BRUNO ANSELMO GUILHEN

Avaliação de algoritmos para a análise preditiva das notas de Engenharia no ENADE utilizando dados socioeconômicos

São Paulo

2024

BRUNO ANSELMO GUILHEN

Avaliação de algoritmos para a análise preditiva das notas de Engenharia no ENADE utilizando dados socioeconômicos

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção do Título de Mestre em Ciências.

Área de Concentração:

Engenharia Elétrica e Sistemas Digitais.

Orientador:

Prof. Dr. Sergio Takeo Kofuji

São Paulo

2024

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 17 de maio de 2024.

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Guilhen, Bruno Anselmo
Avaliação de algoritmos para análise preditiva das notas de Engenhariano ENADE utilizando dados socioeconômicos / B. A. Guilhen -- São Paulo, 2024.
82 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Inteligência Artificial 2.Regressão linear 3.Aprendizado de Máquina 4.Algoritimos para Regressão Linear I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

Dedicatória

À Rosângela,

Aos meus filhos: Diego, Bruninho, Vitor, Emmanuel (*in memoriam)* e a Nicole,

Ao meu irmão Sérgio (*in memoriam) de quem eu sinto enorme falta,*

À minha irmã Suhaila, aquela que eu confidencio minha vida,

Ao meu pai (in memoriam) e a minha mãe,

só vocês sabem o tamanho do esforço para que esse trabalho se tornasse realidade.

AGRADECIMENTOS

Primeiramente à Deus, que sempre me deu forças para chegar até aqui e nos momentos mais difíceis me reergueu e permitiu que eu continuasse.

Agradeço ao meu orientador, o professor Sergio Takeo Kofuji, pela imensa paciência, amizade, pelas broncas sempre na medida certa e pelos puxões de orelha que me despertaram, me orientaram e me fizeram seguir.

Ao grande amigo Anderson, pela imensa ajuda na condução da pesquisa.

Por fim, agradeço à minha esposa, pela tolerância, pela parceria e o incentivo incondicional ao trabalho.

“Longo é o caminho, difícil é a jornada, estreita a porta, mas a fé remove obstáculos; nada temas: é preciso crer somente!”

Jesus

RESUMO

GUILHEN, Bruno Anselmo. Avaliação de algoritmos para a análise preditiva das notas de Engenharia no ENADE utilizando dados socioeconômicos. 2024. Dissertação (Mestrado em Engenharia Elétrica e Sistemas Digitais) – Escola Politécnica da Universidade de São Paulo, 2024.

O processo de análise preditiva consiste em realizar previsões tomando como base alguns algoritmos de aprendizado de máquina. Um tipo muito comum consiste em utilizar algoritmos de regressão linear aplicados em uma base de dados devidamente tratada. O tratamento dos dados consiste em realizar a análise exploratória dos dados através de análise estatística, com objetivo de reduzir a dimensionalidade dos dados, tratar a multicolinearidade e realizar a devida validação dos dados que serão utilizados para treino e teste dos algoritmos de regressão. A escolha da base de dados também é outro fator de extrema importância para obter resultados estatisticamente válidos. O trabalho utilizou a base de dados fornecida pelo governo federal contendo os dados dos estudantes dos cursos de engenharia que realizaram a prova do ENADE em 2019. Após o devido tratamento estatístico, foram escolhidos algoritmos de regressão linear para o processo de análise preditiva. Os algoritmos selecionados foram: LightGBM, XGBoost e o CatBoost. O critério de escolha foi o balanço entre replicabilidade dos resultados com menor erro e menor tempo de treinamento. Levou-se em conta fatores como R^2 (R-squared), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Median Absolute Error. O treinamento dos algoritmos passou por ajustes dos hiperparâmetros que representam a melhor opção para a melhor resposta. Depois de realizados os ajustes foi possível realizar previsões e retirar insights da base. Do ponto de vista da engenharia, o modelo proposto permite analisar dados socioeconômicos e prever o desempenho do estudante no ENADE. Do ponto de vista social, o trabalho permite que uma instituição universitária consiga analisar e programar suas políticas sociais desde os anos iniciais, neste ponto, os algoritmos permitem predizer a nota esperada com base nas condições que a instituição oferece ao estudante, por exemplo, oferecimento de cotas, bolsas de estudos, bolsa de iniciação científica entre outros.

Palavras-chave: aprendizado de máquina, LightGBM, XGBoost, CatBoost, regressão linear.

ABSTRACT

GUILHEN, Bruno Anselmo. Avaliação de algoritmos para a análise preditiva das notas de Engenharia no ENADE utilizando dados socioeconômicos. 2024. Dissertação (Mestrado em Engenharia Elétrica e Sistemas Digitais) – Escola Politécnica da Universidade de São Paulo, 2024.

The predictive analysis process consists of making predictions based on some machine learning algorithms. A very common type consists of using linear regression algorithms applied to a properly treated database. Data processing consists of carrying out exploratory data analysis through statistical analysis, with the aim of reducing the dimensionality of the data, treating multicollinearity, and carrying out due validation of the data that will be used for training and testing the regression algorithms. The choice of database is also another extremely important factor to obtain statistically valid results. The work used the database provided by the federal government containing data from engineering students who took the ENADE test in 2019. After due statistical treatment, linear regression algorithms were chosen for the predictive analysis process. The algorithms selected were: LightGBM, XGBoost and CatBoost. The choice criterion was the balance between replicability of results with lower error and shorter training time. Factors such as R^2 (R-squared), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Median Absolute Error were taken into account. The training of the algorithms involved adjustments of the hyper parameters that represent the best option for the best response. After the adjustments were made, it was possible to make predictions and extract insights from the base. From an engineering point of view, the proposed model allows analyzing socioeconomic data and predicting student performance in ENADE. From a social point of view, the work allows a university institution to analyze and program its socials policies from the initial years, at this point, the algorithms allow predicting the expected grade based on the conditions that the institution offers to the student, for example, offering quotas, scholarships, scientific initiation scholarships, among others.

Keywords: machine learning, LightGBM, XGBoost, CatBoost, linear regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
ENADE	Exame Nacional de Desempenho de Estudantes
ETL	Extração, Transformação e Carga
DL	Deep Learning
IA	Inteligência Artificial
IES	Instituição de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KNN	K-nearest Neighbors Algorithm
MAE	Erro Absoluto Médio
ML	Machine Learning
MLP	Multilayer Percptron
RF	Random Forest
RMSE	Erro Médio Quadrático
RNA	Redes Neurais Artificiais
SVM	Suport Vector Machine
VIF	Variance Inflation Factor

Sumário

1 INTRODUÇÃO

Segundo o relatório *Data Age 2025* da *International Data Corporation* (IDC), a quantidade de dados gerada pela humanidade vai saltar de 33 ZettaByes (ZB) em 2018 para 175 ZB em 2025 (REINSEL; GANTZ; RYDNING, 2018). Os pesquisadores buscam obter conhecimento programando algoritmos de aprendizado de máquina e aplicando técnicas de mineração de dados em uma grande quantidade de dados (WITTEN et al., 2017).

Observando a abundância dos dados gerados e diante da hipótese de se obter um conjunto de dados consolidado e válido, surgem diversas oportunidades para a utilização de algoritmos na realização de análises preditivas buscando respostas que seriam muito complexas se fossem feitas sem a ajuda do aprendizado de máquina (MAULUD; ABDULAZEEZ, 2020a). Um exemplo, é a base de dados oferecida pelo governo federal brasileiro referente ao Exame Nacional de Desempenho de Estudantes (ENADE), que contempla os dados socioeconômicos e o desempenho dos estudantes medido através das notas das provas (BRASIL, 2024).

Na nota técnica n. 20 de 2019/CGCQES/DAES (BRASIL, 2019) o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) explica a importância do processo de realização do Enade, uma vez que, o exame mede o desempenho dos estudantes que concluem os cursos superiores no Brasil. O exame é dividido por áreas de formação e a nota do exame é composta de duas partes, sendo a primeira parte de conhecimentos básicos representando 25% da nota e a segunda parte de conhecimentos específicos representando 75% da nota. Para cada área de conhecimento o ENADE é realizado de três em três anos (BRASIL, 2019).

Em todo o processo avaliativo existe a necessidade de prever o desempenho dos alunos nos exames futuros, ou seja, a instituição de ensino pode querer entender o comportamento das notas dos alunos de um determinado curso (MAULUD; ABDULAZEEZ, 2020a) bem como, estudar quais os principais fatores que influenciam a nota final dos estudantes que realizam o exame (LIMA et al., 2019). Esse tipo de estudo pode ser feito utilizando análise de dados e o aprendizado de máquina (ROMERO; VENTURA, 2020).

No estudo proposto por Ninaus et al. (2019) bem como no trabalho de Trajano (2023), as principais entradas de variáveis na avaliação do comportamento utilizando aprendizado de máquina são ligados a fatores socioeconômicos. Assim, se uma determinada universidade deseja entender a relação da variável nota utilizando esse tipo de dado na entrada, e como consequência, analisar políticas e estratégias educacionais de forma preditiva, é possível implementar tais ações treinando algoritmos que buscam esse tipo de resposta (MORAIS, 2020).

Uma forma para realizar a análise preditiva das notas e consequentemente buscar melhorar os fatores que influenciam negativamente no desempenho dos estudantes é treinar algoritmos de regressão linear para as previsões (MORETTIN; SINGER, 2022) . Segundo Alyahian & Düstergör (2020) o uso desse tipo de algoritmo se justifica por conta da simplicidade e interpretabilidade, principalmente quando se busca entender a relação entre variáveis independentes e dependentes. A regressão linear se mostra eficiente mesmo quando algumas suposições básicas da estatística não são totalmente atendidas, apresentando robustez nos resultados apresentados pelos algoritmos que utilizam essa técnica (SUNDAY et al., 2020).

Para o treinamento dos algoritmos faz-se necessário escolher corretamente os dados de entrada que realmente podem influenciar no resultado a ser estimado (MAULUD; ABDULAZEEZ, 2020a). Segundo Landes & Magalhães (2018), após escolher a base adequada para a análise os dados devem ser preparados estatisticamente para que possam ser inseridos nos algoritmos que serão treinados. Este processo é conhecido como análise exploratória de dados (MORETTIN; SINGER, 2022), envolve a limpeza dos dados, a remoção de multicolinearidade e a análise da influência das variáveis no processo de regressão linear (ANURADHA; VELMURUGAN, 2015).

Para Fávero & Belfiore (2017) somente após o tratamento correto dos dados é possível escolher corretamente os algoritmos utilizados, bem como, estabelecer os hiperparâmetros que permitem o treinamento e geram as melhores respostas na análise preditiva.

Muitos são os algoritmos de regressão linear que podem ser utilizados para busca de resultados, a escolha e avaliação dos algoritmos depende de fatores e do resultado que se pretende alcançar, bem como, do nível de precisão e de acurácia, entre outros itens (SARKER,

2021). Fatores como erro médio quadrado e tempo de treinamento são as sugestões mais aplicadas para o caso concreto e representam excelentes parâmetros para escolha dos algoritmos (MAULUD; ABDULAZEEZ, 2020b).

Este trabalho realiza uma análise exploratória de dados seguido da avaliação e comparação entre algoritmos de regressão linear para realizar as predições. De acordo com Romero & Ventura (2020) os algoritmos de regressão linear utilizados no processo de análise preditiva de dados precisam de diversos ajustes para que a saída não seja comprometida por variáveis que possuem vieses. Para Fávero & Belfiore (2017) o processo de limpeza de dados e o ajuste dos algoritmos representa parte significativa do trabalho. Uma vez que esses parâmetros estão estabelecidos corretamente, os algoritmos podem ser treinados para fornecer as respostas mais adequadas.

Para realizar a mineração de dados e obter as melhores respostas dos algoritmos todas as etapas de análise precisam estar corretamente ajustadas, esse é o motivo de trabalhar uma análise exploratória mais aprofundada, uma vez que, as respostas tendem a ser mais precisas quando os ajustes estão livres de erros e multicolinearidades (WITTEN et al., 2017).

Diante desse contexto e na busca de métodos para validar o aprendizado as principais perguntas que delimitaram o campo da pesquisa são:

- Quais métodos de análise exploratória melhor definem as variáveis para a análise exploratória dos dados?
- Qual a melhor técnica de análise preditiva pode ser utilizada para obter respostas utilizando os dados escolhidos?
- Quais são os algoritmos que melhor representam as respostas esperadas?
- Quais são os resultados mais relevantes a serem observados após o treinamento dos algoritmos?

Essa pesquisa se justifica, uma vez que esse problema tem sido investigado por diversos autores, porém muitos utilizaram técnicas estatísticas com alcance limitado. Nos últimos anos técnicas mais avançadas baseadas no aprendizado de máquina tem sido utilizada, porém com aplicações limitadas. Essa pesquisa busca avaliar técnicas mais modernas baseadas em inteligência artificial.

Tomando como base os questionamentos apresentados, a pesquisa segue com o objetivo de realizar a modelagem do sistema e a experimentação para a validação.

1.1 Objetivos

O objetivo geral da pesquisa é avaliar o desempenho acadêmico dos alunos da engenharia baseado nos fatores socioeconômicos da base de dados do ENADE.

De forma específica podem ser citados os seguintes objetivos para essa pesquisa:

- Analisar de forma exploratória os dados da base do ENADE para entender a sua natureza e especificar quais podem ser utilizados na construção do *dataset*.
- Analisar comparativamente algoritmos de aprendizagem de máquina para o desempenho acadêmico.
- Avaliar o impacto dos dados socioeconômicos relacionados com a nota do ENADE.

1.2 Metodologia

Para o trabalho em curso, é utilizada a base fornecida pelo INEP que contém as notas e os dados socioeconômicos de todos os alunos que fizeram a prova de Engenharia no ano de 2019 (BRASIL, 2024). Para a construção do modelo e análise dos dados, as seguintes sequências de ações são realizadas (SUNDAY et al., 2020):

- Análise exploratória de dados: a análise exploratória consiste em uma profunda análise estatística para limpeza de dados, análise de regressão, testes estatísticos, análise de multicolinearidade e modelagem preditiva;
- Avaliação e escolha dos algoritmos de regressão linear disponíveis: a avaliação ocorre utilizando métricas do tipo Erro Médio Quadrado (RMSE), Erro Absoluto Médio (MAE) e Coeficiente de Determinação (R^2), pois fornecem as métricas mais robustas de precisão do modelo de regressão linear múltipla;
- Análise e comparação dos algoritmos selecionados: utilizando a base de treino e teste os resultados das métricas da etapa anterior serão utilizados para

comparar os algoritmos e verificar quais fornecem a melhor capacidade preditiva, essa etapa também leva em consideração o tempo de treinamento e a replicabilidade do algoritmo;

- Mineração dos dados para obtenção dos resultados: a interpretação dos resultados leva em conta os dados socioeconômicos e a nota final na prova do ENADE que é a variável alvo, os resultados podem ser utilizados para entender a oferta de serviços pela entidade educacional.

A Figura 1 é uma representação gráfica das etapas aplicadas.

Figura 1. Metodologia empregada na construção do trabalho.

Fonte: Adaptado de (SUNDAY et al., 2020)

1.3 Organização da dissertação

Essa dissertação está organizada em 6 capítulos dispostos da seguinte forma: O capítulo 1 apresenta a Introdução, objetivos gerais e específicos e metodologia. O capítulo 2 apresenta a fundamentação teórica, o capítulo 3 apresenta uma comparação com os principais trabalhos referenciados, o capítulo 4 apresenta a proposta deste trabalho, o capítulo 5 apresenta o experimento e discute os resultados do trabalho, enquanto o capítulo 6 lista as conclusões, limitações e rumos que essa pesquisa pode tomar.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo revisar os principais conceitos aplicados na estrutura da pesquisa, são tópicos que se relacionam com a inteligência artificial, a análise exploratória de dados, com análise preditiva e a fundamentação estatística aplicada para validação da proposta, por exemplo, os algoritmos de regressão linear.

2.1 Inteligência Artificial (IA)

A inteligência artificial tornou-se realidade em todos os campos de pesquisa da humanidade, segundo Ciolacu (2018) a IA pode ser comparada à descoberta da eletricidade, ou seja, elemento presente em diversas ações humanas. Neste sentido é que a IA vem desempenhando o papel principal nas tarefas que podem ser automatizadas, por exemplo, sistemas de recomendação de filmes da Netflix, de vídeos do Youtube, sistemas que utilizam dados de entrada para entender as necessidades dos usuários e realizar a previsão de perfis. (TOZADORE, 2020)

Os sistemas inteligentes que funcionam baseados em IA possuem formas clássicas de aprendizagem que estão bem solidificadas e em uso no mercado, conhecida como *machine learning* (ML) ou aprendizado de máquina (AM) ela se divide em (RUSSEL, 2013):

- Aprendizado não supervisionado: clusterização
- Aprendizado supervisionado: por classificação
- Aprendizado supervisionado: por regressão

A Figura 2 a seguir mostra, esquematicamente, como os algoritmos de aprendizado são categorizados.

Figura 2.Tipos de Aprendizado de máquina

Fonte: Adaptado de (FÁVERO; BELFIORE, 2017)

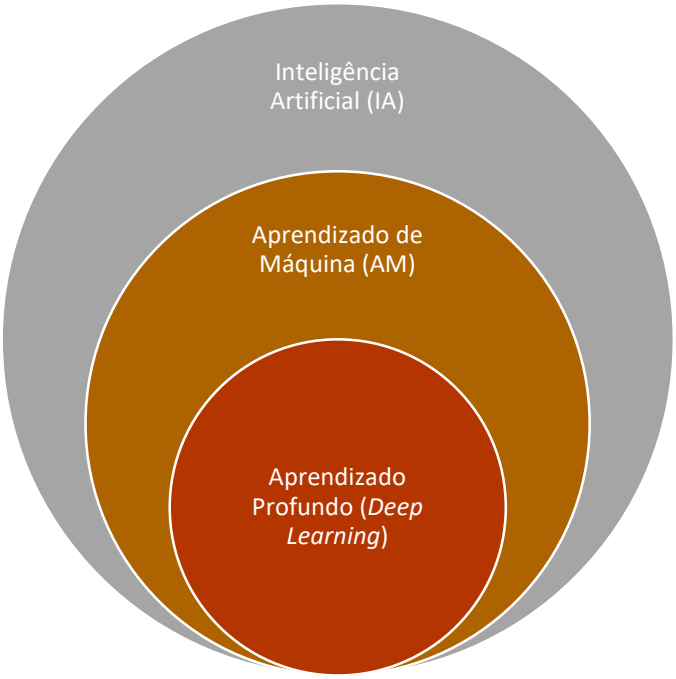
O aprendizado supervisionado é aquele baseado em modelos com evidências prévias. O sistema considera uma entrada padrão de dados com saídas conhecidas. Por exemplo, quando o sistema precisa aprender o que é um cachorro, realiza-se vários *inputs* com imagens de cachorro, mesmo variando alguns fatores como cores, tamanhos, o objeto é sempre o mesmo (patas, pelos, formato do rosto, cauda), esse processo é repetido até que o algoritmo aprenda a reconhecer um cachorro dentre várias outras imagens. Já o aprendizado não supervisionado os algoritmos procuram padrões ocultos, sem nenhuma base de dados anterior, uma base de dados é apresentada e o sistema inicia uma separação dos grupos até conseguir atingir determinados padrões de aprendizado (MORETTIN; SINGER, 2022).

