a estrutura dos dados originais. Esse tipo de redução é particularmente útil em áreas como visão computacional, bioinformática e visualização de dados, onde o alto número de características e dependências complexas pode tornar a análise e modelagem computacionalmente desafiadoras.

Dados de alta dimensionalidade apresentam desafios como redundância, presença de variáveis irrelevantes e aumento da complexidade, dificultando a eficiência dos algoritmos de aprendizado de máquina. Além disso, o fenômeno da "maldição da dimensionalidade" ocorre quando, com o aumento das dimensões, as distâncias entre os pontos tornam-se mais uniformes, dificultando a identificação de padrões relevantes nos dados (Bellman, 1958). Para mitigar esses problemas, técnicas de redução de dimensionalidade, como *Uniform Manifold Approximation and Projection* (UMAP) e *Pairwise Controlled Manifold Approximation Projection* (PaCMAP), vêm sendo amplamente aplicadas, pois otimizam o uso de recursos computacionais e aumentam a precisão em tarefas de classificação em grandes conjuntos de dados complexos (Hirasawa, 2023).

Existem duas abordagens principais na redução de dimensionalidade: a Extração de Características, que transforma o conjunto original em novas variáveis representativas, e a Seleção de Características, que identifica e mantém as variáveis mais relevantes do conjunto de dados existente. Ambas desempenham um papel crucial na construção de modelos mais eficientes e na interpretação dos dados, sendo detalhadas nas próximas seções.

2.1.1 Extração de Características

A Extração de Características transforma os dados originais em um novo conjunto de variáveis, que buscam capturar a variabilidade ou padrões relevantes dos dados em um espaço reduzido. Técnicas populares incluem o *Principal Component Analysis* (PCA), que gera componentes principais ortogonais que maximizam a variância dos dados (Jolliffe, 2002), e *Autoencoders*, redes neurais que aprendem representações compactas em um espaço latente (Wang et al., 2012). No contexto deste estudo, a técnica de Extração de Características foi aplicada usando PCA, SVD, ICA, PaCMAP e *Autoencoders*, visando preservar as principais características dos microdados do ENEM.

2.1.2 Seleção de Características

Diferente da extração, a Seleção de Características não cria novas variáveis, mas seleciona um subconjunto das variáveis originais mais relevantes para a tarefa de modelagem.

Isso ajuda a reduzir a complexidade e aumentar a interpretabilidade dos modelos. A Seleção de Características é especialmente útil quando há variáveis irrelevantes ou redundantes que podem ser descartadas sem afetar o desempenho do modelo (Weikuan et al., 2022).

Neste estudo, a Seleção de Características foi realizada utilizando o algoritmo *XGBoost*, que calcula a importância das variáveis durante o processo de treinamento. Esse cálculo é realizado através de três métricas principais: Ganho, Cobertura e Frequência, permitindo identificar variáveis críticas que influenciam diretamente o desempenho do modelo (Chen et al., 2018).

- **Ganho**: Refere-se ao ganho médio na precisão do modelo ao dividir os dados com uma determinada variável. Quanto maior o ganho associado a uma variável, mais relevante ela é considerada para o desempenho do modelo.
- **Cobertura**: Mede a cobertura média dos dados que uma variável impacta, ou seja, quantos dados passam pelo nó onde a variável foi usada para a divisão. Isso oferece uma medida de quão frequentemente a variável influencia as previsões.
- **Frequência**: Indica a frequência com que uma variável é usada para dividir os dados nas árvores do modelo, proporcionando uma visão adicional sobre a importância da variável com base em sua recorrência.

Essas métricas permitem ao *XGBoost* quantificar a importância das variáveis de forma precisa, auxiliando na construção de modelos mais interpretáveis e otimizados. No presente estudo, o modelo foi inicialmente treinado com todas as variáveis, e as variáveis foram classificadas com base na importância (Ganho). Posteriormente, foram realizados novos treinamentos considerando apenas os subconjuntos mais relevantes de variáveis.

2.1.3 Métodos de Redução Lineares

Os métodos de redução de dimensionalidade lineares assumem que os dados podem ser transformados em um subespaço de menor dimensionalidade através de combinações lineares das variáveis originais. Essa transformação é particularmente eficaz em cenários com variáveis altamente correlacionadas.

Principal Component Analysis (PCA): transforma o conjunto de dados original em um novo conjunto de variáveis, os componentes principais, que são combinações lineares das variáveis originais e são ordenados de forma que o primeiro componente retém a maior parte da variância dos dados, seguido pelos demais componentes (Jolliffe, 2002). Matematicamente, a redução é feita decompondo a matriz de covariância dos dados em autovalores e autovetores, mantendo os componentes com maior variância.

Singular Value Decomposition (SVD): é uma técnica que decompõe a matriz de dados em três componentes fundamentais: duas matrizes ortogonais e uma matriz diagonal composta por valores singulares, que representam a importância de cada dimensão nos dados originais. Essa decomposição permite identificar e descartar componentes de menor relevância, facilitando a redução dimensional sem perder informações essenciais. O SVD é amplamente aplicado em compressão de dados, já que os valores singulares menores (que carregam menos variância) podem ser truncados, reduzindo a dimensionalidade e simplificando o processamento, sendo também muito utilizado em análise de dados textuais e em sistemas de recomendação.(Klema e Laub, 1980)

Independent Component Analysis (ICA): é uma técnica linear que decompõe os dados em componentes estatisticamente independentes, ou seja, variáveis que não influenciam mutuamente seu comportamento. Em outras palavras, o valor de uma variável não fornece informações sobre o valor de outra. Essa característica é especialmente útil em aplicações como a separação de sinais, onde o ICA pode isolar diferentes fontes de áudio (como vozes) em um ambiente misto, assumindo que cada fonte é independente das demais. A técnica é, portanto, ideal para cenários que requerem independência entre os componentes extraídos (Hyvärinen e Oja, 2000).

2.1.4 Métodos de Redução Não-Lineares

Os métodos de redução não-lineares são desenvolvidos para dados cujas estruturas não podem ser adequadamente capturadas por transformações lineares, como relações complexas e não-lineares entre variáveis.

Pairwise Controlled Manifold Approximation Projection (PaCMAP): é uma técnica de redução de dimensionalidade projetada para preservar a estrutura dos dados tanto em nível local quanto global ao mapear para um espaço de menor dimensão. Essa característica torna o PaCMAP especialmente útil em conjuntos de dados com múltiplas estruturas, onde é importante manter relações próximas entre pontos similares sem perder a percepção das estruturas globais. PaCMAP realiza essa preservação ao ajustar pares de pontos para otimizar a proximidade relativa, equilibrando a estrutura dos dados em diferentes escalas. Estudos recentes indicam que essa técnica é altamente eficaz para tarefas de classificação e agrupamento, onde a consistência estrutural é essencial (Wang et al., 2021).

Autoencoders: são redes neurais projetadas para aprender representações compactas de dados de alta dimensionalidade, comprimindo-os em um espaço latente e reconstruindo-os com mínima perda de informação. Comparado a métodos lineares, os *Autoencoders* capturam relações não-lineares complexas, sendo úteis para dados com padrões complexos. Variantes, como *Autoencoders* variacionais, permitem modelagem probabilística para cenários com dados ruidosos (Keser e Töreyin, 2019).

3 METODOLOGIA

Este estudo visa compreender a relação entre fatores socioeconômicos e desempenho acadêmico dos estudantes do Exame Nacional do Ensino Médio (ENEM) de 2022. Para isso, foram aplicadas e comparadas diversas técnicas de redução de dimensionalidade, com o objetivo de extrair e simplificar os dados sem perder informações relevantes. A metodologia abrange desde o pré-processamento dos microdados até a avaliação do impacto dessas técnicas em tarefas de classificação binária e multiclasse, utilizando métricas como acurácia, *F1-Score* e tempo de treinamento para verificar a eficiência e a eficácia dos modelos gerados.

3.1 CONJUNTO DE DADOS E PRÉ-PROCESSAMENTO

Esta seção descreve a origem dos dados utilizados neste estudo, bem como as etapas de pré-processamento aplicadas para viabilizar a análise.

Fonte e Características dos Dados: O conjunto de dados empregado neste trabalho é composto pelos microdados do Exame Nacional do Ensino Médio (ENEM) de 2022, obtidos junto ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Esses dados incluem uma ampla gama de informações pessoais, sociais, econômicas e de desempenho dos participantes do exame, abrangendo variáveis como faixa etária, tipo de escola frequentada no ensino médio, escolaridade dos pais, presença de bens como carro e geladeira na residência, além da localização da escola concluída, entre outras características. Esse conjunto de dados destaca-se pela sua riqueza e complexidade, tornando-o ideal para a aplicação de técnicas de redução de dimensionalidade. O conjunto original contém 3.476.105 amostras e 76 atributos, representando uma base rica para análise e extração de padrões relevantes.

Seleção e Preparação das Variáveis: O processo de seleção e preparação das variáveis foi realizado com o objetivo de criar um conjunto de dados robusto e representativo. As variáveis foram categorizadas em dois grupos principais: Numéricas e Categóricas. Variáveis numéricas são aquelas que podem ser medidas numericamente e expressam uma quantidade, nesse caso, variáveis como idade, notas das provas, ano de conclusão do ensino médio, entre outras. Já as variáveis categóricas são aquelas que não possuem valores quantitativos, são categorias que podem ser atribuídas ao objeto de estudo, por exemplo, sexo, cor/raça, estado civil e as diversas perguntas do questionário socioeconômico.

Durante o pré-processamento, as variáveis foram avaliadas manualmente, e algumas colunas foram descartadas por não agregarem valor ao problema em estudo ou por apresentarem desafios significativos para um tratamento eficiente. Colunas que continham os vetores de respostas individuais de cada prova foram removidas devido à complexidade e ao grande número de combinações possíveis que uma codificação eficaz desses dados exigiria. Além disso, a coluna referente ao número de inscrição do aluno foi descartada por não fornecer informações relevantes para os objetivos do estudo, enquanto a coluna com o código do município da escola foi eliminada devido à alta quantidade de valores nulos. Após essa etapa, o conjunto de dados resultante ficou com 54 colunas.

Tratamento de Dados Faltantes: O tratamento de dados faltantes foi realizado através da remoção de registros incompletos, garantindo a integridade do conjunto de dados para análise.

Codificação e Normalização: As variáveis categóricas foram codificadas utilizando o `OrdinalEncoder` da biblioteca *Scikit-learn*, transformando-as em valores numéricos ordinais. Para as variáveis numéricas, aplicou-se a normalização através do `MinMaxScaler`, escalonando os valores para o intervalo [0, 1]. Este processo é crucial para garantir que todas as variáveis tenham pesos comparáveis nas análises subsequentes.

Definição das Variáveis Alvo: Foram estabelecidos dois cenários de classificação para avaliar a eficácia das técnicas de redução de dimensionalidade:

- **Classificação Binária:** Baseada no tipo de escola (pública ou privada), codificada como 0 para escola pública e 1 para escola privada.
- **Classificação Multiclasse:** Baseada na faixa de renda familiar, categorizada em cinco níveis (A, B, C, D, E), onde A representa a faixa de renda mais alta e E a mais baixa.

A classificação multiclasse foi estabelecida de acordo com o seguinte mapeamento de renda familiar mensal (em múltiplos de salários mínimos):

$$map\_renda = \begin{cases} A : [P, Q] & \text{(15 a 25 salários mínimos)} \\ B : [N, O] & \text{(10 a 15 salários mínimos)} \\ C : [J, K, L, M] & \text{(5 a 10 salários mínimos)} \\ D : [E, F, G, H, I] & \text{(2 a 5 salários mínimos)} \\ E : [A, B, C] & \text{(até 2 salários mínimos)} \end{cases}$$

Essa classificação foi baseada nas regras definidas para a divisão de renda familiar, onde **Classe A** abrange famílias com renda de 15 a 25 salários mínimos (R$18.180,01 a R$30.300,00), **Classe B** de 10 a 15 salários mínimos (R$12.120,01 a R$18.180,00), **Classe C** de 5 a 10 salários mínimos (R$7.272,01 a R$12.120,00), **Classe D** de 2 a 5 salários mínimos (R$2.424,01 a R$7.272,00), e **Classe E** até 2 salários mínimos (R$2.424,00 ou menos). Essas faixas de renda foram derivadas com base nos valores atualizados do salário mínimo em 2022.

Vale destacar que o Instituto Brasileiro de Geografia e Estatística (IBGE) adota outras divisões de renda, como na Pesquisa de Orçamentos Familiares (POF) de 2017-2018, que não segue essa classificação em classes A a E, mas sim em faixas de múltiplos do salário mínimo. No entanto, a nomenclatura utilizada neste estudo é amplamente aplicada em análises socioeconômicas, mesmo sem uma definição oficial pelo IBGE.

Tabela 3.1: Definição das Variáveis Alvo

Cenário	**Variável**	**Codificação**
Classificação Binária	Tipo de Escola	0: Pública, 1: Privada
Classificação Multiclasse	Faixa de Renda Familiar	A: Mais alta B: Alta C: Média D: Baixa E: Mais baixa

3.2 TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

Neste estudo, foram implementadas e comparadas seis técnicas de redução de dimensionalidade, abrangendo métodos lineares, não-lineares e o método de Seleção de Características. A Tabela 3.2 apresenta um resumo dessas técnicas e seus principais parâmetros.

Para cada técnica, testamos diferentes dimensões reduzidas: 2, 3, 5, 10 e 20 componentes, permitindo uma análise comparativa do desempenho em diferentes níveis de redução

Tabela 3.2: Técnicas de Redução de Dimensionalidade e Parâmetros Utilizados

Técnica	**Tipo**	**Parâmetros Principais**	**Valores Utilizados**
PCA	Linear	svd_solver	Valor padrão: "auto"
SVD	Linear	algorithm	Valor padrão: "randomized"
ICA	Linear	max_iter, tol	max_iter: 200 (padrão), tol: 0.0001 (padrão)
PaCMAP	Não-linear	n_neighbors, MN_ratio, FP_ratio	n_neighbors: 10, MN_ratio: 0.5, FP_ratio: 2.0
Autoencoder	Não-linear	learning_rate, epochs, batch_size	learning_rate: 0.001, epochs: 200, batch_size: 32

Já para a técnica de Seleção de Características, foi utilizado o algoritmo *XGBoost*, que avalia a importância das variáveis durante o treinamento do modelo. Esse cálculo de importância é baseado em três métricas principais: *Ganho*, *Cobertura* e *Frequência*. Neste trabalho, o modelo foi inicialmente treinado com todas as variáveis disponíveis no conjunto de dados, e as variáveis foram ranqueadas com base no *Ganho*. Em seguida, novos treinamentos foram realizados utilizando apenas os subconjuntos de variáveis mais relevantes, identificados por essa métrica. Essa abordagem permitiu uma redução significativa na dimensionalidade dos dados, sem comprometer o desempenho do modelo, além de oferecer uma análise interpretável sobre as variáveis mais críticas para a tarefa de classificação.