2.2 Aprendizado de Máquina (AM) ou *Machine Learning (ML)*

O aprendizado de máquina é um campo dentro da IA que objetiva desenvolver e treinar algoritmos utilizados em sistemas inteligentes (JANIESCH; ZSCHECH; HEINRICH, [s.d.]). Enquanto a IA é generalista, utilizando uma diversidade de técnicas e desenvolvimento de sistemas inteligentes para simular tarefas humanas o AM concentra na construção de modelos e no desenvolvimento de algoritmos (SHI, 2023).

Da mesma forma que o AM está contido na IA as técnicas de aprendizado profundo, conhecidas como *deep learning (DL) são uma especificação do AM* (YE, 2021). A Figura 3 a seguir ilustra a relação entre essas ciências.

Figura 3 - Relação entre IA, AM e Aprendizado profundo.



Fonte: Adaptado de (NAIM et al., 2023).

Conforme mostrado na Figura 2 os algoritmos de AM podem ser treinados baseados em sistemas de aprendizado supervisionado ou não supervisionado (LI; ZOU, 2022). O aprendizado supervisionado demanda mais recursos e tempo na preparação dos dados e mais cuidado no tratamento da base de entrada, por outro lado, os sistemas não supervisionados requerem mais atenção na construção dos parâmetros dos algoritmos (NAIM et al., 2023).

O aprendizado de máquina pode ser moldado por meio de um sistema determinístico que busca estabelecer a relação linear entre variáveis independentes e dependentes (FRIEDMAN; HASTIE; TIBSHIRANI, 2010). Um tipo comum de sistema determinístico são os sistemas de aprendizado supervisionado que utilizam dados rotulados para o treinamento (HUANG et al., 2023). Uma variável dependente, também chamada de variável alvo, é previamente conhecida e utilizada para nortear o processo de aprendizado, esses sistemas são conhecidos como sistemas de aprendizado supervisionado baseados em regressão linear (IGHALO; ADENIYI; MARQUES, 2020).

2.3 Sistema de aprendizado supervisionado baseado em regressão linear

Os sistemas de regressão linear são comumente utilizados para predições, uma vez que, são treinados utilizando dados rotulados que aprendem a relação entre as variáveis para realizar previsões precisas com dados novos e não observados (SYAH; NAFSIAH; SADDHONO, 2023). A vantagem de utilizar esse tipo de sistema é a possibilidade de aplicação em diversas bases de dados, desde bases com poucas variáveis até bases mais robustas com muitas variáveis (ALYAHYAN; DÜŞTEGÖR, 2020). Uma outra característica dos sistemas que utilizam regressão linear é que são projetados para a previsão de valores contínuos associando variáveis independentes (SARKER, 2021).

Para construir um sistema de regressão linear os seguintes componentes devem estar presentes (ZHOU; JIANG; LIU, 2021):

- Extração de dados: processo de coleta dos dados relevantes para o problema a ser analisado. A coleta busca os dados de forma bruta, contendo as variáveis independentes e a variável dependente que é o alvo do problema (WITTEN et al., 2017);
- Transformação de dados: etapa correspondente à limpeza e pré-processamento que verifica a presença de valores nulos, em branco, realiza a extração dos *outliers* e a normalização dos dados (WITTEN et al., 2017);
- Escolha do modelo de regressão: etapa que visa escolher o tipo de regressão a ser utilizada na base. Os tipos podem variar entre regressão linear simples, múltipla, multivariada etc. A escolha depende da complexidade existente na relação entre as variáveis (MORETTIN; SINGER, 2022).
- Treinamento do modelo: etapa que corresponde ao treinamento do algoritmo escolhido e a busca dos parâmetros que melhor associam os valores previstos com os valores reais (FÁVERO; BELFIORE, 2017);
- Avaliação do algoritmo: etapa que consiste em avaliar os resultados do modelo tomando como base algumas métricas de desempenho conhecidas, tais como, coeficiente de determinação múltipla (R^2), erro médio quadrático (RMSE), erro absoluto médio (MAE) e tempo de treinamento. Essas métricas são utilizadas para medir o padrão de generalização do modelo (FÁVERO; BELFIORE, 2017).

- Ajuste do Algoritmo: etapa que visa consolidar o que foi extraído da extração e transformação de dados (análise dos dados) seguida dos ajustes dos hiperparâmetros, consiste em modificar variáveis e parâmetros (SYAH; NAFSIAH; SADDHONO, 2023);
- Implementação e resultados: consiste em extrair os resultados do modelo implementado e realizar a validação (MORETTIN; SINGER, 2022).

Segundo Maulud & Abdulazeez (2020) os sistemas baseados em regressão linear podem ser de regressão simples, quando utilizam apenas uma variável independente para prever uma única variável dependente (alvo), os sistemas podem ser de regressão múltipla, quando múltiplas variáveis independentes são utilizadas para prever uma variável dependente. Por fim, os sistemas de regressão linear polinomial que são um caso especial de regressão linear múltipla que modela as relações não lineares entre as variáveis.

Para o correto funcionamento de um sistema de regressão linear faz-se necessário realizar um tratamento estatístico prévio nos dados, processo conhecido como análise exploratória dos dados (MORETTIN; SINGER, 2022). Para Witten et al., (2017) a análise exploratória de dados está dentro do contexto do *Extraction*, *Transformation*, and *Loading (ETL*), termo para designar um processo conhecido em português como extração, transformação e carga. Neste caso, a extração e transformação dos dados é a etapa que valida as informações para serem utilizados em sistemas de aprendizado de máquina, tais como, os baseados em regressão linear (ALDERA et al., 2021).

Assim, antes da escolha dos algoritmos de regressão linear para a criação de um sistema de aprendizado de máquina faz-se necessário realizar a devida limpeza e transformação dos dados (MAPHOSA; DOORSAMY; PAUL, 2023).

2.4 Análise exploratória de dados

Todo sistema baseado em inteligência artificial que utiliza aprendizado supervisionado precisa receber dados para o treinamento dos algoritmos (HUANG et al., 2023). Porém, os dados que alimentam o sistema precisam estar livres de vieses que atrapalham o treinamento e corrompem os resultados (IGHALO; ADENIYI; MARQUES, 2020). Por essa razão, que o processo de análise dos dados e sua consequente limpeza é importante.

As principais etapas realizadas no processo de análise exploratória de dados estão listadas a seguir (ROMERO; VENTURA, 2020):

- Tratamento e limpeza dos dados: análise da base de dados na busca de dados inconsistentes, campos em branco, *outliers* (ZHAI et al., 2022);
- Redução de dimensionalidade: esse processo normalmente tem como resultado a redução da quantidade de variáveis da base escolhida, ou seja, o número de colunas que não possuem relação significativa com a variável alvo (SYAH; NAFSIAH; SADDHONO, 2019);
- Normalização dos dados: para trabalhar com alguns algoritmos de aprendizado de máquina, principalmente os que utilizam regressão linear, é importante que as variáveis tenham a mesma escala. A normalização visa criar um escalonamento para um intervalo determinado de forma a impedir que variáveis com escalas muito grandes ou muito pequenas influenciem a regressão (IGHALO; ADENIYI; MARQUES, 2020);
- Padronização dos dados: a padronização dos dados é a aplicação de técnicas básicas da estatística para a transformação dos dados de forma que os valores tenham média zero e desvio padrão um, validando as inferências feitas na regressão linear (IGHALO; ADENIYI; MARQUES, 2020). A padronização serve para padronizar os dados em torno de zero. Como essa ação não modifica a distribuição dos dados, essa técnica possui aplicação bastante relevante quando as variáveis do modelo possuem escalas muito distintas. Nesse caso, a padronização permite inserir a mesma importância relativa para os dados e evita que variáveis com escalas distintas influenciem no modelo (FRIEDMAN; HASTIE; TIBSHIRANI, 2010);

Na etapa de normalização e padronização dos dados diversos testes são aplicados para garantir estatisticamente que as variáveis do modelo seguem uma distribuição normal e podem ser utilizadas para as etapas de treinamento e teste dos algoritmos (HASSAN; ELKORANY; WASSIF, 2022). Os principais testes utilizados nessas etapas são discutidos no subtópico a seguir.

2.4.1 Testes de hipótese e testes paramétricos

Segundo Fávero & Belfiore (2017) um teste de hipótese representa uma suposição sobre um parâmetro, uma vez que, não é viável testar toda a população. A suposição sobre o determinado parâmetro pode ser a média, o desvio-padrão, o coeficiente de correlação etc.

O teste de hipótese aceita ou rejeita determinada hipótese utilizando uma amostra retirada da população, neste caso, é necessário definir alguns fatores primordiais para que erros da amostra não representem a população (BATCH; ELMQVIST, 2018). O teste de hipótese é construído utilizando basicamente dois parâmetros, a hipótese nula (H_0) e a hipótese alternativa (H_1), nesse caso define a hipótese nula a ser testada da amostra aleatória e o teste serve para comprovar ou rejeitar a hipótese (FÁVERO; BELFIORE, 2017).

Para Morettin & Singer (2022) as principais etapas de um teste de hipótese são:

- Escolha da hipótese: determina-se uma hipótese nula (H_0) e a hipótese alternativa (H_1). A hipótese alternativa é que vai ser testada e a hipótese nula será ou não rejeitada. Neste caso, a hipótese nula indica que a afirmação não possui relação ou que não existe associação entre as variáveis (SALEH; LAYOUS; REPUBLIC, 2022);
- Escolha do teste estatístico: de acordo com o tamanho da amostra e com o nível de associação que se pretende realizar será necessário escolher um tipo de teste para aceitar ou rejeitar a hipótese nula (SILVA-LUGO; WARNER; GALINDO, 2022);
- Escolha do nível de significância: o nível de significância, representado por α, está relacionado com o nível de precisão que se busca ao testar um conjunto de dados e compará-los com a hipótese nula, ou seja, a significância é a probabilidade de se rejeitar a hipótese nula. Normalmente o α pode ser 5% (0,05) ou 1% (0,01) (TRINCHERO; CANAVERO, 2021);
- Escolha da regra de decisão: regra utilizada para verificar se a hipótese nula será rejeitada ou não. Normalmente o teste comparativo utilizado é o p-valor (regra que compara o valor de α com o valor analisado) ou o valor crítico (regra

que estabelece um ponto do espaço amostral e compara com o valor analisado) (TRAWINSKI et al., 2017).

Um teste de hipótese pode ser paramétrico ou não-paramétrico (LANZANTE, 2021). Os testes paramétricos são executados sobre parâmetros da população e são aplicados em dados quantitativos, por exemplo, os testes de Kolmogorov-Smirnov, de Shapiro-Wilk e de Shapiro-Francia que são testes que verificam a normalidade univariada dos dados (HONG et al., 2024).

3 TRABALHOS RELACIONADOS

O objetivo deste capítulo é listar os principais trabalhos que se relacionam com o tema proposto. Seguindo o panorama apresentado na Figura 1, que mostra as etapas estruturadas na construção deste texto, alguns casos de análise exploratória de dados e os mecanismos de validação de variáveis aplicadas na regressão linear são listados. Também são comentados alguns trabalhos que utilizam a mesma base de dados com outras abordagens e as pesquisas que utilizam e discutem as vantagens e aplicações dos algoritmos que podem ser adaptados para esta dissertação.

No trabalho de Alyahyan & Düstergör (2020) é conduzida uma abordagem para predição de sucesso acadêmico baseado na média das notas dos alunos durante a graduação. O trabalho leva em conta diversos fatores atrelados à vida acadêmica como variáveis independentes. Os autores fazem a aquisição e limpeza dos dados utilizando estatística clássica e medidas como média, mediana e p-valor, bem como, a utilização de normalização e discretização para o processo de tratamento dos dados. No trabalho de Alyahyan & Düstergör (2020), a predição utiliza três categorias de algoritmos: classificação (árvore de decisão, algoritmos Bayesianos, rede neural artificial, *K-nearest Neighbors Algorithm)* regressão e clusterização (X-means). No artigo, os dados são generalizados e não passam por uma profunda limpeza estatística, por essa razão que os algoritmos de classificação mostram resultados acima de 60% de acurácia, enquanto os de regressão e clusterização os resultados estão na casa dos 20%. As limitações do trabalho citado encontram-se na realização de um tratamento estatístico parcial e por utilizar uma base de dados pequena, além de utilizar algoritmos de ML de classificação e regressão. Esse ponto difere da pesquisa proposta nesse trabalho, uma vez que o tratamento estatístico é mais robusto, com uma base de dados grande aplicando apenas algoritmos de regressão linear múltipla.

No trabalho proposto por Souza & Cazella (2022), assim como no trabalho proposto por Alyahyan; Düştegör (2020), as etapas utilizadas são muito parecidas. Os trabalhos possuem como variável alvo o desempenho acadêmico dos alunos, realizam o processo de extração, transformação dos dados e realizam análise estatística com normalização. Porém, no trabalho de Souza & Cazella (2022) os algoritmos observados são apenas de regressão linear: árvores de decisão, regressão linear simples, R*andom Forest* (RF*), Suport Vector*

Machine (SVM) e Redes Neurais Artificiais do tipo *Multilayer Perceptron* (MLP). Os resultados do treinamento dos algoritmos apresentam níveis de precisão e acurácia altos, valores acima de 90%. Porém, a base de dados do trabalho, mesmo contando com 33 colunas e 694 linhas, está limitada a notas das disciplinas de língua portuguesa e matemática, o que não exige uma abordagem estatística mais robusta para utilizar os algoritmos.

No estudo conduzido por Maphosa; Doorsamy; Paul (2023) tem-se uma excelente aplicação da análise exploratória de dados para avaliar o desempenho acadêmico dos estudantes de engenharia ao longo da graduação. Nesse ponto os autores percorrem as fases de extração, transformação, carregamento, análise estatística, discretização e normalização dos dados até o momento de treinar a ML. O autor utiliza regressão linear para correlacionar o desempenho acadêmico com as variáveis cor da pele, raça, gênero e idade, entre outras. A base de dados possui 4242 linhas e 16 colunas (variáveis) e o autor realiza a análise exploratória utilizando bibliotecas Python.

Quando se trata de treinar algoritmos de regressão linear com uma base de dados mais robusta o trabalho proposto por Zhou; Jiang; Liu (2021) mostra as etapas de análise de dados mais detalhadas. Mantendo a conhecida estrutura de mineração de dados ETL o trabalho divide em 3 passos o processo de análise e utiliza parâmetros estatísticos baseados em estimativa, média, variância e R-quadrado. Um ponto importante desse trabalho é que a base de dados é massiva, neste caso, as técnicas de regressão linear são mais ajustadas até chegar à etapa de treinamento. A diferença para este trabalho é que os autores avaliaram o desempenho em disciplinas isoladas e não trataram parâmetros dos algoritmos.

No tocante aos critérios de escolha dos algoritmos, o trabalho proposto por Hassan; Elkorany; Wassif (2022) ilustra os principais parâmetros para a avaliação e seleção de algoritmos. Critérios de avaliação, tais como, *Mean Absolut Error* (MAE), R-quadrado e RMSE são listados e desenvolvidos no trabalho. Esse tratamento estatístico fundamenta a base para tratar os dados e selecionar variáveis que são utilizadas nos sistemas de aprendizado de máquina baseados em regressão linear. Essa pesquisa possui uma base de dados pequena e mostra apenas como tratar corretamente parâmetros de algoritmos de regressão linear.

Depois de realizar a análise estatística e escolher a categoria dos algoritmos o presente trabalho precisa entender quais são os melhores algoritmos para realizar as ações e como ajustar os hiperparâmetros para obter os melhores resultados. No trabalho de Kumar;

Dhanalakshmi (2023) o autor realiza treinamento e ajustes de algoritmos de ML utilizando especificamente os algoritmos XGBoost e CatBoost. A comparação dos algoritmos, os ajustes e o treinamento, apesar de utilizar uma base de dados pequena, fornecem importantes parâmetros para a aplicação em outras bases de dados, por exemplo, parâmetros de erro, porcentagem de treino e teste para ajuste dos algoritmos, tempo de resposta e treinamento de cada um deles são os principais dados retirados do trabalho. Embora os autores analisem as notas de todas as disciplinas do curso, os cursos são apenas remotos (EAD) e a base de dados é pequena.

O trabalho de Amin et al. (2023) possui importante contribuição no que se refere ao ajuste de parâmetros para o treinamento de algoritmos de ML no ambiente de aprendizado, tais como, número de camadas para não ocorrer falhas no treinamento (*overfit)*, tamanho da base de treino e teste. Esse trabalho faz uma avaliação de dados de cursos *on-line* e apresenta uma base com grande volume, sendo um fator importante para contribuir com a pesquisa.

O trabalho de Zahedi et al. (2021) oferece uma visão mais aprofundada nos ajustes dos parâmetros e aplica em diversas categorias de algoritmos (classificação, regressão e clusterização). No trabalho, os ajustes nos algoritmos de regressão linear utilizando XGBoost são os que mais contribuem com o trabalho de pesquisa, uma vez que, as técnicas estatísticas e de ajustes são relevantes.

4 PROPOSTA DO TRABALHO E IMPLEMENTAÇÃO

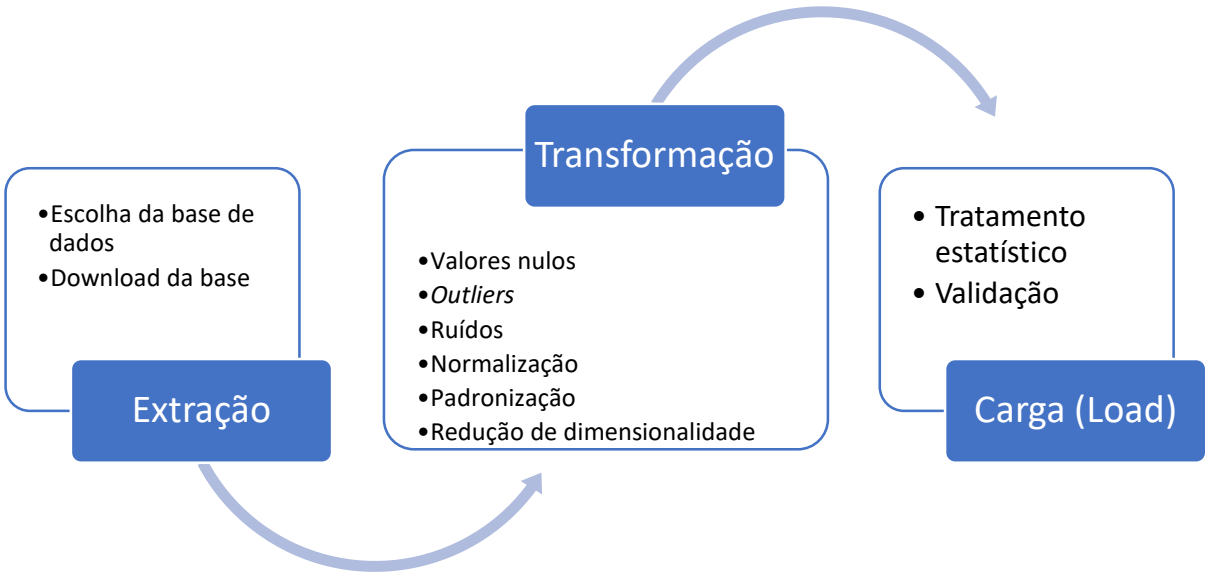
A proposta desse trabalho de pesquisa é realizar uma comparação dos principais algoritmos de aprendizado de máquina que utilizam regressão linear quando aplicados na base de dados dos alunos da engenharia que realizaram a prova do ENADE 2019.