3.3 MODELAGEM E AVALIAÇÃO

Nesta seção, apresentamos os principais pontos referentes à modelagem do problema e de que forma as soluções estudadas foram avaliadas de forma a auxiliar no estudo comparativo dessas soluções. Nos experimentos realizados neste trabalho, todos os métodos de redução de dimensionalidade foram aplicados para reduzir os dados a 2, 3, 5, 10 e 20 componentes. Após a redução dimensional, todos os conjuntos de dados foram submetidos ao mesmo algoritmo de classificação, o *XGBoost*. Além disso, utilizou-se o *XGBoost* com todas as 53 características do conjunto de dados resultante após a etapa de seleção e preparação das variáveis como um modelo "baseline", permitindo a comparação dos resultados dos modelos reduzidos com a totalidade dos dados disponíveis.

Algoritmo de Classificação: O classificador XGBoost foi escolhido como algoritmo base para avaliar o desempenho das diferentes técnicas de redução de dimensionalidade. O XGBoost (*Extreme Gradient Boosting*) é um algoritmo de aprendizado de máquina baseado em árvores de decisão, amplamente reconhecido por sua eficiência e precisão em tarefas de classificação. Ele opera por meio de um conjunto de árvores, ajustando iterativamente novos modelos para corrigir os erros dos modelos anteriores, o que o torna robusto para conjuntos de dados complexos e de alta dimensionalidade. Sua capacidade de lidar com dados heterogêneos e a utilização

de técnicas avançadas de regularização contribuem para um desempenho otimizado e para a prevenção de sobreajuste, sendo uma escolha confiável em uma ampla gama de problemas de classificação(Chen e Guestrin, 2016).

Preparação dos Dados para Modelagem: O conjunto de dados foi dividido em conjuntos de treinamento (80%) e teste (20%) utilizando a função `train_test_split` da Scikit-learn, com uma semente aleatória fixa para garantir a reprodutibilidade.

Processo de Treinamento e Validação: O processo de treinamento e validação seguiu as seguintes etapas:

1. **Validação Cruzada:** Utilizou-se a validação cruzada com 5 *folds* (`KFold`) para avaliar a estabilidade do modelo e evitar *overfitting*.
2. **Treinamento do Modelo:** O *XGBoost* foi treinado com 400 estimadores, utilizando a função objetivo 'binary:logistic" para o cenário de classificação binária e 'multi:softmax" para o cenário multiclasse.
3. **Hiperparâmetros:** Os principais hiperparâmetros do *XGBoost* foram mantidos constantes em todos os experimentos para permitir uma comparação justa entre as técnicas de redução de dimensionalidade.

Métricas de Avaliação: Para avaliar o desempenho dos modelos, foram utilizadas as seguintes métricas:

- **Acurácia:** Proporção de previsões corretas sobre o total de previsões.
- **F1-score:** Média harmônica entre precisão e *recall*, fornecendo uma medida balanceada do desempenho do modelo.

Além das métricas de desempenho, também foi registrado o tempo de treinamento para cada combinação de técnica de redução e número de componentes, permitindo uma análise da eficiência computacional.

Implementação: A implementação foi realizada em Python 3.11, utilizando as seguintes bibliotecas principais:

- Pandas (1.2.4) e NumPy (1.20.2) para manipulação de dados
- Scikit-learn (0.24.2) para pré-processamento, implementação de PCA, SVD e ICA, e métricas de avaliação
- XGBoost (1.4.2) para o algoritmo de classificação
- PaCMAP (0.6.3) para a implementação do método PaCMAP
- TensorFlow (2.5.0) para a implementação do Autoencoder
- Matplotlib (3.4.2) para visualização dos resultados

3.3.1 Reprodutibilidade

Para garantir a reprodutibilidade dos experimentos, foram adotadas as seguintes medidas:

- Utilização de sementes aleatórias fixas em todas as etapas que envolvem aleatoriedade, como a divisão dos dados e a inicialização dos modelos.
- Disponibilização do código-fonte completo, incluindo *scripts* de pré-processamento, treinamento e avaliação, em um repositório público no GitHub[1].

Esta abordagem metodológica permite uma análise comparativa robusta das diferentes técnicas de redução de dimensionalidade aplicadas aos dados do ENEM 2022, fornecendo resultados que tornam-se valiosos sobre a eficácia e eficiência de cada método no contexto de classificação educacional.

[1] https://github.com/cristianomendieta/DimensionalityReductionENEM

4 RESULTADOS

Os resultados da aplicação das diferentes técnicas de redução de dimensionalidade foram avaliados com base em três métricas principais: acurácia, *F1-Score* e tempo de treinamento. Para cada técnica, a redução foi realizada para 2, 3, 5, 10 e 20 componentes, e os dados dimensionalmente reduzidos foram utilizados para treinar modelos de classificação utilizando o algoritmo XGBoost. A análise foi conduzida em cenários de classificação binária e multiclasse, permitindo uma comparação abrangente do desempenho das técnicas em diferentes configurações de redução de dimensionalidade e cenários de classificação.

4.1 CLASSIFICAÇÃO MULTICLASSE: PREVISÃO DA FAIXA DE RENDA FAMILIAR

Nesta seção, apresentamos uma análise detalhada do desempenho das técnicas de redução de dimensionalidade aplicadas ao cenário de classificação multiclasse. Serão discutidas as métricas de acurácia de teste, *F1-Score* e tempo de treinamento, permitindo uma compreensão acerca da eficácia e eficiência dos métodos estudados neste trabalho.

4.1.1 Acurácia de Teste

Na Figura 4.1, observamos a comparação da acurácia de teste para os diferentes métodos no cenário multiclasse. O *Principal Component Analysis* (PCA) demonstra ser uma escolha robusta, oferecendo uma acurácia competitiva em diferentes números de componentes. O *Autoencoder* aproxima-se do PCA, especialmente com um número maior de componentes, refletindo uma capacidade de retenção de informações mesmo após a redução dimensional. O *Singular Value Decomposition* (SVD) também apresenta um desempenho consistente, com resultados semelhantes ao PCA em todos os cenários, reforçando seu potencial como uma técnica linear eficaz. O *Independent Component Analysis* (ICA) mostrou um desempenho similar ao PCA e SVD, especialmente com um número maior de componentes, indicando sua eficácia na captura de variações relevantes para a tarefa de classificação. O PaCMAP, uma técnica não-linear, não apresenta um desempenho tão consistente quanto os métodos lineares, sugerindo uma limitação para cenários multiclasse, onde as relações entre as variáveis podem não ser tão complexas a ponto de exigir uma abordagem não-linear. A técnica de Seleção de Características apresentou resultados competitivos, alcançando acurácia próxima ao *baseline* e, em alguns casos, superando os outros métodos, demonstrando sua eficácia em selecionar as características mais relevantes e manter uma acurácia elevada mesmo com uma redução significativa no número de variáveis.