A base de dados é composta dos dados socioeconômicos extraídos dos questionários aplicados quando os alunos se habilitam para realizar a prova. Essa base possui 136.471 linhas e 194 colunas (BRASIL, 2024).

Por ser uma base grande e adversa a primeira etapa proposta para o trabalho é uma validação estatística feita através da análise exploratória dos dados, seguida da normalização e padronização dos dados, da redução de dimensionalidade, da análise de multicolinearidade e da validação.

A análise exploratória de dados pode ser vista, do ponto de vista da mineração de dados (TRINCHERO; CANAVERO, 2021), como a etapa preparatória do processo de ETL. Nas fases de extração e transformação de dados é que devem ocorrer esse tipo de processamento (SILVA-LUGO; WARNER; GALINDO, 2022). A Figura 4 faz uma ilustração desse processo.

Figura 4 - Processo de ETL



Fonte: Adaptado de (ALYAHYAN; DÜŞTEGÖR, 2020)

A base de dados da pesquisa é coletada do site do INEP e a variável alvo é a nota bruta da prova composta das seguintes partes: 25% correspondente da média ponderada da

formação geral e 75% da média ponderada do componente específico. Para explicar essa nota, inicialmente a base conta com 194 variáveis independentes.

A base de dados está dividida nas seguintes partes (BRASIL, 2024):

- Informações da instituição de ensino superior e do curso;
- Dados do curso;
- Informações do estudante;
- Tipos de presença;
- Notas na formação geral e componente específico;
- Questionário do estudante;
- Licenciaturas.

Depois da coleta dos dados, o trabalho seguiu para estabelecer as variáveis que devem constar na análise de regressão. No primeiro momento, aplica-se a técnica de redução de dimensionalidade e a nova base é reduzida de 194 para 34 colunas. A redução de dimensionalidade remove as colunas da base, porém o número de linhas permanece o mesmo, outra observação é que valores nulos e com erros foram tratados pelo governo federal antes da publicação dos dados, não sendo necessário realizar esse tipo de limpeza nos dados.

O critério usado para a redução é a separação das variáveis que se relacionam com as notas e os dados socioeconômicos do restante da base. A Tabela 1 ilustra o resultado da primeira redução de dimensionalidade.

Tabela 1 –Aplicação de redução de dimensionalidade na base de dados do ENADE.

Variável	Categoria
1. Nota bruta da prova (y) - Média ponderada da formação geral (25%) e componente específico (75%)	Valor de 0 a 100
2. Tipo da categoria administrativa da IES	Pública federal, pública estadual, pública municipal, privada com fins lucrativos ou especial

Variável	Categoria
3. Nome do Estado à qual pertence a IES	RO, AC, AM, RR, PA, AP, TO, MA, PI, CE, RN, PB, PE, AL, SE, BA, MG, ES, RJ, SP, PR, SC, RS, MS, MT, GO e DF
4. Descrição da modalidade de ensino do curso	EaD ou presencial
5. Gratuidade do curso	Sim ou não
6. Idade que o aluno completa no ano de referência do ENADE.	até 17, 18 – 21, 22 – 25, 26 – 29, 30 – 33, 34 – 37, 38 – 41, 42 – 45, 46 – 49, 50 – 53, 54 – 57, 58 – 61, 62 – 65, maior do que 65
7. Sexo do aluno	Feminino ou masculino
8. Ano de início da graduação	Faixa de 1990 até 2019
9. Estado civil	Solteiro, casado, divorciado, viúvo ou outro
10. Cor ou raça	Branca, preta, amarela, parda, indígena ou não declarado
11. Nacionalidade	Brasileiro, brasileiro naturalizado ou estrangeiro
12. Escolarização do pai	Nenhum, Ensino Fundamental até 5º ano, Ensino Fundamental do 6º ao 9º ano, Ensino Médio, Graduação, Pós-graduação
13. Escolarização da mãe	Nenhum, Ensino Fundamental até 5º ano, Ensino Fundamental do 6º ao 9º ano, Ensino Médio, Graduação, Pós-graduação
14. Onde e com quem o aluno mora	Casa ou apartamento, sozinho, Casa ou apartamento, com pais e/ou parentes, Casa ou

Variável	Categoria
	apartamento, com cônjuge e/ou filhos, Casa ou apartamento, com outras pessoas, Alojamento universitário da própria instituição ou outros tipos de habitação
15. Número de pessoas que moram com o aluno	Nenhuma, uma, duas, três, quatro, cinco, seis, sete ou mais
16. Renda total da família, incluindo o aluno	Até 1,5 salários-mínimos (sm), de 1,5 a 3 sm, de 3 a 4,5 sm, de 4,5 a 6 sm, 6 a 10 sm, de 10 a 30 sm, acima de 30 sm
17. Renda e forma de financiamento dos gastos do aluno	Sem renda, com gastos financiados por programas governamentais; sem renda, com gastos financiados pela minha família ou por outras pessoas; com renda e ajuda a família ou outras pessoas para financiar os gastos; com renda, sem ajuda para financiar os gastos; com renda, contribui com o sustento da família e; principal responsável pelo sustento da família.
18. Situação e forma de trabalho do aluno (exceto estágio ou bolsas)	Não está trabalhando, trabalho eventual, trabalho com até 20 horas semanais, trabalho entre 21 a 39 horas semanais, trabalho de 40 horas semanais ou mais
19. Bolsa de estudos ou forma de financiamento para custear todas ou a maior parte das mensalidades	Nenhum, curso gratuito; nenhum, curso não gratuito; ProUni integral; ProUni parcial, apenas; FIES, apenas; ProUni Parcial e FIES; bolsa governamental; bolsa própria da instituição; bolsa oferecida por outra entidade;

Variável	Categoria
	financiamento pela própria instituição ou; financiamento bancário
20. Tipo de bolsa de permanência recebida durante a graduação	Nenhum, auxílio moradia, auxílio alimentação, auxílio moradia e alimentação, auxílio permanência ou outro
21. Tipo de bolsa de acadêmica recebida durante a graduação	Nenhum, iniciação científica, extensão, monitoria/tutoria, PET ou outro
22. Participação de programas e ou atividades curriculares no exterior durante a graduação	Não participou, Programa Ciência sem Fronteiras, intercâmbio financiado pelo governo federal, intercâmbio financiado pelo governo estadual, intercâmbio da instituição ou intercâmbio não institucional
23. Ingresso por políticas de ação afirmativa ou inclusão social	Não; sim, étnico-racial; sim, renda; sim, escola pública ou particular com bolsa de estudo; sim, por dois ou mais critérios anteriores; sim, por sistema diferente dos anteriores.
24. Unidade da Federação onde concluiu o ensino médio	RO, AC, AM, RR, PA, AP, TO, MA, PI, CE, RN, PB, PE, AL, SE, BA, MG, ES, RJ, SP, PR, SC, RS, MS, MT, GO e DF
25. Tipo de escola onde cursou o ensino médio	Todo em escola pública, todo em escola privada, todo no exterior, maior parte em escola pública, maior parte em escola privada, parte no exterior, metade em escola pública e metade em privada.

Variável	Categoria
26. Modalidade de ensino médio cursado	Tradicional, Profissionalizante Técnico, Profissionalizante Magistério, Educação de Jovens e Adultos ou Outro
27. Quem mais incentivou cursar a graduação	Ninguém, pais, outros membros da família que não os pais, professores, líder ou representante religioso, colegas/amigos, outras pessoas.
28. Algum grupo determinante para enfrentar dificuldades durante o curso	Não tive dificuldade, não recebi apoio para enfrentar dificuldades, Pais, Avós; Irmãos, primos ou tios; Líder ou representante religioso, colegas de curso ou amigos, professores do curso, profissionais do serviço de apoio ao estudante da IES, colegas de trabalho, outro grupo.
29. Pelo menos um membro da família com curso superior concluído	Sim ou não
30. Quantidade de livros, diferentes da bibliografia do curso, lidos no ano	Nenhum, um ou dois, três a cinco, seis a oito ou mais de oito
31. Quantidade de horas por semana, excetuando as de aula, dedicadas aos estudos	Apenas assiste às aulas, uma a três, quatro a sete, oito a doze ou mais de doze
32. Oportunidade de aprendizado de idioma estrangeiro na IES	Sim, somente presencial; sim, somente semipresencial; sim, parte presencial e parte semipresencial; sim, a distância ou; não

Variável	Categoria
33. Principal motivo de escolha do curso	Inserção no mercado de trabalho, influência familiar, valorização profissional, prestígio social, vocação, oferecido a distância, baixa concorrência para ingresso, ou outro
34. Principal razão de escolha da IES	Gratuidade, preço da mensalidade, proximidade da residência, proximidade do trabalho, facilidade de acesso, qualidade/reputação, única onde obteve aprovação, possibilidade de obter bolsa de estudo ou outro

Após a redução de dimensionalidade que resultou na Tabela 1, os dados passam pelas etapas de análise estatística listadas a seguir. Destaca-se que este ponto possui grande importância para o treinamento do sistema de aprendizado de máquina uma vez que, dados estatisticamente incorretos geram ruídos difíceis de corrigir na saída.

- Cálculo da média e mediana para validar o processo de regressão linear;
- Teste de normalidade Kolmogorov-Smirnov, que avalia a hipótese nula de que os dados seguem uma distribuição normal;
- Análise de regressão linear múltipla com a aplicação dos testes de p-valor, R-quadrado,
- Análise dos resíduos para verificar se os resíduos são independentes e não correlacionados;
- Teste de normalidade nos resíduos utilizando o modelo de teste de Shapiro-Wilk para avaliar a hipótese nula de que os resíduos do modelo ajustados são normalmente distribuídos;
- Avaliação da presença de multicolinearidade utilizando a técnica estatística que calcula o *Variance Inflation Factor* (VIF) para detectar quais variáveis violam os pressupostos do modelo linear;

- Limpeza das variáveis que apresentam multicolinearidades, ou seja, VIF maior que 5;
- Nova análise dos resíduos para verificar se são independentes e não correlacionados;
- Validação do conjunto final de dados.

A análise de regressão linear termina com uma nova tabela composta de 16 colunas (ou variáveis independentes) e as mesmas 136471 linhas (ou registros). O passo seguinte é selecionar e avaliar os algoritmos que melhor descrevem as análises preditivas para essas variáveis.

Um código em Python é usado para teste da base de dados e para avaliar os principais algoritmos de regressão linear. O primeiro passo antes de inserir a base nas rotinas do Python é a normalização dos dados. A técnica em Python utilizada é a normalização de máximos e mínimos. A normalização é necessária pois existem diferenças entre as escalas das variáveis, por exemplo, a escala da variável idade e a nota, conforme apresentado a seguir:

Variável idade:

- idade mínima: 20;
- idade máxima: 81;
- média das idades: 27.31.

Variável nota:

- nota mínima: 0.0;
- nota máxima: 91.6;
- média das notas: 39.41.

Na sequência, uma porção aleatória dos dados precisa ser separada para treino e teste, e os algoritmos de regressão linear escolhidos para aplicação e treinamento. Inicialmente é utilizado o AutoML, técnica automatizada para estabelecer os parâmetros iniciais dos algoritmos. A escolha inicial conta com a ativação dos 12 algoritmos listados a seguir:

- 'KNeighborsUnif';
- 'KNeighborsDist';
- 'LightGBMXT';

- 'LightGBM';
- 'RandomForestMSE';
- 'CatBoost';
- 'ExtraTreesMSE',
- 'NeuralNetFastAI',
- 'XGBoost',
- 'NeuralNetTorch',
- 'LightGBMLarge',
- 'WeightedEnsemble_L2'.

Como o objetivo é realizar a predição de uma possível nota na prova do ENADE tomando como base as variáveis validadas no processo de regressão linear torna-se necessário estabelecer critérios para selecionar os melhores algoritmos que contribuem com a resposta buscada. Os critérios adotados são:

- A métrica do menor erro apresentado no treino e teste utilizando a técnica de RMSE;
- O menor tempo de treinamento, tomando como base um computador pessoal ou uma conta comum no Colab (GOOGLE CO., 2019);
- O tempo de validação;
- A facilidade de replicabilidade dos dados.

Com base nos critérios apresentados os três algoritmos selecionados são:

- LightGBM
- XGBoost
- CatBoost

Uma vez escolhidos os algoritmos o passo seguinte é estabelecer os ajustes nos parâmetros para realizar o processo de retirada de resultados. Após o ajuste ocorre o processo de mineração dos dados e busca de respostas, nesse ponto, é possível predizer a nota do aluno utilizando como entrada uma ou mais variáveis dentro das 12 que estão listadas no processo de análise de regressão. Por exemplo, é possível prever a nota observando apenas uma entrada (critério de bolsa de estudos, ou cota, ou instrução dos pais etc.) ou realizar a combinação entre elas.

5 RESULTADOS

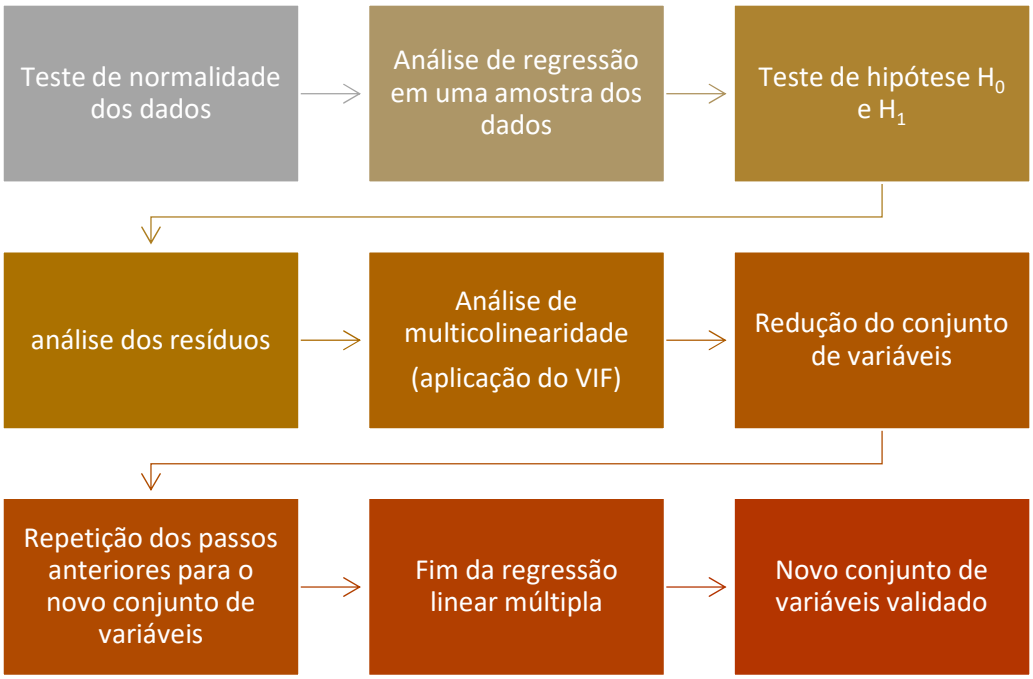
Este capítulo mostra os resultados obtidos no processo executado na pesquisa e está dividido em: análise do processo de regressão para a seleção de variáveis e avalição e seleção dos algoritmos e resultados da análise exploratória.

5.1 Análise de regressão para seleção de variáveis

Para a correta aplicação do processo de regressão linear uma sequência de passos precisa ser seguida. Esses passos envolvem a aplicação de técnicas estatísticas para realizar a redução da dimensionalidade, eliminar a multicolinearidade e garantir o teste de normalidade.

A Figura 5 ilustra a sequência de eventos aplicados nesta etapa.

Figura 5. Diagrama de passos da etapa de análise de regressão linear.

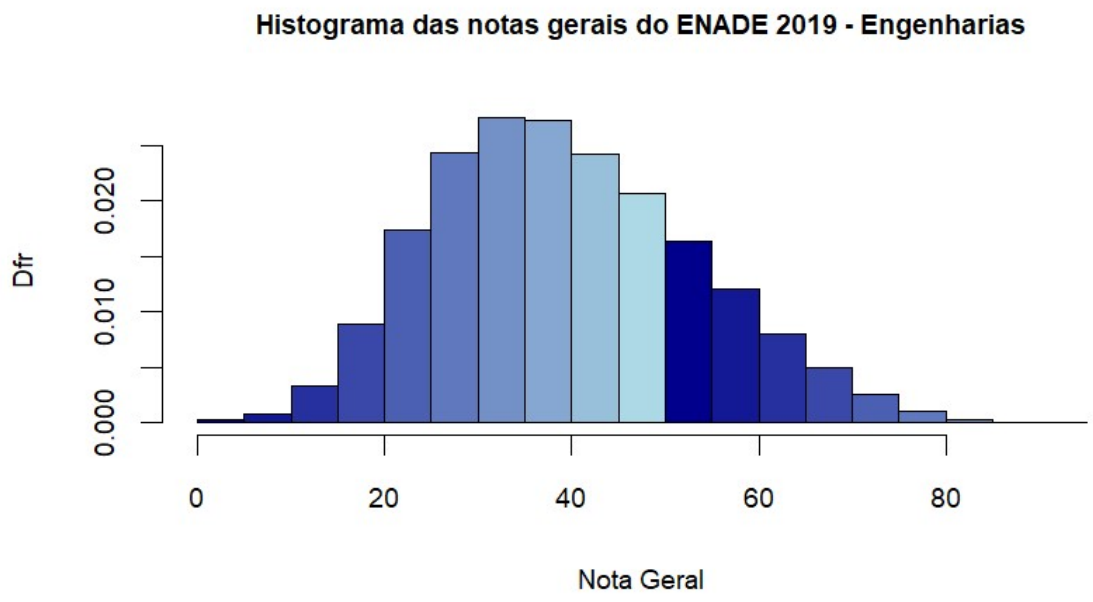


Fonte: Autoria própria.

Ao analisar a base de dados (BRASIL, 2024), o primeiro passo é criar um histograma contendo a variável alvo (nota da prova) para iniciar o processo de validação da normalidade dos dados, mostrado na Figura 6. A nota é composta da seguinte forma:

- Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%)

Figura 6. Histograma da nota geral do ENADE.



Fonte: Autoria própria.

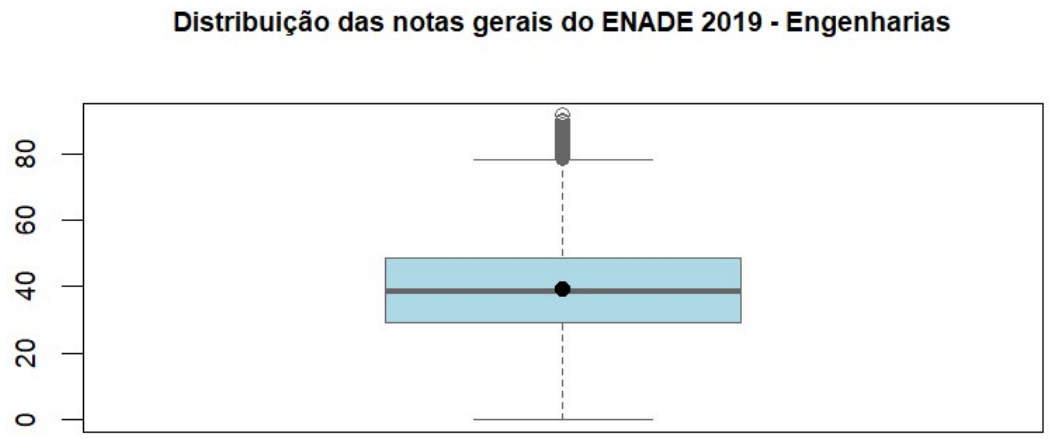
Observando a Figura 6, tem-se as medidas resumo da nota geral, percebe-se que a mediana e a média não coincidem, indicando a presença de assimetria na distribuição desta variável. Os valores calculados para a média, mediana, mínimo, máximo, primeiro quartil e 3 quartil estão mostrados na Tabela 2.

Tabela 2 – Média e mediana da variável nota geral

Min.	1st Qu.	Mediana	Média	3rd Qu.	Max.
0.00	29.10	38.30	39.41	48.80	91.60

No gráfico *boxplot* observado na Figura 7, é possível perceber uma diferença mínima, de aproximadamente 1 ponto, entre a média e a mediana. Embora a diferença apresentada não seja significativa, o passo seguinte é aplicar testes de normalidade para validar a hipótese.

Figura 7 - Boxplot com os valores de média e mediana.



Fonte: Autoria própria.

Utilizando o software R aplica-se o teste de normalidade Kolmogorov-Smirnov (LANDES; MANHÃES, 2018), que avalia a hipótese nula de que os dados seguem uma distribuição normal. O nível de significância escolhido é o p-valor 5%, neste caso, se o teste aplicado na variável apresentar valor menor que 0,05 rejeita-se a hipótese nula em favor da hipótese alternativa de que os dados não seguem distribuição normal. Na simulação em R a nota geral recebe o nome de "y_nota" e o p-valor é um valor próximo de zero, conforme visto na Tabela 3.

Tabela 3 - Teste de Kolmogorov-Smirnov.

Asymptotic one-sample Kolmogorov-Smirnov test
data: y_nota
D = 0.99892, p-value < 2.2e-16 alternative hypothesis: two-sided

Ao analisar os resultados é possível assumir, pelo teorema central do limite, e por se tratar de grande volume de dados, que os dados seguem uma distribuição normal.

O passo seguinte na análise de regressão é extrair uma amostra aleatória de tamanho m = 4.091, o que representa 3% do conjunto de dados.

O modelo adotado para descrever a variável resposta nota bruta da prova (y), é o de regressão linear múltipla.

Dadas as variáveis explicativas (independentes) $X_1, X_2, \dots, X_k$, e a variável resposta (dependente) Y, tomando-se uma amostra de tamanho n, o modelo de regressão linear múltipla é dado por

$$E[x_{1i}, x_{2i}, \dots, x_{ki}] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \ \varepsilon_i, \ i = 1 \dots n \qquad (1)$$

onde $\varepsilon_i \sim iid\ N(0, \sigma^2)$.

Na forma matricial, tem-se:

$$\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}. \qquad (2)$$

Os estimadores são obtidos pelo método de mínimos quadrados, ao se obter o mínimo para:

$$S = \underline{\varepsilon}'\underline{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2, \qquad (3)$$

ou seja,

$$\min S(\beta) = y'y - 2y'X\beta + \beta'X'X\beta \qquad (4)$$

onde a indicação de matriz foi suprimida por simplicidade.

Assim, assumindo que $X'X$ admite inversa, a solução do sistema de equações normais é dada por:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (5)$$

Essa solução representa o vetor de estimadores de mínimos quadrados do vetor de parâmetros de interesse.

O primeiro modelo a ser considerado foi o completo conforme observado a seguir.

5.1.1 Modelo Completo

Inicialmente fez-se um ajuste no modelo de regressão linear múltipla para a variável nota bruta da prova (y), considerando todas as variáveis explicativas (2-34) descritas na Tabela 1, conforme mostrado na Tabela 4.

Tabela 4 - Ajuste de variáveis

y_nota ~ tp_adm + uf + tp_modalidade + gratuito +idade+ tp_sexo +ano_inicio+ q1 + q2 + q3 + q4 + q5 + q6 + q7 + q8 + q9 + q10 + q11 + q12 + q13 + q14 + q15 + q16 + q17 + q18 + q19 + q20 + q21 + q22 + q23 + q24 + q25 + q26

Os coeficientes obtidos no ajuste estão na Tabela 5. A Tabela 5 possui tamanho significativo, por essa razão, optou-se por mostrar parte dos dados e o restante encontra-se no Anexo A.

Tabela 5 - Coeficientes obtidos no ajuste das variáveis.

Coeficientes	Estimado	Std. Error	t value	Pr(>|t|)	Significância
(Intercept)	-901,7392	250,61097	-3,59800	0,000324	***
as.factor(tp_adm)2	0,44148	0,94777	0,46600	0,641383	
as.factor(tp_adm)3	-4,80507	5,19068	-0,92600	0,354654	
as.factor(tp_adm)4	-5,79437	4,86517	-1,19100	0,23373	
as.factor(tp_adm)5	-5,02830	4,83862	-1,03900	0,298776	
as.factor(tp_adm)7	-0,32704	6,68415	-0,04900	0,96098	
as.factor(uf)AL	16,24838	13,77195	1,18000	0,238145	
as.factor(uf)AM	6,64049	13,20280	0,50300	0,61502	
as.factor(uf)AP	-3,15476	14,48488	-0,21800	0,827599	
as.factor(uf)BA	6,65924	12,87740	0,51700	0,605098	
as.factor(uf)CE	10,82379	13,19663	0,82000	0,412156	
as.factor(uf)DF	0,54551	12,99200	0,04200	0,96651	
as.factor(uf)ES	4,93707	13,03906	0,37900	0,704978	
as.factor(uf)GO	5,42026	12,90003	0,42000	0,674382	
as.factor(uf)MA	0,27106	13,09269	0,02100	0,983483	
as.factor(uf)MG	6,37358	12,78099	0,49900	0,618036	
as.factor(uf)MS	5,39463	13,10919	0,41200	0,680717	
as.factor(uf)MT	2,30314	13,07521	0,17600	0,860189	
as.factor(uf)PA	5,83840	13,02875	0,44800	0,654094	
as.factor(uf)PB	6,32315	13,09888	0,48300	0,629319	
as.factor(uf)PE	7,43170	13,01173	0,57100	0,567928	
as.factor(uf)PI	2,14637	13,32610	0,16100	0,87205	
as.factor(uf)PR	4,67767	12,80112	0,36500	0,714824	
as.factor(uf)RJ	4,68504	12,84215	0,36500	0,715268	
as.factor(uf)RN	6,34576	13,17795	0,48200	0,630157	
as.factor(uf)RO	7,34970	13,46303	0,54600	0,585154	
as.factor(uf)RR	-2,56336	16,87223	-0,15200	0,879252	
as.factor(uf)RS	3,29925	12,93324	0,25500	0,79866	
as.factor(uf)SC	4,14374	12,92090	0,32100	0,748454	
as.factor(uf)SE	14,46708	13,20945	1,09500	0,273494	
as.factor(uf)SP	4,97223	12,73387	0,39000	0,696208	
as.factor(uf)TO	8,72607	13,27597	0,65700	0,511038	

Os coeficientes assinalados com asterisco (*) possuem p-valor < 0,05 para o teste que avalia a hipótese de que o coeficiente é nulo, o que indica que ele é estatisticamente diferente de zero.

O teste de hipótese é:

$$H_0: \beta_j = 0 \; vs \; H_1: \beta_j \neq 0 \tag{6}$$

A estatística do teste é:

$T = \frac{\beta_j}{\sqrt{\sigma^2 C_{jj}}}$, onde C_{jj} é o j-ésimo elemento da diagonal principal de $(X'X)^{-1}$. (7)

Rejeita-se H_0 se $|T| > t_{\frac{\alpha}{2}}, n - p, p$ parâmetros.

A Tabela 6 , apresenta os resultados do ajuste do modelo feitos em linguagem R.

Tabela 6 - Resultado dos ajustes do modelo feito em R.

Residual standard error: 11,85 on 3896 degrees of freedom
Multiple R-squared: 0,3041 Adjusted R-squared: 0,2695
F-statistic: 8,777 on 194 and 3896 DF, p-value: < 2,2e-16

O coeficiente de determinação R^2 avalia o ajuste do modelo, medindo a proporção da variabilidade presente nas observações da variável resposta Y que pode ser explicada pelas variáveis X. Como o valor de R^2 aumenta quando são incluídas variáveis no modelo, uma alternativa é o R^2 ajustado, cuja fórmula penaliza a inclusão de variáveis que não sejam significativas ao modelo. Se R^2 e R^2 ajustado são muito diferentes, há variáveis que não estão contribuindo para o ajuste do modelo.

Os valores dos coeficientes de determinação do modelo são:

- R^2 = 0,30 e;
- R^2 ajustado = 0,27.

Esses resultados indicam que o modelo linear ajustado explica aproximadamente 30% da variabilidade dos dados.

Para determinar se existe um relacionamento linear entre a variável resposta Y e o conjunto de variáveis explicativas, $X_1, X_2, \ldots, X_k$, considera-se o teste de significância da regressão, com as hipóteses $H_0: \beta_1 = \beta_2 = \cdots = X\beta_k = 0 \; vs \; H_1: \beta_j \neq 0$, para pelo menos um j.

A estatística do teste é dada por F_{obs}. Rejeita-se H_0 se $F_{obs} > F_{\alpha,k,n-p}$, indicando que significa que pelo menos uma das variáveis independentes, $X_1, X_2, \ldots, X_k$, tem contribuição significativa no modelo.

Como F_{obs} = 8,777 e p-valor > 0,05, rejeita-se a hipótese nula, em favor da alternativa de que pelo menos uma variável tem contribuição significativa no modelo.

A análise dos resíduos tem como objetivo verificar o pressuposto do modelo de regressão linear múltipla, ou seja, verificar se os resíduos $\varepsilon_i \sim iid\ N(0, \sigma^2)$ são independentes e identicamente distribuídos (não correlacionados) e com variância constante.

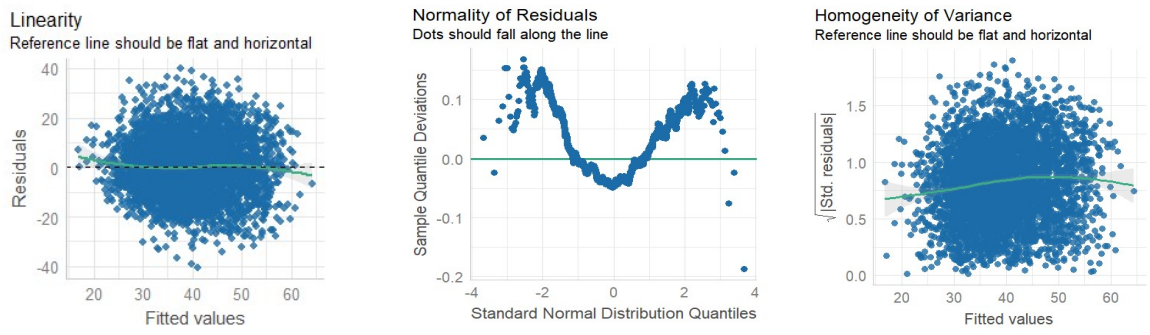
A Tabela 7 apresenta os valores resumo para os resíduos do modelo ajustado, indicando que sua distribuição é um pouco assimétrica à esquerda, comparada a distribuição normal.

Tabela 7 - Resumo de resíduos do modelo ajustado.

Resíduos				
Min	**1Q**	**Mediana**	**3Q**	**Max**
-40,735	-8,001	-0,490	7,727	39,885

Os gráficos dos resíduos estão representados nas Figura 8.

Figura 8. Gráfico de linearidade do resíduo



Fonte: Autoria própria

Ao observar a Figura 8, que mostra os gráficos dos resíduos, percebe-se que o modelo é linear e os resíduos não são lineares, e exatamente por essa razão foram retirados do modelo. No primeiro gráfico da Figura 8, que avalia a linearidade do modelo, para ser linear a linha que acompanha a pontilhada deveria ser coincidente, o que não ocorre. O segundo gráfico, avalia a normalidade residual, e mostra que os pontos não estão distribuídos aleatoriamente em torno de uma reta, indicando que os resíduos não seguem a distribuição normal, o que viola o pressuposto do modelo linear. Por fim, o gráfico das variáveis ajustadas

pelos resíduos padronizados avalia a homogeneidade da variância, ou seja, se a variância é constante. Observa-se que este pressuposto não é atendido.

O teste de normalidade dos resíduos avalia a hipótese nula de que os resíduos do modelo ajustado são normalmente distribuídos. Observa-se na Tabela 8 que a estatística do teste de Shapiro-Wilk (W) apresenta p-valor < 0,05, indicando que a hipótese nula é rejeitada em favor da hipótese alternativa, de que os resíduos não são normais.

Tabela 8 - Teste de normalidade dos resíduos.

Shapiro-Wilk normality test
data: lm01$residuals
W = 0.9979, p-value = 2.394e-05

Quando as estimativas para os coeficientes de regressão apresentam sinais algébricos opostos ao que seriam esperados, há um indício de presença de multicolinearidade. Neste caso, as estimativas dos coeficientes dos parâmetros são insignificantes, visto que, se duas variáveis são fortemente correlacionadas, é muito difícil haver variação em uma sem que haja em outra.

O passo seguinte é detectar a presença de multicolinearidade entre as variáveis. O método adotado é o fator inflacionário da variância (VIF) que avalia a presença de multicolinearidade entre as variáveis explicativas, ou seja, avalia se as variáveis explicativas apresentam correlação entre si. Seja R a matriz dos resíduos do modelo ajustado. O VIF é definido como $VIF_j =$ j-ésimo elemento diagonal de R^{-1} , j = 1, ..., k.

A Tabela 9 apresenta os valores de VIF calculados para o modelo. Os coeficientes da Tabela 8 representam as variações de cada variável e o campo detecção indica a presença ou não da multicolinearidade sempre que o VIF é maior que 5.

Tabela 9 - Valores de VIF calculados.

Coeficientes	VIF	Detecção	Variável
as.factor(tp_adm)2	1,3587	0	tp_adm
as.factor(tp_adm)3	8,7844	1	tp_adm
as.factor(tp_adm)4	164,1004	1	tp_adm
as.factor(tp_adm)5	146,9418	1	tp_adm
as.factor(tp_adm)7	2,2338	0	tp_adm
as.factor(uf)AL	36,4970	1	uf
as.factor(uf)AM	77,5969	1	uf

Coeficientes	VIF	Detecção	Variável
as.factor(uf)AP	15,0134	1	uf
as.factor(uf)BA	202,6829	1	uf
as.factor(uf)CE	120,6939	1	uf
as.factor(uf)DF	60,9889	1	uf
as.factor(uf)ES	83,9720	1	uf
as.factor(uf)GO	112,9174	1	uf
as.factor(uf)MA	88,1894	1	uf
as.factor(uf)MG	678,4025	1	uf
as.factor(uf)MS	40,3548	1	uf
as.factor(uf)MT	64,1503	1	uf
as.factor(uf)PA	77,8675	1	uf
as.factor(uf)PB	83,5142	1	uf
as.factor(uf)PE	161,9207	1	uf
as.factor(uf)PI	59,1886	1	uf
as.factor(uf)PR	348,9402	1	uf
as.factor(uf)RJ	415,7918	1	uf
as.factor(uf)RN	72,4685	1	uf
as.factor(uf)RO	22,0250	1	uf
as.factor(uf)RR	6,1094	1	uf
as.factor(uf)RS	247,7628	1	uf
as.factor(uf)SC	220,3137	1	uf
as.factor(uf)SE	44,6734	1	uf
as.factor(uf)SP	906,6538	1	uf
as.factor(uf)TO	37,6564	1	uf
as.factor(tp_modalidade)1	1,5074	0	tp_modalidade
as.factor(gratuito)1	129,6926	1	gratuito
idade	2,3071	0	idade
as.factor(tp_sexo)M	1,1524	0	tp_sexo
as.factor(ano_inicio)2002	3,2337	0	ano_inicio
as.factor(ano_inicio)2004	3,1981	0	ano_inicio
as.factor(ano_inicio)2005	4,3707	0	ano_inicio
as.factor(ano_inicio)2006	8,5578	1	ano_inicio
as.factor(ano_inicio)2007	6,4792	1	ano_inicio
as.factor(ano_inicio)2008	10,8425	1	ano_inicio
as.factor(ano_inicio)2009	12,9230	1	ano_inicio
as.factor(ano_inicio)2010	60,1177	1	ano_inicio
as.factor(ano_inicio)2011	109,2611	1	ano_inicio
as.factor(ano_inicio)2012	178,4526	1	ano_inicio
as.factor(ano_inicio)2013	417,2827	1	ano_inicio
as.factor(ano_inicio)2014	782,9815	1	ano_inicio
as.factor(ano_inicio)2015	1.100,1308	1	ano_inicio
as.factor(ano_inicio)2016	264,9333	1	ano_inicio
as.factor(ano_inicio)2017	151,3653	1	ano_inicio
as.factor(ano_inicio)2018	111,2819	1	ano_inicio
as.factor(ano_inicio)2019	49,5542	1	ano_inicio
as.factor(q1)B	3,5925	0	q1
as.factor(q1)C	1,1828	0	q1

Coeficientes	VIF	Detecção	Variável
as.factor(q1)D	1,0418	0	q1
as.factor(q1)E	1,4084	0	q1
as.factor(q2)B	1,2365	0	q2
as.factor(q2)C	1,0877	0	q2
as.factor(q2)D	1,4008	0	q2
as.factor(q2)E	1,1180	0	q2
as.factor(q2)F	1,0981	0	q2
as.factor(q3)B	1,0501	0	q3
as.factor(q3)C	1,4767	0	q3
as.factor(q4)B	7,2718	1	q4
as.factor(q4)C	6,5364	1	q4
as.factor(q4)D	10,6257	1	q4
as.factor(q4)E	7,8397	1	q4
as.factor(q4)F	4,5848	0	q4
as.factor(q5)B	9,8816	1	q5
as.factor(q5)C	9,0761	1	q5
as.factor(q5)D	17,5254	1	q5
as.factor(q5)E	14,0869	1	q5
as.factor(q5)F	9,7725	1	q5
as.factor(q6)B	9,3855	1	q6
as.factor(q6)C	8,6277	1	q6
as.factor(q6)D	2,2432	0	q6
as.factor(q6)E	1,1906	0	q6
as.factor(q6)F	1,1787	0	q6
as.factor(q7)B	4,2321	0	q7
as.factor(q7)C	6,1295	1	q7
as.factor(q7)D	6,7477	1	q7
as.factor(q7)E	4,4641	0	q7
as.factor(q7)F	2,4222	0	q7
as.factor(q7)G	1,5498	0	q7
as.factor(q7)H	1,3689	0	q7
as.factor(q8)B	2,4164	0	q8
as.factor(q8)C	2,5873	0	q8
as.factor(q8)D	2,2851	0	q8
as.factor(q8)E	2,5982	0	q8
as.factor(q8)F	2,3486	0	q8
as.factor(q8)G	1,4586	0	q8
as.factor(q9)B	4,4014	0	q9
as.factor(q9)C	5,7758	1	q9
as.factor(q9)D	3,6761	0	q9
as.factor(q9)E	3,9082	0	q9
as.factor(q9)F	3,5811	0	q9
as.factor(q10)B	1,2103	0	q10
as.factor(q10)C	1,2686	0	q10
as.factor(q10)D	1,6379	0	q10
as.factor(q10)E	2,8790	0	q10
as.factor(q11)B	18,1377	1	q11