4.1.2 F1-Score

A Figura 4.2 mostra o desempenho em termos de *F1-Score* para o cenário multiclasse. Os métodos lineares, como PCA, SVD e ICA, mantêm robustez em diferentes níveis de redução dimensional, apresentando *F1-Scores* elevados e similares entre si. O *Autoencoder* oferece um *F1-Score* próximo aos métodos lineares, especialmente quando o número de componentes é maior, demonstrando sua eficiência em manter um desempenho elevado. O PaCMAP, enquanto útil para preservação de relações locais e globais em dados altamente complexos, mostrou-se menos eficiente no cenário multiclasse, sugerindo que sua estrutura não-linear pode não capturar adequadamente a complexidade desse tipo de classificação. A Seleção de Características também mostrou um desempenho competitivo em termos de *F1-Score*, especialmente com um número

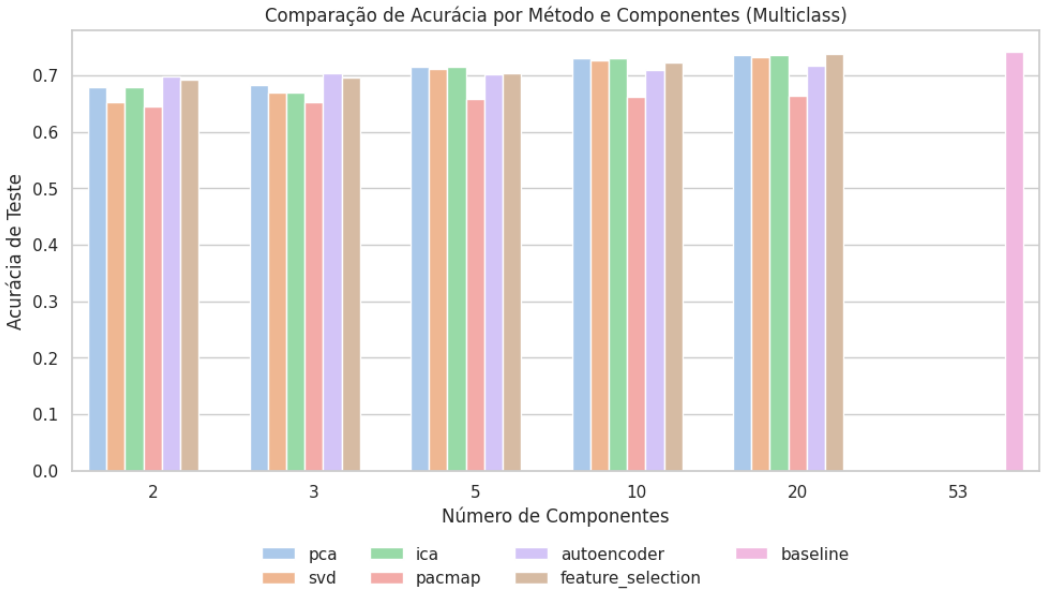


Figura 4.1: Comparação de Acurácia por Método e Componentes no Cenário Multiclasse.

maior de componentes, sendo comparável e até superior aos métodos lineares, demonstrando sua capacidade de priorizar características relevantes para a classificação multiclasse.

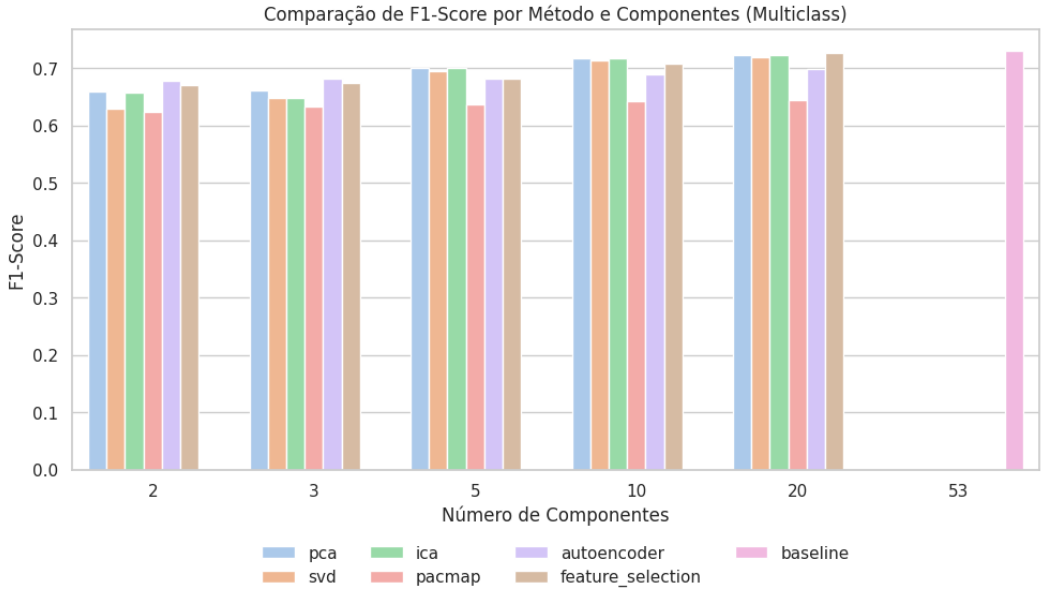


Figura 4.2: Comparação de F1-Score por Método e Componentes no Cenário Multiclasse.

4.1.3 Tempo de Treinamento

A Figura 4.3 exibe o tempo de treinamento para cada técnica no cenário multiclasse. Os métodos lineares, como PCA, SVD e ICA, apresentaram tempos de treinamento similares entre si. O *Autoencoder* possui tempos de treinamento comparáveis aos métodos lineares, indicando uma eficiência aceitável em termos computacionais. O PaCMAP, embora projetado para preservação de estruturas complexas, não apresentou vantagem significativa em termos de tempo de processamento. A Seleção de Características destacou-se pelo tempo de processamento

mais elevado, principalmente em cenários com poucos componentes, onde o processo de seleção de características demandou recursos computacionais significativos. Para aplicações onde o tempo de treinamento é crítico, métodos como PCA, SVD e ICA são mais recomendados.

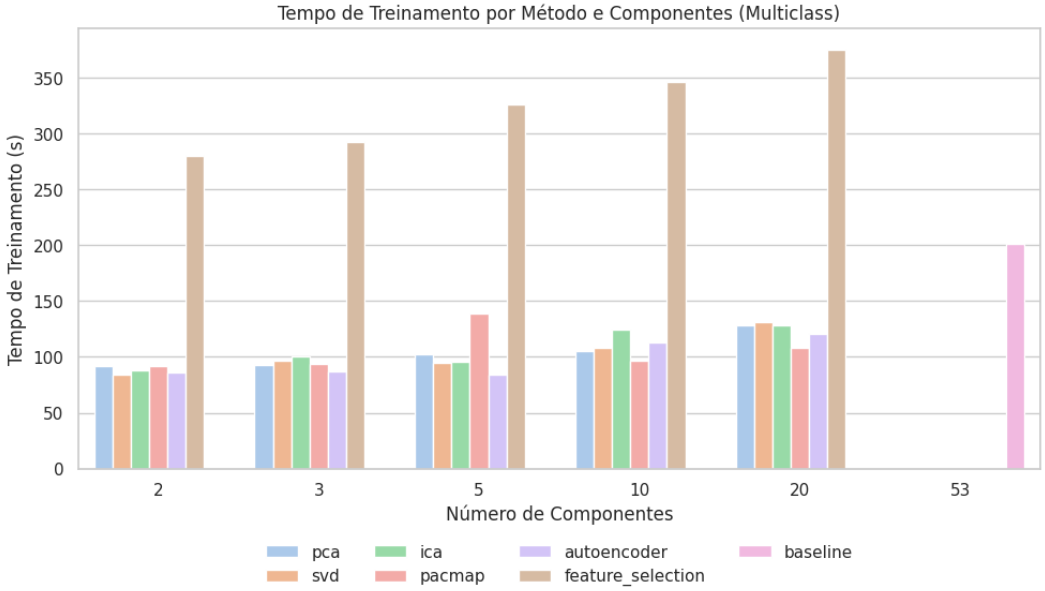


Figura 4.3: Comparação de Tempo de Treinamento por Método e Componentes no Cenário Multiclasse.

4.2 CLASSIFICAÇÃO BINÁRIA: PREVISÃO DO TIPO DE ESCOLA

Nesta seção, exploramos o desempenho das técnicas de redução de dimensionalidade aplicadas ao cenário de classificação binária. Serão discutidas as métricas de acurácia de teste, *F1-Score* e tempo de treinamento, visando a análise da performance e a viabilidade dos métodos analisados para lidar com problemas de classificação binária.

4.2.1 Acurácia de Teste

Na Figura 4.4, observamos os resultados de acurácia para o cenário binário. Tanto o PCA quanto o *Autoencoder* continuam a mostrar resultados sólidos em termos de acurácia com diferentes números de componentes. O PCA é consistentemente competitivo, mantendo uma alta acurácia, especialmente em níveis mais baixos de componentes. O SVD também demonstra resultados semelhantes ao PCA, reforçando sua eficácia em cenários com estrutura de dados mais simples. O *Autoencoder*, embora tenha um desempenho comparável ao PCA, mostra uma variação maior com números reduzidos de componentes, refletindo sua dependência de um maior número de dimensões para capturar padrões. O ICA apresenta desempenho similar aos métodos mencionados, indicando que a independência dos componentes consegue capturar efetivamente as relações no cenário binário. O PaCMAP, similarmente ao cenário multiclasse, não se destaca, o que sugere uma limitação na captura das relações simplificadas neste cenário binário. A Seleção de Características mostrou acurácia competitiva, alcançando resultados próximos ao *baseline* e superando os demais métodos em certos casos, indicando sua capacidade de selecionar as características mais relevantes para a classificação binária.

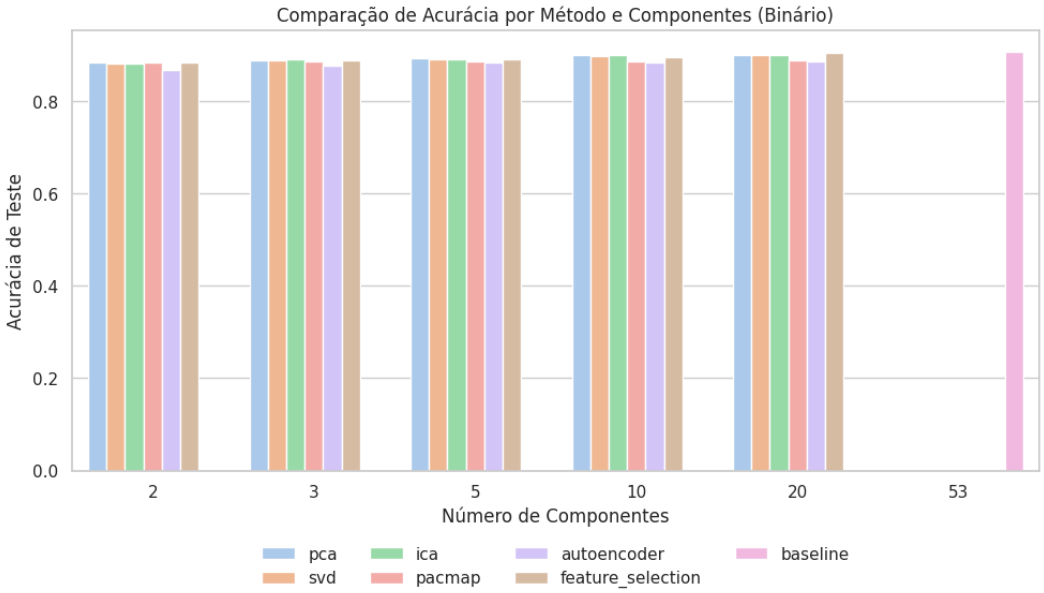


Figura 4.4: Comparação de Acurácia por Método e Componentes no Cenário Binário.

4.2.2 F1-Score

A Figura 4.5 apresenta o desempenho das técnicas em termos de *F1-Score* no cenário binário. Os métodos PCA, SVD e ICA são eficazes, com pontuações elevadas, especialmente para 5 ou mais componentes, indicando sua capacidade de capturar características relevantes para a classificação. O *Autoencoder* apresenta *F1-Scores* inferiores aos métodos lineares, sugerindo que, no cenário binário, as técnicas lineares são mais adequadas. O PaCMAP, embora projetado para preservar tanto relações locais quanto globais, não obteve um *F1-Score* competitivo neste cenário mais simples, sugerindo que sua complexidade não linear pode ser excessiva para o cenário binário. A Seleção de Características apresentou os melhores *F1-Scores*, especialmente com um número maior de componentes, aproximando-se do *baseline* e superando os demais métodos, o que demonstra sua eficácia em selecionar características relevantes para a classificação binária.

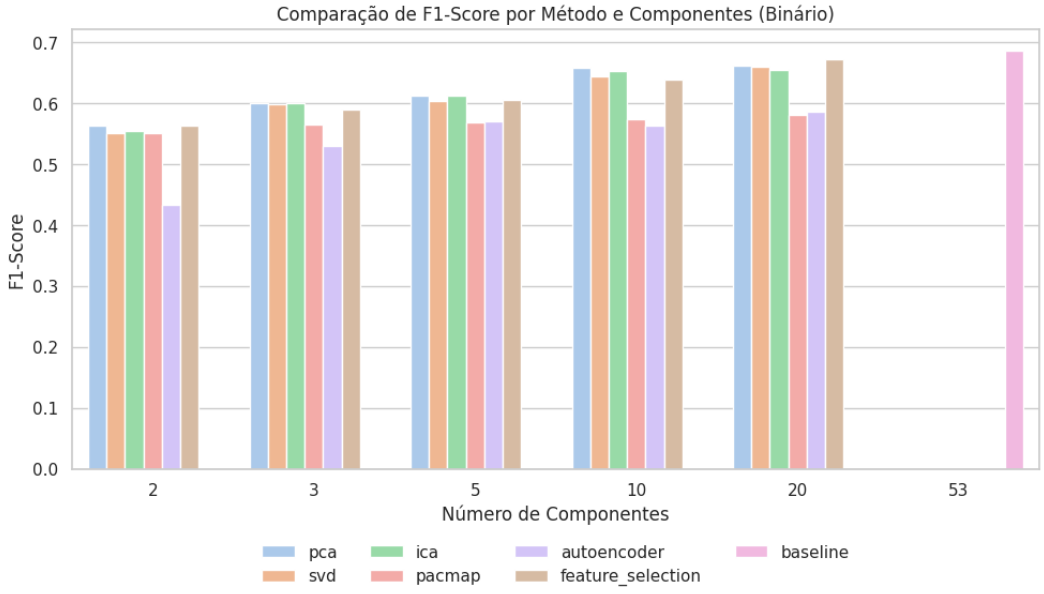


Figura 4.5: Comparação de F1-Score por Método e Componentes no Cenário Binário.

4.2.3 Tempo de Treinamento

A Figura 4.6 mostra o tempo de treinamento para o cenário binário. Os métodos lineares, como PCA, SVD e ICA, continuam a ser os mais rápidos. O *Autoencoder* apresenta tempos de treinamento comparáveis aos métodos lineares, indicando eficiência computacional aceitável. O PaCMAP, especialmente com um número maior de componentes, requer um tempo de treinamento significativamente maior, devido à sua complexidade computacional na preservação de relações não lineares. A Seleção de Características apresentou um tempo de treinamento mais elevado em comparação com os demais métodos, indicando que o processo de seleção é computacionalmente intensivo. Para aplicações onde o tempo de treinamento é crítico, métodos como PCA, SVD, ICA são mais recomendados.