Coeficientes	VIF	Detecção	Variável
as.factor(q11)C	9,8192	1	q11
as.factor(q11)D	3,0761	0	q11
as.factor(q11)E	18,4211	1	q11
as.factor(q11)F	2,9729	0	q11
as.factor(q11)G	2,2245	0	q11
as.factor(q11)H	6,7393	1	q11
as.factor(q11)I	2,9671	0	q11
as.factor(q11)J	2,1846	0	q11
as.factor(q11)K	1,5562	0	q11
as.factor(q12)B	1,1258	0	q12
as.factor(q12)C	1,1724	0	q12
as.factor(q12)D	1,3156	0	q12
as.factor(q12)E	1,2300	0	q12
as.factor(q12)F	1,1110	0	q12
as.factor(q13)B	1,3816	0	q13
as.factor(q13)C	1,1578	0	q13
as.factor(q13)D	1,1900	0	q13
as.factor(q13)E	1,0895	0	q13
as.factor(q13)F	1,1195	0	q13
as.factor(q14)B	1,1375	0	q14
as.factor(q14)C	1,0678	0	q14
as.factor(q14)D	1,0433	0	q14
as.factor(q14)E	1,0983	0	q14
as.factor(q14)F	1,0729	0	q14
as.factor(q15)B	1,1359	0	q15
as.factor(q15)C	1,1918	0	q15
as.factor(q15)D	1,3671	0	q15
as.factor(q15)E	1,3659	0	q15
as.factor(q15)F	1,0853	0	q15
as.factor(q16)12	10,1801	1	q16
as.factor(q16)13	11,9565	1	q16
as.factor(q16)14	3,0819	0	q16
as.factor(q16)15	11,8093	1	q16
as.factor(q16)16	4,6093	0	q16
as.factor(q16)17	5,8853	1	q16
as.factor(q16)21	12,9863	1	q16
as.factor(q16)22	9,9610	1	q16
as.factor(q16)23	18,5321	1	q16
as.factor(q16)24	9,4341	1	q16
as.factor(q16)25	11,0343	1	q16
as.factor(q16)26	21,0555	1	q16
as.factor(q16)27	8,0594	1	q16
as.factor(q16)28	6,5886	1	q16
as.factor(q16)29	25,6197	1	q16
as.factor(q16)31	63,6758	1	q16
as.factor(q16)32	12,3174	1	q16
as.factor(q16)33	39,9098	1	q16

Coeficientes	VIF	Detecção	Variável
as.factor(q16)35	82,5020	1	q16
as.factor(q16)41	26,3155	1	q16
as.factor(q16)42	21,4188	1	q16
as.factor(q16)43	27,3868	1	q16
as.factor(q16)50	5,0136	1	q16
as.factor(q16)51	8,2463	1	q16
as.factor(q16)52	13,7817	1	q16
as.factor(q16)53	7,1454	1	q16
as.factor(q16)99	3,3376	0	q16
as.factor(q17)B	1,8905	0	q17
as.factor(q17)C	1,6584	0	q17
as.factor(q17)D	1,1343	0	q17
as.factor(q17)E	1,1417	0	q17
as.factor(q17)F	1,2040	0	q17
as.factor(q18)B	1,1469	0	q18
as.factor(q18)C	1,0532	0	q18
as.factor(q18)D	1,1217	0	q18
as.factor(q18)E	1,1251	0	q18
as.factor(q19)B	2,2405	0	q19
as.factor(q19)C	1,6076	0	q19
as.factor(q19)D	1,2541	0	q19
as.factor(q19)E	1,2518	0	q19
as.factor(q19)F	1,2962	0	q19
as.factor(q19)G	1,2536	0	q19
as.factor(q20)B	1,3223	0	q20
as.factor(q20)C	1,8884	0	q20
as.factor(q20)D	1,1407	0	q20
as.factor(q20)E	1,1547	0	q20
as.factor(q20)F	1,2560	0	q20
as.factor(q20)G	1,4757	0	q20
as.factor(q20)H	1,2583	0	q20
as.factor(q20)I	1,1346	0	q20
as.factor(q20)J	1,1421	0	q20
as.factor(q20)K	1,3932	0	q20
as.factor(q21)B	1,4384	0	q21
as.factor(q22)B	2,0285	0	q22
as.factor(q22)C	1,9278	0	q22
as.factor(q22)D	1,3217	0	q22
as.factor(q22)E	1,4007	0	q22
as.factor(q23)B	6,1073	1	q23
as.factor(q23)C	5,8354	1	q23
as.factor(q23)D	3,8170	0	q23
as.factor(q23)E	3,5955	0	q23
as.factor(q24)B	1,1271	0	q24
as.factor(q24)C	1,2645	0	q24
as.factor(q24)D	1,3717	0	q24
as.factor(q24)E	1,6246	0	q24

Coeficientes	VIF	Detecção	Variável
as.factor(q25)B	1,2410	0	q25
as.factor(q25)C	1,4350	0	q25
as.factor(q25)D	1,1070	0	q25
as.factor(q25)E	1,5073	0	q25
as.factor(q25)F	1,1648	0	q25
as.factor(q25)G	1,0883	0	q25
as.factor(q25)H	1,3732	0	q25
as.factor(q26)B	1,8806	0	q26
as.factor(q26)C	2,7998	0	q26
as.factor(q26)D	1,3466	0	q26
as.factor(q26)E	1,7059	0	q26
as.factor(q26)F	3,4393	0	q26
as.factor(q26)G	1,2538	0	q26
as.factor(q26)H	2,1665	0	q26
as.factor(q26)I	2,0843	0	q26

O fato de haver presença de multicolinearidade explica a violação dos pressupostos do modelo linear. Para corrigir a multicolinearidade, foram eliminadas as variáveis com VIF > 5, que são aquelas altamente correlacionadas a outras variáveis.

As variáveis eliminadas são apresentadas na Tabela 10 a seguir.

Tabela 10 - Tabela com variáveis eliminadas depois da análise do VIF.

Variável	Descrição
tp_adm	Tipo da categoria administrativa da IES
uf	Nome do Estado à qual pertence a IES
gratuito	Gratuidade do curso
ano_inicio	Ano de início da graduação
q4	Escolarização do pai
q5	Escolarização da mãe
q6	Onde e com quem mora
q7	Número de pessoas que moram na mesma casa com o aluno
q9	Situação financeira (incluindo bolsas)
q11	Tipo de bolsa de estudos ou financiamento recebido para custear mensalidades
q16	Unidade da Federação onde concluiu o ensino médio
q23	Quantidade de horas por semana, excetuando as de aula, dedicadas aos estudos

Depois da exclusão das variáveis com VIF > 5 ajusta-se um modelo de regressão linear múltipla para a variável nota bruta da prova (y), considerando todas as variáveis explicativas com VIF < 5, conforme mostrado na Tabela 11.

Tabela 11 - Modelo ajustado de regressão com eliminação das variáveis com VIF > 5.

y_nota ~ tp_modalidade + idade+ tp_sexo + q1 + q2 + q3 + q8 + q10 + q12 + q13 + q14 + q15 + q17 + q18 + q19 + q20 + q21 + q22 + q24 + q25 + q26

Os coeficientes obtidos no ajuste estão na Tabela 12.

Tabela 12 - Coeficientes obtidos pós ajuste.

Coefficientes	**Estimativa**	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif.**
(Intercept)	44,90240	2,20344	20,378	2,00E-16	***
idade	-0,22239	0,04479	-4,965	7,15E-07	***
as.factor(q1)B	0,67092	0,66574	1,008	0,313623	
as.factor(q1)C	0,64168	1,64384	0,390	0,696294	
as.factor(q1)D	2,04731	7,14710	0,286	0,774546	
as.factor(q1)E	-0,16454	1,24298	-0,132	0,894694	
as.factor(q10)B	-1,24883	0,85557	-1,460	0,144468	
as.factor(q10)C	-0,97888	0,96586	-1,013	0,310894	
as.factor(q10)D	-0,47752	0,64873	-0,736	0,461718	
as.factor(q10)E	-2,83029	0,49924	-5,669	1,54E-08	***
as.factor(q12)B	0,74803	2,37907	0,314	0,753217	
as.factor(q12)C	4,05135	1,31895	3,072	0,002143	**
as.factor(q12)D	5,38738	1,52866	3,524	0,000429	***
as.factor(q12)E	3,27805	1,40558	2,332	0,019741	*
as.factor(q12)F	2,69966	1,62009	1,666	0,095719	.
as.factor(q13)B	5,48586	0,69310	7,915	3,18E-15	***
as.factor(q13)C	2,52176	1,34625	1,873	0,061116	.
as.factor(q13)D	5,17869	0,91382	5,667	1,55E-08	***
as.factor(q13)E	6,02054	2,55003	2,361	0,018275	*
as.factor(q13)F	-0,63785	0,83370	-0,765	0,444263	
as.factor(q14)B	1,26924	1,58494	0,801	0,423287	
as.factor(q14)C	2,22791	2,72200	0,818	0,41313	
as.factor(q14)D	-3,84642	8,71846	-0,441	0,659106	
as.factor(q14)E	0,72112	1,63414	0,441	0,659033	
as.factor(q14)F	-1,69541	1,39697	-1,214	0,22496	
as.factor(q15)B	0,48181	1,66044	0,290	0,771701	
as.factor(q15)C	1,81434	0,82773	2,192	0,028441	*
as.factor(q15)D	3,49390	0,68974	5,066	4,26E-07	***
as.factor(q15)E	5,43274	0,88373	6,147	8,64E-10	***
as.factor(q15)F	-1,35084	1,73792	-0,777	0,437042	
as.factor(q17)B	3,59472	0,51840	6,934	4,74E-12	***
as.factor(q17)C	8,50004	8,53945	0,995	0,319609	
as.factor(q17)D	0,02853	1,02626	0,028	0,977827	
as.factor(q17)E	0,98813	0,98673	1,001	0,316684	
as.factor(q17)F	9,13195	3,12959	2,918	0,003543	**
as.factor(q18)B	1,97069	0,55403	3,557	0,000379	***
as.factor(q18)C	-1,25139	3,45418	-0,362	0,71716	

Coefficientes	Estimativa	Std. Error	t value	Pr(>|t|)	Signif.
as.factor(q18)D	-0,62292	1,28500	-0,485	0,62787	
as.factor(q18)E	-5,12126	2,58399	-1,982	0,047557	*
as.factor(q19)B	-1,14788	0,61689	-1,861	0,062851	.
as.factor(q19)C	-2,05269	0,95546	-2,148	0,031743	*
as.factor(q19)D	3,90472	1,38964	2,810	0,00498	**
as.factor(q19)E	-5,95703	5,11796	-1,164	0,244516	
as.factor(q19)F	-1,58226	1,21841	-1,299	0,194147	
as.factor(q19)G	-0,26307	1,34454	-0,196	0,844887	
as.factor(q2)B	-2,68725	0,79964	-3,361	0,000785	***
as.factor(q2)C	-1,28121	1,17992	-1,086	0,277612	
as.factor(q2)D	-2,03807	0,44609	-4,569	5,05E-06	***
as.factor(q2)E	-5,99878	3,64572	-1,645	0,099959	.
as.factor(q2)F	-0,56266	1,41806	-0,397	0,691549	
as.factor(q20)B	-0,71016	0,92149	-0,771	0,440952	
as.factor(q20)C	0,09500	0,51765	0,184	0,854397	
as.factor(q20)D	0,32304	1,61641	0,200	0,841606	
as.factor(q20)E	-2,04384	1,43198	-1,427	0,153577	
as.factor(q20)F	-1,98320	2,83216	-0,700	0,483816	
as.factor(q20)G	2,01742	0,71881	2,807	0,005031	**
as.factor(q20)H	-0,68137	1,04098	-0,655	0,512799	
as.factor(q20)I	-2,38424	3,01284	-0,791	0,428781	
as.factor(q20)J	1,54762	2,04476	0,757	0,449172	
as.factor(q20)K	0,07497	0,95894	0,078	0,937691	
as.factor(q21)B	-1,16755	0,46432	-2,515	0,011958	*
as.factor(q22)B	-1,78792	0,53732	-3,328	0,000884	***
as.factor(q22)C	-0,89644	0,59097	-1,517	0,129374	
as.factor(q22)D	-0,70285	0,94418	-0,744	0,456674	
as.factor(q22)E	-0,50382	0,84466	-0,596	0,550893	
as.factor(q24)B	-2,29795	1,82713	-1,258	0,20858	
as.factor(q24)C	0,63554	0,87679	0,725	0,468587	
as.factor(q24)D	1,44593	0,86240	1,677	0,093694	.
as.factor(q24)E	-0,50061	0,49322	-1,015	0,310176	
as.factor(q25)B	-1,80517	0,81807	-2,207	0,027397	*
as.factor(q25)C	-0,98400	0,56615	-1,738	0,082275	.
as.factor(q25)D	-2,32523	1,65386	-1,406	0,159819	
as.factor(q25)E	1,00609	0,51164	1,966	0,049322	*
as.factor(q25)F	4,76098	3,88672	1,225	0,220672	
as.factor(q25)G	-3,72453	3,23262	-1,152	0,24932	
as.factor(q25)H	-0,70145	0,64903	-1,081	0,279871	
as.factor(q26)B	-3,98134	1,03191	-3,858	0,000116	***
as.factor(q26)C	-1,26171	0,76185	-1,656	0,097778	.
as.factor(q26)D	-4,78536	1,53470	-3,118	0,001833	**
as.factor(q26)E	-5,26501	1,05909	-4,971	6,93E-07	***
as.factor(q26)F	-0,07143	0,65690	-0,109	0,913421	
as.factor(q26)G	-2,73144	1,50159	-1,819	0,068981	.
as.factor(q26)H	-0,98051	0,93713	-1,046	0,295487	
as.factor(q26)I	-2,94830	0,90063	-3,274	0,001071	**

Coefficientes	Estimativa	Std. Error	t value	Pr(>|t|)	Signif.
as.factor(q3)B	1,40384	2,58484	0,543	0,587088	
as.factor(q3)C	2,67800	4,21306	0,636	0,525046	
as.factor(q8)B	1,46895	0,67227	2,185	0,028943	*
as.factor(q8)C	1,98963	0,70140	2,837	0,004582	**
as.factor(q8)D	4,03715	0,79176	5,099	3,57E-07	***
as.factor(q8)E	5,25806	0,79777	6,591	4,94E-11	***
as.factor(q8)F	6,67253	0,91010	7,332	2,74E-13	***
as.factor(q8)G	8,03610	1,62034	4,960	7,36E-07	***
as.factor(tp_modalidade)1	-0,73865	1,20378	-0,614	0,53951	
as.factor(tp_sexo)M	0,89322	0,43888	2,035	0,041896	*

Os coeficientes assinalados com asterisco (*) possuem p-valor < 0,05 para o teste que avalia a hipótese de que o coeficiente é nulo, o que indica que esse coeficiente é estatisticamente diferente de zero.

A Tabela 13 apresenta os resultados do ajuste do modelo.

Tabela 13 - Resultados do ajuste do modelo.

Residual standard error: 12.24 on 3997 degrees of freedom
Multiple R-squared: 0.2383, Adjusted R-squared: 0.2206
F-statistic: 13.45 on 93 and 3997 DF, p-value: < 2.2e-16

Os valores dos coeficientes de determinação do modelo são:

- R^2 = 0,24;
- R^2 ajustado = 0,22.

Isso indica que o modelo linear ajustado explica aproximadamente 24% da variabilidade dos dados.

Como Fobs = 13,45 e p-valor < 0,05, rejeita-se a hipótese nula, em favor da alternativa de que pelo menos uma variável tem contribuição significativa no modelo.

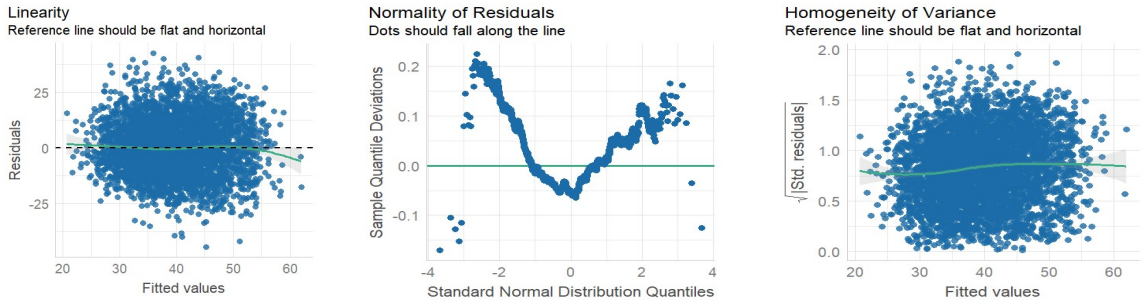
A Tabela 14 apresenta os valores resumo para os resíduos do modelo ajustado, indicando que sua distribuição é um pouco assimétrica à esquerda, comparada a distribuição normal.

Tabela 14 - Valores resumo para o modelo ajustado.

Min	1Q	Mediana	3Q	Max
-44,792	-8,503	-0,633	8,124	42,673

Os gráficos dos resíduos estão na Figura 9, a seguir.

Figura 9. Gráfico das variáveis ajustadas pelos resíduos.



Fonte: Autoria própria.

A Figura 9 mostra o gráfico das variáveis ajustadas pelos resíduos e avalia a linearidade do modelo. Observa-se, na primeira parte da Figura 9, que a linha variável é próxima da pontilhada horizontal. A segunda parte da Figura 9 mostra o gráfico de probabilidade normal indicando que os pontos não estão distribuídos aleatoriamente em torno de uma reta, indicando que os resíduos não seguem a distribuição normal. O gráfico das variáveis ajustadas pelos resíduos padronizados avalia a homogeneidade da variância, ou seja, se a variância é constante. Observa-se a distribuição próxima a linear no eixo y =0

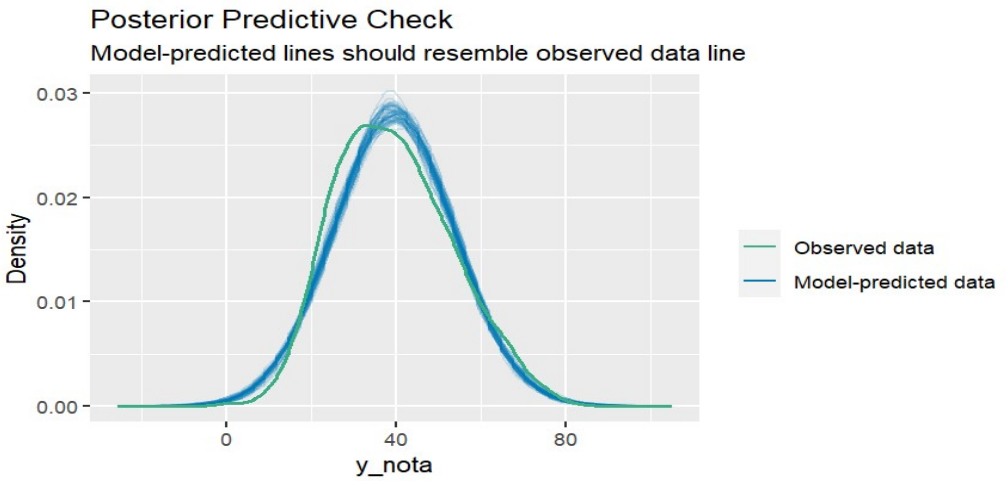
Observa-se na Tabela 15 a seguir que a estatística do teste de Shapiro-Wilk (W) apresenta p-valor < 0,05, indicando que a hipótese nula é rejeitada em favor da hipótese alternativa, de que os resíduos não são normais.