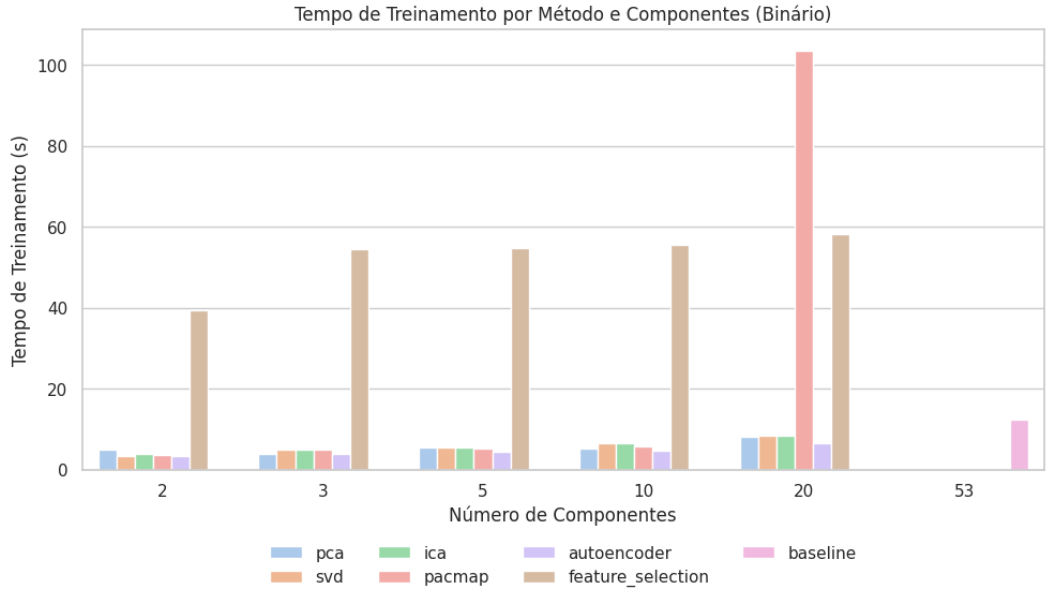


Figura 4.6: Comparação de Tempo de Treinamento por Método e Componentes no Cenário Binário.

4.3 ANÁLISE DO MÉTODO DE SELEÇÃO DE CARACTERÍSTICAS

A técnica de Seleção de Características foi aplicada para identificar as variáveis mais influentes no desempenho de classificação tanto no cenário binário quanto no multiclasse. As Figuras 4.7 e 4.8 apresentam as variáveis mais significativas para cada cenário, permitindo uma interpretação detalhada dos fatores que mais impactam o modelo.

4.3.1 Classificação Binária: Previsão do Tipo de Escola

No cenário de classificação binária, que visa distinguir entre escolas públicas e privadas, destacam-se algumas variáveis relacionadas ao desempenho acadêmico e ao contexto socioeconômico dos alunos. As variáveis mais relevantes foram:

- **faixa_renda_familiar**: Esta variável representa a faixa de renda familiar e aparece como a mais influente, sugerindo que o nível socioeconômico dos estudantes está fortemente associado ao tipo de escola (pública ou privada) frequentada. Isso é consistente com a literatura, que aponta que o acesso a escolas privadas está frequentemente relacionado a famílias com maior renda (Sampaio e Guimarães, 2009).

- **NU_NOTA_COMP4**: Nota na competência 4 da redação, que avalia a habilidade dos alunos em demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação. Esta competência reflete habilidades de estruturação do texto e uso da linguagem formal, possivelmente indicando diferenças nos níveis de preparação entre alunos de diferentes tipos de escola.
- **CO_UF_PROVA**: Código da Unidade da Federação da aplicação da prova. Esta variável pode capturar diferenças regionais no desempenho dos estudantes ou na oferta educacional, que podem influenciar o tipo de escola frequentado. Diferenças entre regiões podem ser significativas em contextos educacionais, influenciando o desempenho e as oportunidades de acesso a diferentes tipos de escolas.
- **NU_NOTA_COMP2**: Nota da competência 2 da redação, que mede a habilidade de compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema dentro do formato dissertativo-argumentativo. Essa competência pode refletir o preparo geral dos estudantes em interpretação de texto e construção de ideias, áreas que podem variar com o tipo de escola.
- **Q007 ("Em sua residência trabalha empregado(a) doméstico(a)?")**: Esta questão do questionário socioeconômico avalia a presença de trabalhadores domésticos na residência do aluno, um indicador indireto de classe social e poder aquisitivo. Essa variável, sendo relevante para a classificação binária, sugere que alunos de escolas privadas podem ter um perfil socioeconômico diferenciado, refletido em aspectos como o emprego de ajuda doméstica.
- **Q008 ("Na sua residência tem banheiro?")**: Esta questão simples, mas reveladora, está relacionada às condições de moradia e higiene, sendo um indicador socioeconômico importante. A presença ou ausência de banheiro pode revelar muito sobre o nível de pobreza e, indiretamente, sobre o acesso ao tipo de escola.
- **Q024 ("Na sua residência tem computador?")**: A presença de um computador em casa pode refletir o nível de acesso a recursos educativos, tecnologias e oportunidades de aprendizado adicionais. Esta variável é particularmente relevante, pois pode indicar um fator diferenciador entre alunos de escolas públicas e privadas, onde o acesso a tecnologias pode ser um diferencial na preparação acadêmica.

Essas variáveis indicam uma combinação de fatores acadêmicos e socioeconômicos que influenciam o tipo de escola frequentado. O destaque para variáveis de desempenho acadêmico, como as notas em competências específicas da redação, sugere que habilidades como argumentação e compreensão textual diferenciam os alunos de escolas públicas e privadas. Por outro lado, as variáveis socioeconômicas, como a faixa de renda familiar, a presença de um computador em casa e a condição de moradia, reforçam a correlação entre o nível socioeconômico e o tipo de escola. Esses achados são consistentes com a literatura, que sugere que a escolha entre escolas públicas e privadas no Brasil é influenciada tanto pelo desempenho escolar quanto pelas condições socioeconômicas do aluno e sua família.

4.3.2 Classificação Multiclasse: Previsão da Faixa de Renda Familiar

No cenário de classificação multiclasse, onde os candidatos são divididos em faixas de renda familiar, observa-se uma maior influência de variáveis relacionadas ao contexto socioeconômico e demográfico. As principais características identificadas incluem:

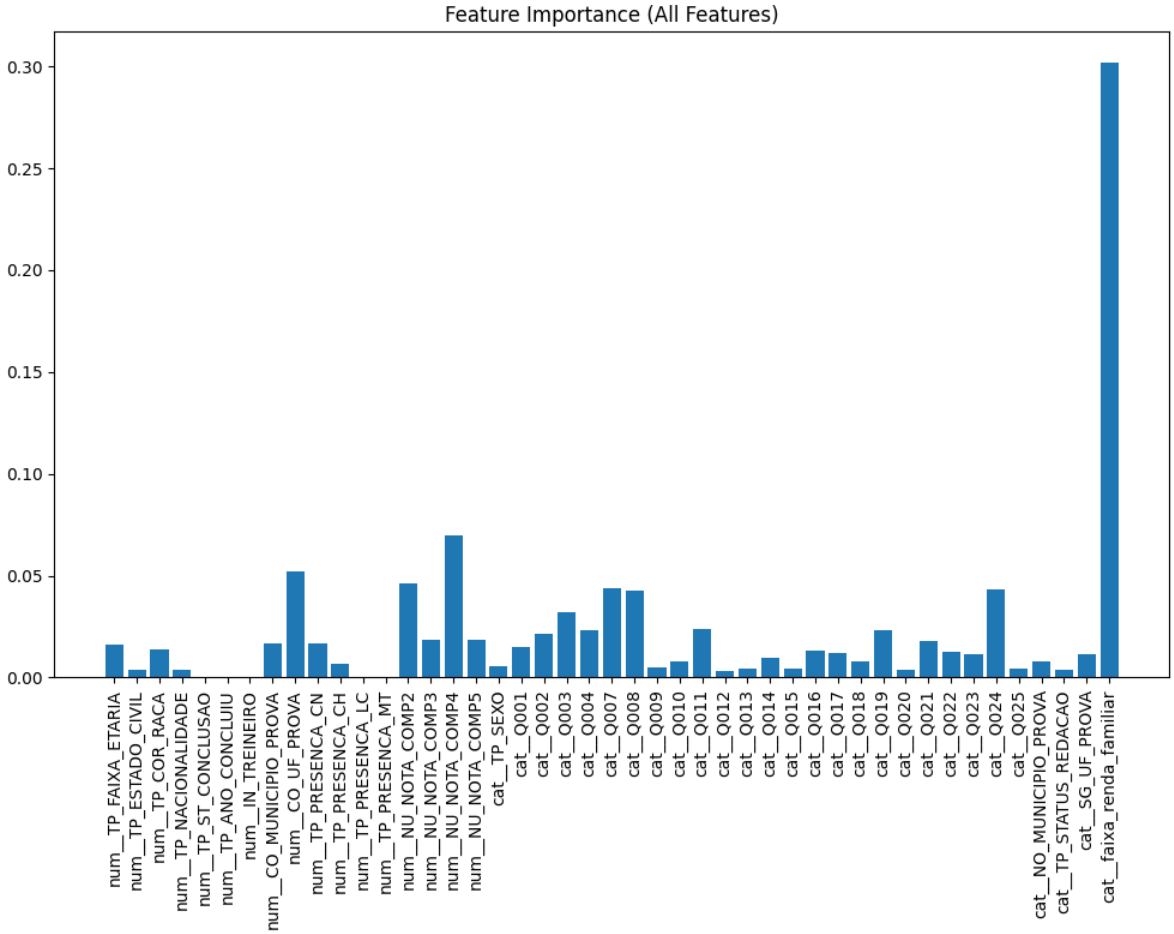


Figura 4.7: Importância das Características para o Cenário Binário.

- **Q010 (“Na sua residência tem carro?”)**: Esta variável indica a posse de um carro pela família, um importante indicador de poder aquisitivo e nível socioeconômico. A presença de um carro é associada a uma renda familiar mais alta, podendo, portanto, discriminar efetivamente entre as faixas de renda.

- **Q008 (“Na sua residência tem banheiro?”)**: A existência de um banheiro em casa reflete condições básicas de moradia e saneamento, que, por sua vez, estão fortemente ligadas ao nível socioeconômico. A presença ou ausência de banheiro contribui para diferenciar as famílias de renda mais baixa das demais.

- **Q024 (“Na sua residência tem computador?”)**: A presença de um computador na residência representa acesso a recursos tecnológicos, que é mais comum em faixas de renda mais altas. Esta variável é relevante para classificar a renda familiar, pois o acesso à tecnologia pode refletir maiores oportunidades educacionais e acesso a informação.

- **Q014 (“Na sua residência tem máquina de lavar roupa?”)**: A posse de uma máquina de lavar roupa reflete conforto e poder aquisitivo, sendo um item comum em famílias de classe média e alta. Sua presença contribui para diferenciar as faixas de renda mais baixas das intermediárias e altas.

- **Q018 (“Na sua residência tem aspirador de pó?”)**: O aspirador de pó é um eletrodoméstico menos comum em residências de baixa renda, sendo encontrado principalmente

em lares de classes média e alta. Essa variável, portanto, contribui para a distinção das famílias em faixas de renda superiores.

- **CO_UF_PROVA**: O código da Unidade da Federação onde a prova foi realizada. Esta variável pode capturar variações regionais que influenciam a renda e as oportunidades econômicas dos candidatos, contribuindo para uma classificação mais precisa das faixas de renda, considerando as disparidades econômicas regionais no Brasil.

Essas variáveis reforçam a relação entre as condições socioeconômicas e a faixa de renda familiar. A inclusão de variáveis relacionadas à posse de bens de consumo, como carro, computador, e eletrodomésticos (máquina de lavar roupa e aspirador de pó), indica que o método de Seleção de Características conseguiu identificar aspectos-chave que refletem o nível de conforto e poder aquisitivo das famílias.

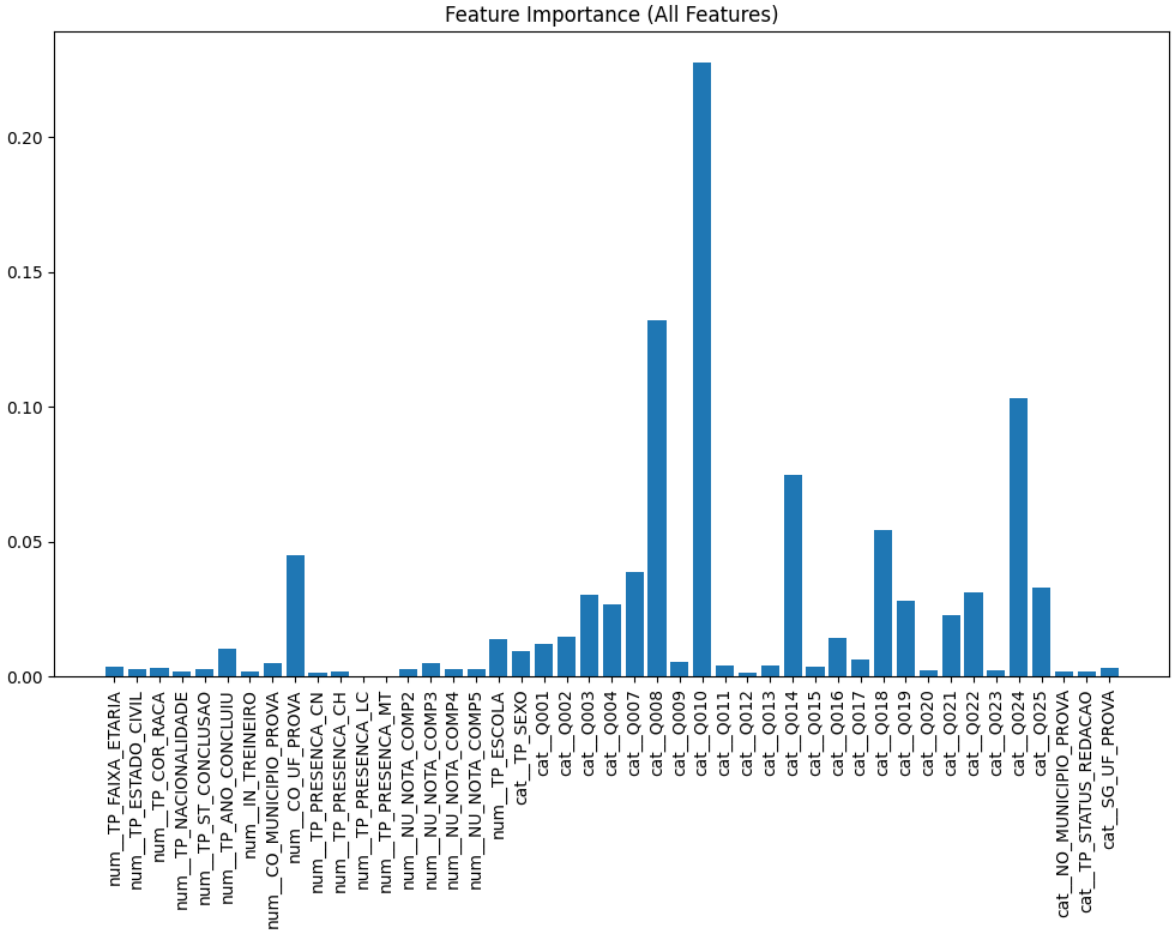


Figura 4.8: Importância das Características para o Cenário Multiclasse.