Tabela 15 - Resultado do teste de Shapiro-Wilk.

Shapiro-Wilk normality test
data: lm02$residuals
W = 0.99726, p-value = 9.711e-07

O gráfico mostrado na Figura 10 apresenta as distribuições dos valores da variável resposta predita pelo modelo e a observada.

Figura 10. Gráfico da distribuição normal das variáveis preditas e observadas.



Fonte: autoria própria.

Verifica-se que o modelo reduzido ajustado se aproxima do observado nas caudas, sendo centrado na nota 40, diferentemente do observado que se encontra mais à esquerda.

Embora, o ajuste não tenha sido o mais apropriado, a análise de regressão linear múltipla foi mantida por ser aplicada como uma ferramenta exploratória, com o objetivo de identificar as variáveis mais significativas para seu emprego em *machine learning.*

As variáveis significativas do modelo reduzido a serem empregadas encontram-se na Tabela 16. A coluna variável da Tabela 16 indica as variáveis significativas que não apresentam multicolinearidade e que podem ser utilizadas no modelo de regressão linear.

Tabela 16 - Variáveis significativas do modelo reduzido.

Variável	Descrição da Variável
idade	Idade
tp_sexo	Sexo
q2	Cor ou raça
q8	Renda total da família, incluindo o aluno
q10	Situação de trabalho (exceto estágio ou bolsas)
q12	Tipo de bolsa de permanência recebida durante a graduação
q13	Tipo de bolsa de acadêmica recebida durante a graduação
q15	Ingresso na graduação meio de políticas de ação afirmativa ou inclusão social

q17	Tipo de escola onde cursou o ensino médio
q18	Modalidade de ensino médio cursado
q19	Quem mais incentivou cursar a graduação
q20	Algum grupo determinante para enfrentar dificuldades durante o curso
q21	Pelo menos um membro da família com curso superior concluído
q22	Quantidade de livros, diferentes da bibliografia do curso, lidos no ano
q25	Principal motivo de escolha do curso
q26	Principal razão de escolha da instituição de educação superior

Com base nos resultados apresentados, as principais sugestões de replicabilidade desse modelo estão listadas a seguir:

- Exclusão de variáveis com VIF > 5: Como identificado, algumas variáveis apresentam alta multicolinearidade com outras. Sugere-se remover essas variáveis do modelo para evitar problemas de multicolinearidade.
- Avaliação dos coeficientes: Após ajustar o modelo de regressão, é importante avaliar a significância estatística dos coeficientes. Aqueles com p-valor maior que 0,05 podem não ser estatisticamente significativos e podem ser considerados para remoção do modelo.
- Análise dos resíduos: Os resíduos do modelo devem ser analisados para verificar se atendem aos pressupostos do modelo de regressão linear. Se os resíduos não seguem uma distribuição normal ou se há padrões nos resíduos, pode ser necessário investigar mais a fundo.
- Interpretação dos coeficientes significativos: Para os coeficientes que são estatisticamente significativos, é importante interpretar o impacto prático dessas variáveis na resposta.
- Revisão do modelo: Com base nos resultados da análise dos resíduos e da significância dos coeficientes, pode ser necessário revisar o modelo, considerando adicionar novas variáveis, transformar variáveis existentes ou remover variáveis irrelevantes.

5.2 Análise dos algoritmos

Diversas opções de algoritmos existiam para a tarefa preditiva a ser utilizada na base de dados que foi tratada. Depois da análise exploratória, a melhor opção encontrada foram os algoritmos de regressão e classificação.

As principais métricas de análise para a escolha dos algoritmos foram o erro médio quadrado (R2), raiz do erro quadrático médio (RMSE), o tempo de treinamento, tempo de validação e a facilidade de replicação dos resultados, essas são as métricas que fornecem a avaliação mais robusta dos modelos que utilizam regressão linear múltipla. Embora existam outros fatores, tais como, *recall, F1-score* e a área sob a curva ROC eles não fornecem as melhores métricas para a análise e treinamento de algoritmos de regressão linear múltipla.

A variável alvo do modelo era a NOTA FINAL do ENADE. Conforme mostrado no item 5.1, os dados de entrada recebem o devido tratamento estatístico de regressão linear múltipla.

Em seguida, utilizando o Python e a biblioteca Autogluon foram testados os 12 algoritmos para avaliar a capacidade de realizar predições das notas com base nas variáveis de entrada. A função Python que estrutura a função preditora e o respectivo teste pode ser acessada pelo link do GitHub a seguir:

- Link https://github.com/brunoguilhen/Dissertacao_Mestrado.git

A lista dos algoritmos testados está descrita a seguir:

- 'KNeighborsUnif',
- 'KNeighborsDist',
- 'LightGBMXT',
- 'LightGBM',
- 'RandomForestMSE',
- 'CatBoost',
- 'ExtraTreesMSE',
- 'NeuralNetFastAI',
- 'XGBoost',
- 'NeuralNetTorch',
- 'LightGBMLarge',

- 'WeightedEnsemble_L2'

O resultado dos testes com os parâmetros de cada um está listado na Figura 11 a seguir:

Figura 11. Resultado da avaliação dos algoritmos.

	model	score_val	eval_metric	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	WeightedEnsemble_L2	-0.130702	root_mean_squared_error	0.154295	372.276434	0.000897	0.562222	2	True	12
1	LightGBMXT	-0.130814	root_mean_squared_error	0.052535	6.243080	0.052535	6.243080	1	True	3
2	XGBoost	-0.130898	root_mean_squared_error	0.031791	6.162736	0.031791	6.162736	1	True	9
3	LightGBM	-0.130911	root_mean_squared_error	0.048150	4.666954	0.048150	4.666954	1	True	4
4	LightGBMLarge	-0.131040	root_mean_squared_error	0.066877	6.742016	0.066877	6.742016	1	True	11
5	CatBoost	-0.131313	root_mean_squared_error	0.024044	49.985348	0.024044	49.985348	1	True	6
6	NeuralNetTorch	-0.131411	root_mean_squared_error	0.021088	160.900724	0.021088	160.900724	1	True	10
7	NeuralNetFastAI	-0.131612	root_mean_squared_error	0.047983	198.407672	0.047983	198.407672	1	True	8
8	ExtraTreesMSE	-0.134207	root_mean_squared_error	0.483575	848.855968	0.483575	848.855968	1	True	7
9	RandomForestMSE	-0.134763	root_mean_squared_error	0.476312	896.149978	0.476312	896.149978	1	True	5
10	KNeighborsUnif	-0.158543	root_mean_squared_error	0.492260	0.091172	0.492260	0.091172	1	True	1
11	KNeighborsDist	-0.158544	root_mean_squared_error	0.511240	0.084980	0.511240	0.084980	1	True	2

Fonte: autoria própria

Ao observar os resultados do teste e avaliar os critérios de escolha, os algoritmos selecionados são:

- LightGBM;
- XGBoost;
- CatBoost;

Nos algoritmos de regressão linear existem diversos critérios de seleção, tais como, a linearidade dos dados, a presença de multicolinearidade, as métricas de erro, a escalabilidade do algoritmo, o tempo de treinamento etc (FÁVERO; BELFIORE, 2017). Conforme observado na Figura 5, os caminhos para os testes de normalidade, linearidade e multicolinearidade foram devidamente testados na base de dados. Assim, para refinar a escolha dos algoritmos e, em seguida, ajustar os parâmetros, optou-se por realizar uma análise das medidas dos erros e do tempo de treinamento apresentados por um conjunto de algoritmos de regressão, selecionando os que apresentaram as melhores métricas, conforme explicado a seguir. O critério de erro para a escolha dos algoritmos está listado a seguir:

- Erro médio usando a métrica RMSE (raiz do erro quadrático médio);
- Tempo de treinamento (coluna fit_time) sendo 100 vezes mais rápidos;
- Tempo da validação;

- Facilidade da replicação dos resultados.

Um outro critério de seleção foi a quantidade de ferramentas utilizadas pela comunidade acadêmica (maior popularidade de uso).

A analisar os resultados da Figura 11 e os critérios de escolha dos algoritmos tem-se, na Tabela 17, a lista dos algoritmos classificados por posição que mostram o menor erro médio quadrado (RMSE), que é o primeiro critério utilizado na avaliação.

Tabela 17 - Tabela com a classificação dos algoritmos

Posição	Modelo	RMSE	Tempo de treinamento (fit_time)
1	NeuralNetTorch	2,1%	160,90000
2	CatBoost	2,4%	49,95000
3	XGBoost	3,7%	6,16000
4	NeuralNetFastAI	4,7%	198,40000
5	LightGBM	4,8%	4,66600
6	LightGBMXT	5,3%	6,24300
7	LightGBMLarge	6,7%	6,74200
8	WeightedEnsemble_L2	15,4%	372,27000
9	RandomForestMSE	47,0%	896,14000
10	ExtraTreesMSE	48,0%	848,85000
11	KNeighborsUnif	49,0%	0,09110
12	KNeighborsDist	51,1%	0,08490

Os cinco primeiros algoritmos da Tabela 17 possuem RMSE menores que 5% (primeiro critério de seleção). Porém, ao observar o tempo de treinamento verifica-se que o tempo de

treinamento do *NeuralNetTorch* é muito superior aos outros, nesse caso, a opção por descartar esse algoritmo reside no fato de que o tempo para treinar em computadores comuns com bases de dados grandes não é interessante. Neste caso, a primeira opção de escolha está no XGBoost que possui RMSE de 3,7%.

Os algoritmos nas posições 5, 6 e 7 são do tipo *Light Gradiente-Boosting Machine* (LGBM) com algumas variações entre eles, por essa razão, a outra opção de escolha é o LightGBM que possui RMSE 4,8%.

Para gerar mais relevância na contribuição, optou-se por selecionar o algoritmo CATBoost que possui RMSE 2,4%, porém, um tempo de treinamento razoável. A justificativa para o uso é a variação que esse algoritmo permite trazer para a pesquisa.

Entre o LightGBM e o XGBoost, o LightGBM pode ser considerado ligeiramente melhor com base nos resultados apresentados. Isso porque ambos os algoritmos demonstraram desempenho bastante semelhante em termos de métricas como RMSE e R-quadrado. No entanto, o LightGBM teve um tempo de treinamento um pouco mais rápido em comparação com o XGBoost, o que pode ser vantajoso, especialmente para conjuntos de dados maiores. Portanto, considerando uma precisão comparável e um tempo de treinamento um pouco mais rápido, o LightGBM se destaca como uma escolha ligeiramente superior.

Depois da escolha dos algoritmos, cada um deles passa por ajustes nos hiperparâmetros e inicia-se a fase de treinamento e teste.

5.3 Treinamento e teste dos algoritmos escolhidos

No código em Python, atribui-se o nome de “y_nota” para a variável de saída (preditora), considerando que o algoritmo vai receber a base de dados e ter como saída a predição da nota do aluno, os algoritmos recebem o treinamento e a avaliação segue as seguintes métricas:

- R-quadrado (R^2);
- Erro médio quadrado (RMSE);
- Erro médio absoluto (MAE);
- Erro mediano absoluto.

5.3.1 Resultados para LightGBM

Depois do ajuste de parâmetros, o algoritmo passa pela etapa de teste que consiste em rodar a base de treino para predizer a nota. A variável resposta é y_nota. A Figura12 mostra o resultado das métricas.

Figura 12. Resultado do treino e teste para o LGBM.

Treino

```
[ ] print("r2: ", r2_score(X_enade_base_treino['y_nota'], predict_lightGBM_treino))
    print("RMSE: ", mean_squared_error(X_enade_base_treino['y_nota'], predict_lightGBM_treino))
    print("Median Absolute Error: ", median_absolute_error(X_enade_base_treino['y_nota'], predict_lightGBM_treino))
    print("Mean Absolute Error: ", mean_absolute_error(X_enade_base_treino['y_nota'], predict_lightGBM_treino))

    r2:  0.25375469729731115
    RMSE:  0.017110127444253928
    Median Absolute Error:  0.08956976540744568
    Mean Absolute Error:  0.1048173590856212
```

Teste

```
[ ] print("r2: ", r2_score(X_enade_base_teste['y_nota'], predict_lightGBM))
    print("RMSE: ", mean_squared_error(X_enade_base_teste['y_nota'], predict_lightGBM))
    print("Median Absolute Error: ", median_absolute_error(X_enade_base_teste['y_nota'], predict_lightGBM))
    print("Mean Absolute Error: ", mean_absolute_error(X_enade_base_teste['y_nota'], predict_lightGBM))

    r2:  0.2346315266072072
    RMSE:  0.017518285526641104
    Median Absolute Error:  0.0904455752351919
    Mean Absolute Error:  0.10669588490621441
```

Fonte: Autoria própria.

Ao observar a Figura 12, percebe-se que o R-quadrado no treino e no teste estão próximos de 25% e 23%. Como o R-quadrado explica a variabilidade dos dados, isso indica que aproximadamente um quarto da dispersão dos dados (25%) em torno da média da variável dependente pode ser explicada pelas variáveis independentes. O valor de R-quadrado varia de 0 até 1, onde zero não explica nada e 1 indica que o modelo é capaz de explicar toda a variação nos dados. Quando a base de dados é pequena e com poucas variáveis esse valor tende a estar mais perto da unidade. Porém, ao aplicar na base de dados da pesquisa, o valor de 25% é considerado aceitável, por essa razão é que as métricas de RMSE, MAE e erro absoluto da mediana foram inseridas.

Ao observar o resultado dos outros 3 fatores percebe-se que eles não passam de 10%, sendo que o RMSE, ou seja, o erro quadrático médio, medida de precisão do modelo, está na casa de 1%. Esse valor indica que o erro entre as previsões e os valores reais são muito baixos, bem como, que o R-quadrado é aceitável e o algoritmo possui excelente nível de precisão preditiva.

Outro fator a ser observado é que o resultado do teste indica qual a capacidade de aprendizado do algoritmo depois do treino, ao observar os valores é possível perceber que o

algoritmo manteve os padrões do treino, neste caso, conclui-se que ganhou capacidade preditiva.

5.3.2 Resultados para o XGBoost

As mesmas observações feitas para o LGBM podem ser reproduzidas no XGBoost, a Figura 13 ilustra os resultados de treino e teste.

Figura 13. Resultado de treino e teste para o XGBoost.

Treino

```
[ ]  print("r2: ", r2_score(X_enade_base_treino['y_nota'], predict_XGBoost_treino))
     print("RMSE: ", mean_squared_error(X_enade_base_treino['y_nota'], predict_XGBoost_treino))
     print("Median Absolute Error: ", median_absolute_error(X_enade_base_treino['y_nota'], predict_XGBoost_treino))
     print("Mean Absolute Error: ", mean_absolute_error(X_enade_base_treino['y_nota'], predict_XGBoost_treino))

     r2:  0.2662793778612571
     RMSE:  0.016822957957395444
     Median Absolute Error:  0.08882886041199795
     Mean Absolute Error:  0.10389757827591448
```

Teste

```
[ ]  print("r2: ", r2_score(X_enade_base_teste['y_nota'], predict_XGBoost))
     print("RMSE: ", mean_squared_error(X_enade_base_teste['y_nota'], predict_XGBoost))
     print("Median Absolute Error: ", median_absolute_error(X_enade_base_teste['y_nota'], predict_XGBoost))
     print("Mean Absolute Error: ", mean_absolute_error(X_enade_base_teste['y_nota'], predict_XGBoost))

     r2:  0.2334149044620022
     RMSE:  0.017546132419815653
     Median Absolute Error:  0.0902471479890648
     Mean Absolute Error:  0.10653694055094665
```

Fonte: autoria própria

A diferença a ser observada dos resultados no LGBM para o XGBoost é que no primeiro algoritmo o resultado de treino é de 25%, no segundo, o resultado é de 26%. É uma ligeira melhora, para o tamanho da base de dados e a quantidade de variáveis essa diferença de 1% representa melhora. Porém, no teste os resultados estão na casa dos 23%, mantendo-se equivalentes os outros elementos e a precisão preditiva medida pelo RMSE.

5.3.3 Resultados para o CatBoost

O CatBoost apresenta valores similares os apresentados pelo XGBoost. O R-quadrado e a precisão observada pelo RMSE são idênticas mantendo-se as mesmas observações feitas para os dois algoritmos anteriores.

A Figura 14 ilustra os resultados de treino e teste que estão na casa dos 26% e 23% para o R-quadrado e de 1% para o RMSE.

Figura 14. Resultado de treino e teste para o CatBoost.

Treino

```
[ ] print("r2: ", r2_score(X_enade_base_treino['y_nota'], predict_CatBoost_treino))
    print("RMSE: ", mean_squared_error(X_enade_base_treino['y_nota'], predict_CatBoost_treino))
    print("Median Absolute Error: ", median_absolute_error(X_enade_base_treino['y_nota'], predict_CatBoost_treino))
    print("Mean Absolute Error: ", mean_absolute_error(X_enade_base_treino['y_nota'], predict_CatBoost_treino))

    r2:  0.2614336237838898
    RMSE:  0.016934062803921093
    Median Absolute Error:  0.0890072610831156
    Mean Absolute Error:  0.10425099214973045
```

Teste

```
[ ] print("r2: ", r2_score(X_enade_base_teste['y_nota'], predict_CatBoost))
    print("RMSE: ", mean_squared_error(X_enade_base_teste['y_nota'], predict_CatBoost))
    print("Median Absolute Error: ", median_absolute_error(X_enade_base_teste['y_nota'], predict_CatBoost))
    print("Mean Absolute Error: ", mean_absolute_error(X_enade_base_teste['y_nota'], predict_CatBoost))

    r2:  0.23607973165540774
    RMSE:  0.017485138003039778
    Median Absolute Error:  0.09081147123111905
    Mean Absolute Error:  0.10645925340445853
```

Fonte: autoria própria

Conforme visto na Figura 11 e na Tabela 9, no processo de seleção dos algoritmos, o CatBoost apresentou um tempo de treinamento maior que os outros escolhidos. Porém, a sua escolha se deve, além do baixo valor de RMSE, ou seja, alta precisão no aprendizado e na capacidade preditiva, pelo fato de ser um algoritmo com alta capacidade de trabalhar com dados categóricos sem exigir pré-processamento extensivo, sendo uma opção para bases de dados grandes que não recebem tratamento adequado.

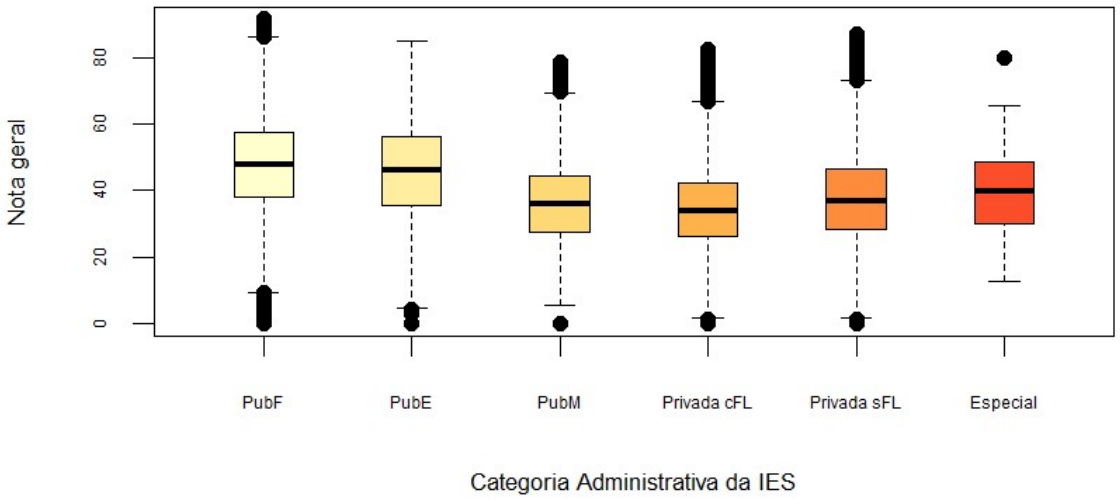
Convém notar que os 3 algoritmos selecionados e analisados são algoritmos de aprendizado de máquina baseados em gradiente de *boosting*, ou seja, um método onde múltiplos modelos de aprendizado de máquina são combinados para criar um modelo mais poderoso, por essa razão, é que os resultados de treino e teste são muito parecidos.

5.4 Análise das predições

Depois das etapas de treino e teste dos algoritmos o passo final é realizar a análise exploratória dos dados relacionando as entradas com as saídas. Neste ponto, seleciona-se apenas os resultados mais relevantes e que pode contribuir com respostas para as instituições de ensino.

A primeira análise, mostrada da Figura 15, lista a nota dos alunos por categoria de Instituição de Ensino Superior (IES).

Figura 15. Gráfico das notas por categoria de Instituição de Ensino.

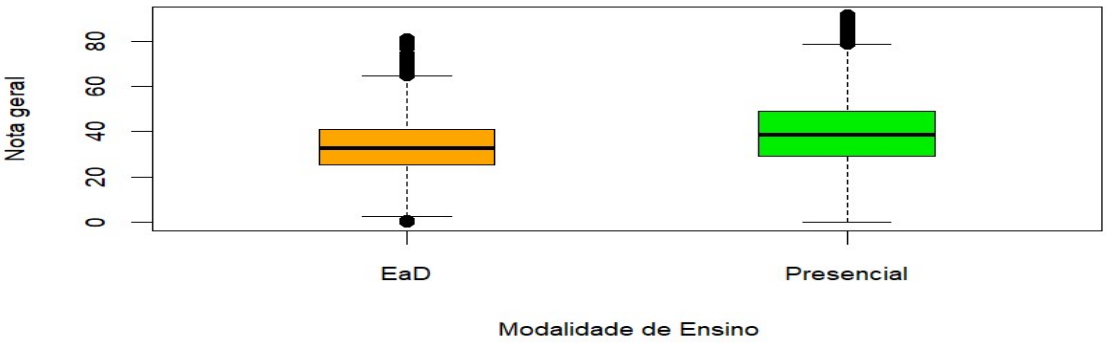


Fonte: autoria própria.

Observa-se no gráfico da Figura 15 que as universidades classificadas como públicas federais e públicas estaduais possuem desempenho melhor que a pública municipal, as privadas com e sem fins lucrativos e as especiais. Além disso, comparando com a média geral de notas, que é o valor numérico 40, as duas citadas estão acima da média. Um ponto interessante a ser observado e que dificilmente os algoritmos conseguem prever é a quantidade de *outliers* apresentados. Tomando como exemplo os alunos das públicas federais, percebe-se um conjunto de *outliers* na parte superior da média, indicando que tiraram notas muito acima do padrão e muitas notas abaixo da média, fato que pode ser entendido como abandono ou entrega da prova em branco. Ao calcular a média para todas as instituições o valor é aproximadamente 40.

Na Figura 16 observa-se a análise pela modalidade do curso, dividido em presencial e remoto.

Figura 16. Notas por modalidade do curso.

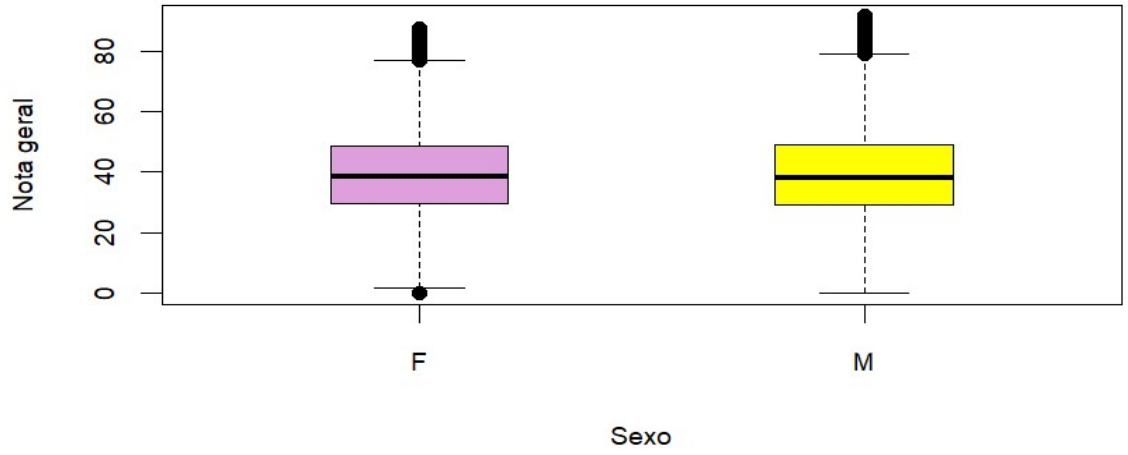


Fonte: autoria própria.

A Figura 16 mostra uma ligeira vantagem do ensino presencial relacionado à nota final. Percebe-se que os alunos do curso presencial possuem notas que estão em torno da média e os alunos do ensino remoto notas ligeiramente abaixo da média geral. Nesse caso, entende-se que, embora o ensino presencial seja o mais tradicional aplicado nos cursos de engenharia, o ensino remoto não ficou com médias muito inferiores.

A Figura 17 ilustra a variável nota geral separada por sexo conforme visto a seguir.

Figura 17. Nota geral por sexo.

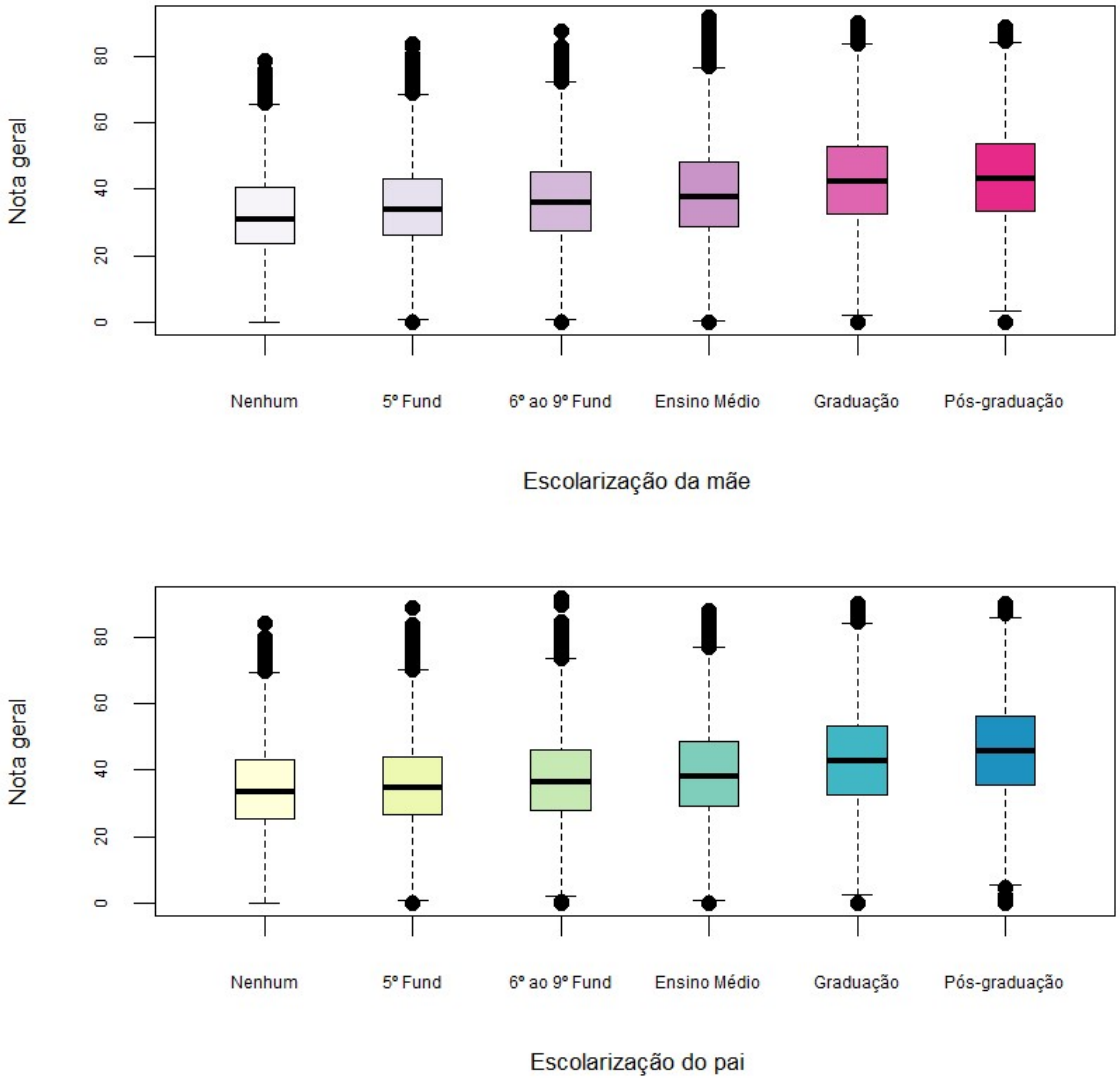


Fonte: autoria própria

A Figura 17 mostra que não há significativa diferença na nota final entre os sexos. Nota-se que o sexo masculino, por estar em maior número nos cursos de engenharia também possui maior quantidade de elementos de *outliers* acima da média. Porém, na média geral, as notas são semelhantes.

A Figura 18 mostra o gráfico de escolarização do pai e da mãe. Essa variável possui forte correlação linear com a nota final e quando inserida dentro da regressão exerce muita influência no ambiente de treino, por essa razão é que houve a exclusão na análise da multicolinearidade. No trabalho de Bordieri (2021) é feita uma abordagem sobre a influência da escolaridade dos pais no sucesso do aluno na graduação, o que corrobora com o que foi encontrado na análise de regressão.

Figura 18. Gráfico da nota final baseado na escolaridade dos pais.

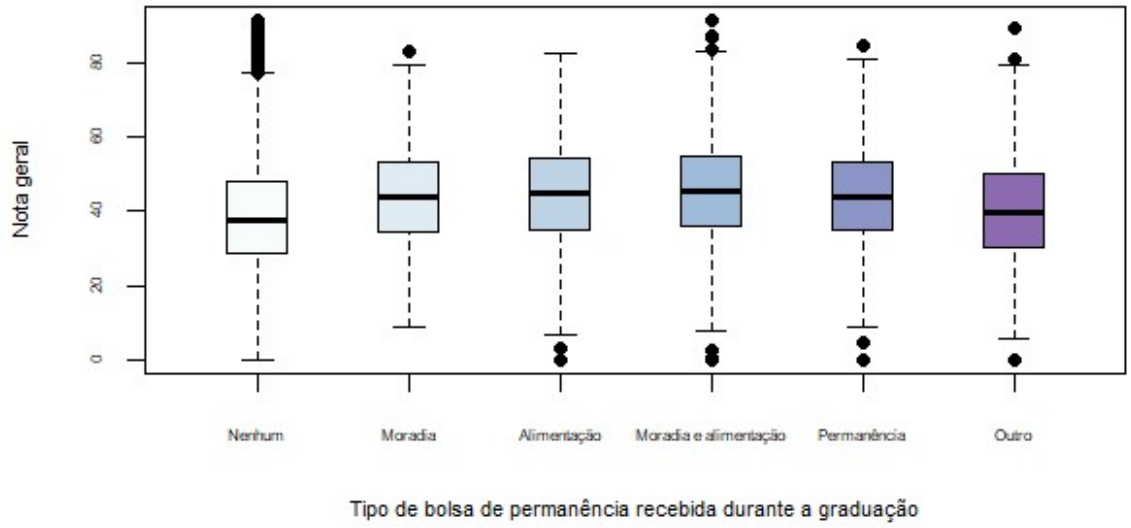


Fonte: autoria própria.

Conforme já mencionado, a Figura 18 mostra que tanto o nível de escolaridade do pai quanto o da mãe exercem influência na nota final do aluno, quer seja na nota do ENADE como mostrado nessa pesquisa, quer seja no desempenho da graduação como mostra Bordieri (2021), percebe-se que conforme o nível de instrução melhora, o resultado da nota final acompanha este crescimento.

Outro ponto relevante observado nessa pesquisa diz respeito ao recebimento de alguns tipos de bolsa durante o curso e a política de cotas para acesso à universidade. A Figura 19 mostra que as notas finais dos alunos são ligeiramente melhores para aqueles que receberam bolsa de permanência de moradia e alimentação, isso indica que é uma política a ser mantida caso a universidade queira melhores resultados no ENADE.

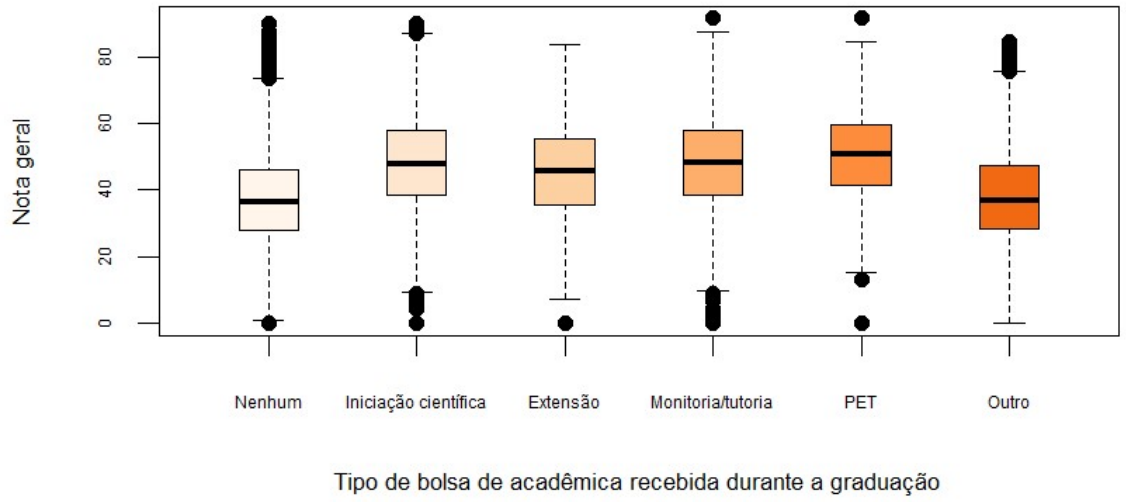
Figura 19. Gráfico de nota por tipo de bolsa permanência.



Fonte: autoria própria

Na mesma linha mostrada na Figura 19, a Figura 20 mostra a nota final dos alunos que receberam bolsas acadêmicas conforme visto a seguir.

Figura 20. Nota dos alunos que receberam bolsa acadêmica.

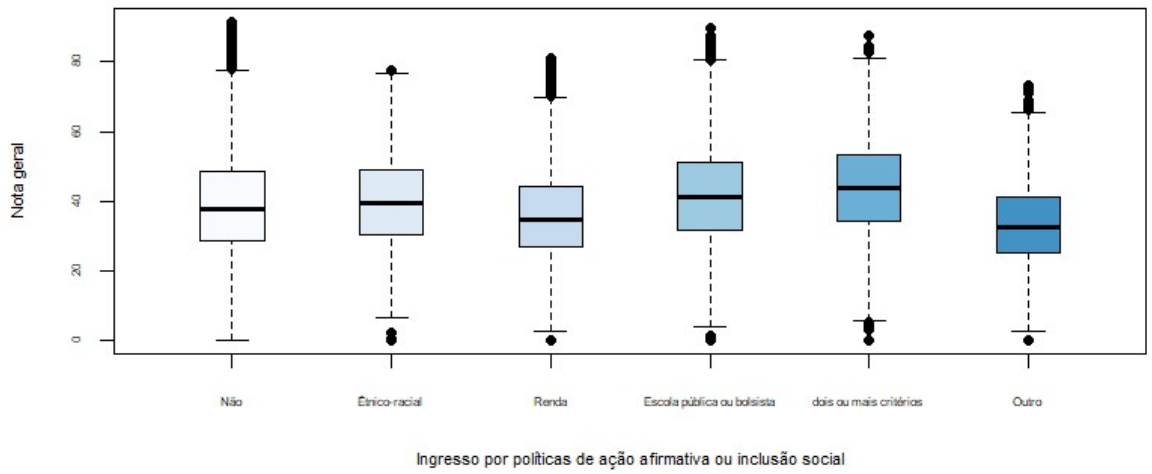


Fonte: autoria própria.

Depreende-se da Figura 20 que bolsas acadêmicas que incentivam a pesquisa, a extensão, monitoria e o Programa Especial de Treinamento (PET) influenciam positivamente a nota final do aluno. Nota-se que os alunos que não receberam bolsa nenhuma apresentam média inferior a 40 que é a média geral calculada e os alunos que receberam o incentivo apresentam nota superior a 40.

Por fim, a análise da nota final sendo descrita pela política de cotas (ações afirmativas ou inclusão social) está registrada na Figura 21 conforme visto a seguir.

Figura 21. Notas finais baseadas nas ações afirmativas.



Fonte: autoria própria

Um ponto relevante que pode ser percebido ao descrever a nota final tomando como base as ações afirmativas é que as notas dos alunos que entraram na graduação sem nenhuma ação afirmativa e a nota dos alunos que possuem uma ou mais ações afirmativas são muito semelhantes. Isso indica que as ações afirmativas contribuem muito no processo de entrada, uma vez que, o ponto de corte dos alunos que entram na universidade sem ações afirmativas é maior que os alunos com ações afirmativas, porém, ao final do curso, na nota do ENADE percebe-se um reflexo positivo. Entende-se, dessa análise, que essa política é muito positiva do ponto de vista do desempenho.

O que se pode destacar das análises de predições está descrito a seguir:

- Desempenho por categoria de Instituição de Ensino Superior (IES): A análise revela que as universidades públicas federais e estaduais tendem a ter um desempenho melhor em comparação com as instituições municipais e privadas, com ou sem fins lucrativos. Isso sugere a importância do financiamento público e do modelo de ensino nessas instituições.
- Modalidade do curso: Embora haja uma ligeira vantagem para os alunos de cursos presenciais em termos de nota final, os alunos de cursos remotos não ficam muito atrás. Isso indica que o ensino remoto pode ser uma alternativa viável, especialmente em situações em que o acesso ao ensino presencial é limitado.