A técnica de Seleção de Características, portanto, demonstrou-se eficaz para reduzir a dimensionalidade dos dados enquanto preserva variáveis críticas para a precisão dos modelos. Embora o custo computacional desta técnica seja relativamente elevado, a análise sugere que ela pode ser uma opção vantajosa em cenários onde a interpretação das variáveis mais relevantes é essencial para os objetivos da pesquisa, especialmente ao explorar a relação entre condições de vida e nível socioeconômico dos candidatos.

4.4 DISCUSSÃO

Os experimentos demonstram que métodos lineares, como PCA, SVD e ICA, proporcionam um bom equilíbrio entre acurácia, *F1-Score* e tempo de treinamento, sendo adequados tanto para cenários binários quanto multiclasse. O ICA, diferentemente do esperado em alguns contextos, mostrou desempenho similar ao PCA e SVD, especialmente com um número maior de componentes, indicando sua eficácia na captura de variações relevantes para as tarefas de classificação abordadas neste estudo. Métodos não-lineares, como *Autoencoder* e PaCMAP, apresentam um custo computacional mais elevado. O *Autoencoder* mostrou-se competitivo em termos de acurácia e *F1-Score*, especialmente com um número maior de componentes, demonstrando sua capacidade de manter um desempenho elevado mesmo após a redução dimensional. No entanto, o PaCMAP, apesar de seu potencial para capturar relações complexas, não se mostrou vantajoso em termos de acurácia e *F1-Score* nos cenários analisados, sugerindo que sua aplicação pode não ser ideal em problemas onde as relações não-lineares não são predominantes. A Seleção de Características destacou-se na acurácia e no *F1-Score*, muitas vezes alcançando resultados próximos ou superiores aos métodos lineares, demonstrando sua eficácia em selecionar as características mais relevantes para a classificação. Contudo, seu tempo de processamento elevado é um ponto a ser considerado em cenários com restrições de recursos computacionais.

5 CONSIDERAÇÕES FINAIS

Este estudo apresentou uma análise comparativa de seis técnicas de redução de dimensionalidade, aplicadas a um conjunto de dados de alta dimensionalidade derivado dos microdados do ENEM 2022. Através de experimentos sistemáticos, demonstramos que métodos lineares, como o PCA, SVD e ICA, são altamente eficazes na redução da dimensionalidade dos dados, preservando tanto a acurácia quanto o *F1-Score* do modelo de classificação, além de serem eficientes em termos de tempo de treinamento. O ICA, que pode apresentar desempenho inferior em alguns contextos devido à sua suposição de independência entre componentes, neste estudo mostrou desempenho similar ao PCA e SVD, especialmente com um número maior de componentes. Isso pode ser atribuído ao fato de que as características independentes extraídas pelo ICA foram eficazes na captura das variações relevantes nos dados do ENEM, indicando sua capacidade de representar adequadamente as informações necessárias para as tarefas de classificação abordadas.

A Seleção de Características, uma técnica que seleciona as variáveis mais relevantes de forma direta, mostrou-se competitiva em acurácia e *F1-Score*, frequentemente alcançando resultados próximos ou superiores aos métodos lineares, tanto no cenário binário quanto no multiclasse. Contudo, seu tempo de treinamento foi significativamente maior, indicando uma exigência computacional mais elevada quando comparado a métodos lineares.

Métodos não lineares, como o *Autoencoder* e o PaCMAP, mostraram-se úteis para situações em que a preservação de estruturas complexas nos dados é crítica, embora apresentem custos computacionais mais elevados. O *Autoencoder* teve um desempenho próximo aos métodos lineares em termos de acurácia e *F1-Score*, principalmente com um número maior de componentes, demonstrando sua capacidade de manter um desempenho elevado mesmo após a redução dimensional. Por outro lado, o PaCMAP não se mostrou vantajoso em termos de acurácia e *F1-Score* nos cenários analisados, sugerindo que sua aplicação pode não ser ideal em problemas onde as relações não lineares não são predominantes.

Em ambos os cenários de classificação, a técnica de Seleção de Características permitiu uma análise detalhada dos padrões socioeconômicos dos candidatos, destacando variáveis importantes para a tarefa de classificação. Além disso, os métodos de redução de dimensionalidade mostraram-se eficazes na simplificação dos dados sem perda significativa de informação, evidenciando sua eficiência e potencial para simplificar e interpretar grandes conjuntos de dados no contexto educacional.

REFERÊNCIAS

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F. e Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38–44.

Bellman, R. (1958). Dynamic programming and stochastic control processes. *Information and Control*, 1(3):228–239.

Binois, M. e Wycoff, N. (2022). A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Trans. Evol. Learn. Optim.*, 2(2).

Chen, C.-Y., Leu, J.-S. e Prakosa, S. W. (2018). Using autoencoder to facilitate information retention for data dimension reduction. Em *2018 3rd International Conference on Intelligent Green Building and Smart Grid (IGBSG)*, páginas 1–5.

Chen, T. e Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, página 785–794, New York, NY, USA. Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K. e Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Em *North American Chapter of the Association for Computational Linguistics*.

Hinton, G. E. e Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hirasawa, J. G. V. (2023). Aplicação de métodos de redução de dimensionalidade não lineares em classificadores paramétricos e não paramétricos. Trabalho de Conclusão de Curso, Universidade Federal de São Carlos.

Hyvärinen, A. e Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430.

Jia, W., Sun, M., Lian, J. e Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8(3):2663–2693.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, 2nd edition.

Keser, R. K. e Töreyin, B. U. (2019). Autoencoder based dimensionality reduction of feature vectors for object recognition. Em *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, páginas 577–584.

Klema, V. e Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176.

Köppen, M. (2000). The curse of dimensionality. Em *Proceedings of the International Conference on Soft Computing*.

McInnes, L., Healy, J. e Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G. e Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Oliveira, E., JUSTO, W. e Lucena, M. (2024). The dynamics of public high school student performance in ceará: A study of the case of sobral in brazil. *IOSR Journal Of Humanities And Social Science*, 29:25–33.

Sampaio, B. e Guimarães, J. (2009). Diferenças de eficiência entre ensino público e privado no brasil. *Economia Aplicada*, 13(1):45–68.

Santos, B., Saporetti, C. M. e Macedo, B. S. (2023). Analysis of the impact of the pandemic on social inequalities in enem 2019 and 2020 using machine learning. *Semina: Exact and Technological Sciences*, 44(2):1–12.

Wang, J., He, H. e Prokhorov, D. V. (2012). A folded neural network autoencoder for dimensionality reduction. *Procedia Computer Science*, 13:120–127. Proceedings of the International Neural Network Society Winter Conference (INNS-WC2012).

Wang, Y., Huang, H., Rudin, C. e Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

Weikuan, Z., Qiang, L. e Jiawei, W. (2022). A comprehensive review on feature selection strategies for high-dimensional data. *Journal of Machine Learning Research*, 23(5):1–45.