- Diferenças de gênero: A análise não revela diferenças significativas nas notas finais entre os sexos, sugerindo que ambos têm desempenho semelhante nos cursos de engenharia.
- Influência da escolaridade dos pais: Fica evidente que o nível de escolaridade dos pais exerce uma forte influência nas notas finais dos alunos, destacando a importância do apoio familiar e do ambiente educacional.
- Impacto das bolsas: Os resultados mostram que alunos que receberam bolsas de permanência e bolsas acadêmicas apresentaram notas finais mais altas, indicando o impacto positivo desses programas de incentivo no desempenho dos alunos.
- Política de cotas: A análise revela que as ações afirmativas têm um impacto positivo no desempenho dos alunos, mostrando que essas políticas são eficazes em promover a inclusão e melhorar os resultados acadêmicos.

O estudo analisou diversas outras métricas, por não representar diferenças significativas nos resultados optou-se por deixar as análises no link do GitHub https://github.com/brunoguilhen/Dissertacao_Mestrado.git.

6 CONCLUSÕES

A pesquisa realiza uma análise da base de dados do ENADE para os alunos da engenharia, tratando os dados do questionário socioeconômico como variáveis independentes e relacionando-os com a variável dependente chamada de nota geral.

Um dos objetivos é comparar algoritmos que realizam predições que levam o aluno a uma determinada nota e, em seguida, realizar uma análise exploratória para observar como os fatores de entrada se relacionam com a nota.

O resultado da análise de regressão linear permite reduzir a dimensionalidade de 194 colunas para 12 colunas utilizando um tratamento bastante ajustado e profundo dos dados. Assim, com o tratamento correto a etapa seguinte recebe as variáveis já preparadas para serem utilizadas na comparação dos algoritmos baseados em gradiente de *boosting*. Dos 12 algoritmos testados apenas 5 estavam com as métricas de erro abaixo do esperado, além de possuir bons tempos de treinamento.

Os algoritmos também passam por alguns ajustes nos parâmetros para otimizar, melhorar e refinar a capacidade preditiva, mesmo sendo parte do processo de implementação, o ajuste de hiperparâmetros é um processo trabalhoso que merece cuidado e rigor científico, uma vez que, são esses ajustes que melhoram o desempenho, acurácia no treinamento.

A pesquisa segue observando os dados através da análise exploratória com a criação de diversos gráficos. Para cada variável inserida existe uma resposta e um gráfico de análise do comportamento, por questão de comodidade, alguns desses gráficos estão no link do GitHub (https://github.com/brunoguilhen/Dissertacao_Mestrado.git

Da análise final infere-se algumas ações importantes que merecem destaque. Por exemplo, a diferença que a política de cotas cria na vida do acadêmico, uma vez que, quem entrou por cota possui nota do ENADE igual ou levemente superior que os não cotistas.

Outro ponto que as Instituições de Ensino Superior (IES) devem observar é a política de bolsas para as ações afirmativas e a política de bolsas acadêmicas. Os gráficos das Figuras 19 e 20 ilustram que os bolsistas normalmente apresentam notas ligeiramente superior à média.

Um ponto que foi debatido no item 5 refere-se ao grau de escolarização do pai e da mãe, nos gráficos mostrados na Figura 18 percebe-se que quanto maior o grau de escolarização do pai e da mãe melhor a nota final do aluno no exame. Mas a análise vai além,

uma vez que, essa variável possui grande interferência no treinamento do algoritmo, ou seja, possui multicolinearidade, fazendo com que os dados sejam enviesados, por essa razão essas variáveis não foram inseridas no modelo, sendo necessário um tratamento separado utilizando somente esses dados, o que está fora do escopo do trabalho.

A Figura 16 mostra a nota final separada por modalidade de curso, presencial e EAD, o ensino presencial possui maior média de notas. No *boxplot* da Figura 16 é possível perceber que além da média ser ligeiramente maior, o ponto de máximo e quantidade de *outliers* na parte superior, ou seja, bem acima da média, também é maior no ensino presencial.

A Figura 17 mostra a média geral por sexo, a nota do sexo masculino é ligeiramente superior, porém não é uma diferença significativa em termos de média, nesse caso, é importante notar que os homens representam 14014 alunos e as mulheres 6888, o número de pessoas do sexo masculino é o dobro e a média das notas não foi muito diferente.

Uma limitação encontrada para o trabalho é a possibilidade de replicar os testes para um novo ENADE. Nos anos anteriores não houve uma coleta de dados socioeconômicos como ocorreu no ano de 2019, que representa a base de dados da pesquisa, e ainda não ocorreu um novo exame. Isso destaca a importância de uma coleta de dados consistente e padronizada para permitir análises comparativas ao longo do tempo.

No ano de 2025, se o questionário socioeconômico se mantiver com as mesmas variáveis talvez seja possível realizar a comparação.

Uma sugestão de melhoria para futuras pesquisas será explorar técnicas de tratamento de multicolinearidade para incluir variáveis como o nível de escolaridade dos pais no modelo de análise preditiva. Isso poderia proporcionar uma compreensão mais abrangente dos fatores que influenciam o desempenho dos alunos no ENADE.

Outra consideração importante sobre trabalhos futuros é a realização de estudos longitudinais que avaliam a eficácia de intervenções feitas pelas instituições, por exemplo, na política de cotas, de bolsas e outras que foram sugeridas no trabalho e comparar com a nota final dos próximos exames.

Uma das principais vantagens deste estudo é a aplicação de técnicas avançadas de aprendizado de máquina na análise dos dados do ENADE, proporcionando insights valiosos sobre os fatores que influenciam o desempenho dos alunos. Além disso, a abordagem detalhada na análise exploratória dos dados e na comparação de algoritmos permite uma compreensão mais aprofundada do problema em questão.

Esse trabalho não esgota o estudo da aplicação de aprendizado de máquina na base do ENADE. Porém, entende-se que a uma linha para a continuação da pesquisa é criar um sistema de análise preditiva que recebe as variáveis de entrada, aplica um dos algoritmos já treinados e emite corretamente respostas estatisticamente validadas e com boa precisão e acurácia.

REFERÊNCIAS BIBLIOGRÁFICAS

ALDERA, S. et al. Exploratory data analysis and classification of a new arabic online extremism dataset. **IEEE Access**, v. 9, p. 161613–161626, 2021.

ALYAHYAN, E.; DÜŞTEGÖR, D. **Predicting academic success in higher education: literature review and best practices**. **International Journal of Educational Technology in Higher Education**, v. 17, n. 11, Dec. 2020. Aricle nº 3. DOI: 10.1186/s41239-020-0177-7.

AMIN, S. et al. Developing a personalized e-Learning and MOOC recommender System in IoT-enabled smart education. **IEEE Access**, v. 11, p. 136437–136455, 2023.

ANURADHA, C.; VELMURUGAN, T. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. **Indian Journal of Science and Technology**, v. 8, n. 15, 2015. Article nº IPL057.

BATCH, A.; ELMQVIST, N. The Interactive Visualization Gap in Initial Exploratory Data Analysis. **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 1, p. 278–287, 1 jan. 2018.

BORDIERI, E. D. Prevendo desempenho no ENADE: uma aplicação de algoritmos de aprendizado de máquina. **DIVERSITÀ – Revista Multidisciplinar do Centro Universitário Cidade Verde,** p. 59–67, 2021.

BRASIL. **Instituto Nacional de Estudos E Pesquisas Educacionais Anísio Teixeira (INEP). 2019**. Disponível em: <https://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2019/nota_tecnica_n20-2019_CGCQES-DAES_calculo_NF_Enade.pdf>. Acesso em: 13 fev. 2023.

BRASIL. **Instituto Nacional de Estudos E Pesquisas Educacionais Anísio Teixeira (INEP)**. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade>. Acesso em: 13 fev. 2023.

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®**. 1. ed. São Paulo: [s.n.]. v. 1

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **Regularization Paths for Generalized Linear Models via Coordinate DescentJSS Journal of Statistical Software**. [s.l: s.n.]. Disponível em: <http://www.jstatsoft.org/>.

GOOGLE CO. **Google Colab**. Disponível em: <https://research.google.com/colaboratory/intl/pt-BR/faq.html#:~:text=O%20Colab%20permite%20que%20qualquer,an%C3%A1lise%20de%20dados%20e%20educa%C3%A7%C3%A3o.>. Acesso em: 4 jan. 2024.

HASSAN, Y. M. I.; ELKORANY, A.; WASSIF, K. Utilizing social clustering-based regression model for predicting student's GPA. **IEEE Access**, v. 10, p. 48948–48963, 2022.

HONG, Y. et al. Kolmogorov–Smirnov type testing for structural breaks: A new adjusted-range based self-normalization approach. **Journal of Econometrics**, v. 238, n. 2, 1 jan. 2024.

HUANG, C. et al. Examining the relationship between peer feedback classified by deep learning and online learning burnout. **Computers and Education**, v. 207, 1 dez. 2023.

IGHALO, J. O.; ADENIYI, A. G.; MARQUES, G. Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. **Biofuels, Bioproducts and Biorefining**, v. 14, n. 6, p. 1286–1295, 1 nov. 2020.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, v. 31, n. 3, p. 685-695, Sept. 2021.

KUMAR, G. S.; DHANALAKSHMI, R. Performance analysis of Catboost algorithm and XGboost algorithm for prediction of CO 2 emission rating. In: INTERNATIONAL CONFERENCE ON CONTEMPORARY COMPUTING AND INFORMATICS, 6th. 14-16 Sept. 2023, Gautam Buddha Nagar, India. **IC3I**. [Piscataway]: IEEE, 2023. p. 1-5. DOI: 10.1109/IC3I59117.2023.10398160. Disponível em: https://ieeexplore.ieee.org/document/10398160/

LANDES, F. B.; MANHÃES, L. M. B. Análise dos exames do ENADE para os cursos de Computação utilizando o software R. **Revista do Seminário Internacional de Estatística com R**, v. 3, n. 1, p. 1–14, 2018. Apresentado ao 3. Seminário Internacional de Estatística com R, 2018, maio 2018.

LANZANTE, J. R. Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper's tests. **International Journal of Climatology**, v. 41, n. 14, p. 6314–6323, 30 nov. 2021.

LI, T.; ZOU, X. **Informatization of Constructive English Learning Platform Based on Artificial Intelligence Algorithm**. Proceedings - 2022 International Conference on Frontiers of Artificial Intelligence and Machine Learning, FAIML 2022. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2022.

LIMA, P. DA S. N. et al. Análise de dados do Enade e Enem: uma revisão sistemática da literatura. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 24, n. 1, p. 89–107, maio 2019.

MAPHOSA, M.; DOORSAMY, W.; PAUL, B. S. Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis. **IEEE Access**, v. 11, p. 48977–48987, 2023.

MAULUD, D.; ABDULAZEEZ, A. M. A Review on Linear Regression Comprehensive in Machine Learning. **Journal of Applied Science and Technology Trends**, v. 1, n. 4, p. 140–147, 31 dez. 2020a.

MAULUD, D.; ABDULAZEEZ, A. M. A Review on Linear Regression Comprehensive in Machine Learning. **Journal of Applied Science and Technology Trends**, v. 1, n. 4, p. 140–147, 31 dez. 2020b.

MORAIS, M. M. DE O. **O Uso de Aprendizado de Máquina para Identificar Alunos em Risco de Evasão na Educação a Distância**. Rio Grande: Universidade Federal do Rio Grande, 2020.

MORETTIN, P. A.; SINGER, J. DA M. **Estatística e Ciência de Dados**. 1. ed. São Paulo: [s.n.]. v. 1

NAIM, A. et al. **Effective e-Learning practices by machine learning and artificial intelligence**. 2023 International Conference on Artificial Intelligence and Smart Communication, AISC 2023. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2023.

NINAUS, M. et al. Increased emotional engagement in game-based learning – A machine learning approach on facial emotion detection data. **Computers and Education**, v. 142, 1 dez. 2019.

REINSEL, D.; GANTZ, J.; RYDNING, J. **The digitization of the world from Edge to core**. [s.l: s.n.].

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 10, n. 3, 1 maio 2020.

SALEH, H.; LAYOUS, J. A.; REPUBLIC, S. A. Machine learning-regression. **Higher Institute for Applied Sciences and Technology**, v. 1, p. 1–25, 22 jan. 2022.

SARKER, I. H. **Machine Learning: Algorithms, Real-World Applications and Research Directions**. **SN Computer Science**Springer, , 1 maio 2021.

SHI, G. **Technical analysis of machine learning algorithms in artificial intelligence image recognition**. 2023 3rd Asian Conference on Innovation in Technology, ASIANCON 2023. **Anais**. Institute of Electrical and Electronics Engineers Inc., 2023.

SILVA-LUGO, J. L.; WARNER, L. A.; GALINDO, S. **From parametric to non-parametric statistics in education and agricultural education research**. **Journal of Agricultural Education and Extension.** Routledge, 2022.

SOUZA, V. F. DE; CAZELLA, S. Mineração de dados educacionais com algoritmos de regressão: um estudo sobre a predição do desempenho. **Revista Educar Mais**, v. 6, p. 183–198, 18 fev. 2022.

SUNDAY, K. et al. Analyzing student performance in programming education using classification techniques. **International Journal of Emerging Technologies in Learning**, v. 15, n. 2, p. 127–144, 2020.

SYAH, L. Y.; NAFSIAH, S. N.; SADDHONO, K. **Linear regression statistic from accounting information system application for Employee integrity**. Journal of Physics: Conference Series. **Anais**. Institute of Physics Publishing, 16 dez. 2019.

SYAH, L. Y.; NAFSIAH, S. N.; SADDHONO, K. Retraction: application of linear regression mathematical model in the evaluation of teachers informatization quality. **Journal of Physics: Conference Series**, v. 1339, 4 out. 2023.

TRAJANO, F. M. V. **Análise preditiva e exploratória de dados para auxiliar no combate à evasão estudantil nos cursos superiores do IFPB**. Cajazeiras: [s.n.].

TRAWINSKI, B. et al. **Comparison of expert algorithms with machine learning models for real estate appraisal**. 2017 IEEE International Conference on INnovations in Intelligent SysTems

and Applications (INISTA). **Anais**. Gdynia, Poland: IEEE, 3 jul. 2017. Disponível em: <http://ieeexplore.ieee.org/document/8001131/>

TRINCHERO, R.; CANAVERO, F. Machine Learning Regression Techniques for the Modeling of Complex Systems: An Overview. **IEEE Electromagnetic Compatibility Magazine**, v. 10, n. 4, p. 71–79, abr. 2021.

WITTEN, I. H. et al. **Data Mining**. [s.l.] Elsevier, 2017. v. 1

YE, C. **Evolution and Application of Artificial Intelligence Art Design Based on Machine Learning Algorithm**. 2021 IEEE 4th International Conference on Information Systems and Computer Aided Education, ICISCAE 2021. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2021.

ZAHEDI, L. et al. **OptABC: An Optimal Hyperparameter Tuning Approach for Machine Learning Algorithms**. Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2021.

ZHAI, D. et al. Effective Evaluation of Green and High-Quality Development Capabilities of Enterprises Using Machine Learning Combined with Genetic Algorithm Optimization. **Systems**, v. 10, n. 5, 1 out. 2022.

ZHOU, H.; JIANG, S.; LIU, X. Regression analysis of intelligent education based on linear mixed effect model. **Journal of Ambient Intelligence and Humanized Computing**, 2021.

ANEXO A – TABELA DOS COEFICIENTES DE AJUSTES DAS VARIÁVEIS

A tabela a seguir representa a Tabela 5 presente no texto e mostra, de forma completa, todos os ajustes das variáveis. Os valores assinalados com (*) representam os valores com p-valor < 0,05.

Coeficientes	Estimado	Std. Error	t value	Pr(>|t|)	Significância
(Intercept)	-901,7392	250,61097	-3,59800	0,000324	***
as.factor(tp_adm)2	0,44148	0,94777	0,46600	0,641383	
as.factor(tp_adm)3	-4,80507	5,19068	-0,92600	0,354654	
as.factor(tp_adm)4	-5,79437	4,86517	-1,19100	0,23373	
as.factor(tp_adm)5	-5,02830	4,83862	-1,03900	0,298776	
as.factor(tp_adm)7	-0,32704	6,68415	-0,04900	0,96098	
as.factor(uf)AL	16,24838	13,77195	1,18000	0,238145	
as.factor(uf)AM	6,64049	13,20280	0,50300	0,61502	
as.factor(uf)AP	-3,15476	14,48488	-0,21800	0,827599	
as.factor(uf)BA	6,65924	12,87740	0,51700	0,605098	
as.factor(uf)CE	10,82379	13,19663	0,82000	0,412156	
as.factor(uf)DF	0,54551	12,99200	0,04200	0,96651	
as.factor(uf)ES	4,93707	13,03906	0,37900	0,704978	
as.factor(uf)GO	5,42026	12,90003	0,42000	0,674382	
as.factor(uf)MA	0,27106	13,09269	0,02100	0,983483	
as.factor(uf)MG	6,37358	12,78099	0,49900	0,618036	
as.factor(uf)MS	5,39463	13,10919	0,41200	0,680717	
as.factor(uf)MT	2,30314	13,07521	0,17600	0,860189	
as.factor(uf)PA	5,83840	13,02875	0,44800	0,654094	
as.factor(uf)PB	6,32315	13,09888	0,48300	0,629319	
as.factor(uf)PE	7,43170	13,01173	0,57100	0,567928	
as.factor(uf)PI	2,14637	13,32610	0,16100	0,87205	
as.factor(uf)PR	4,67767	12,80112	0,36500	0,714824	
as.factor(uf)RJ	4,68504	12,84215	0,36500	0,715268	
as.factor(uf)RN	6,34576	13,17795	0,48200	0,630157	
as.factor(uf)RO	7,34970	13,46303	0,54600	0,585154	
as.factor(uf)RR	-2,56336	16,87223	-0,15200	0,879252	
as.factor(uf)RS	3,29925	12,93324	0,25500	0,79866	
as.factor(uf)SC	4,14374	12,92090	0,32100	0,748454	
as.factor(uf)SE	14,46708	13,20945	1,09500	0,273494	
as.factor(uf)SP	4,97223	12,73387	0,39000	0,696208	
as.factor(uf)TO	8,72607	13,27597	0,65700	0,511038	
as.factor(tp_modalidade)1	-0,48021	1,28890	-0,37300	0,709486	
as.factor(gratuito)1	2,12859	4,63328	0,45900	0,645964	
idade	-0,13792	0,04700	-2,93400	0,003361	**
as.factor(tp_sexo)M	0,90034	0,43444	2,07200	0,03829	*
ano_inicio	0,46306	0,12406	3,73200	0,000192	***
as.factor(q1)B	1,54877	0,96655	1,60200	0,109156	

Coeficientes	Estimado	Std. Error	t value	Pr(>|t|)	Significância
as.factor(q1)C	1,03528	1,63100	0,63500	0,525629	
as.factor(q1)D	0,01917	6,96398	0,00300	0,997804	
as.factor(q1)E	1,10349	1,36080	0,81100	0,417466	
as.factor(q2)B	-1,84239	0,80227	-2,29600	0,021702	*
as.factor(q2)C	-0,52226	1,16536	-0,44800	0,654071	
as.factor(q2)D	-1,30587	0,46984	-2,77900	0,005472	**
as.factor(q2)E	-2,41737	3,61995	-0,66800	0,504307	
as.factor(q2)F	-0,58400	1,39774	-0,41800	0,676105	
as.factor(q3)B	0,85340	2,53561	0,33700	0,736463	
as.factor(q3)C	1,91779	4,15717	0,46100	0,644594	
as.factor(q4)B	0,47347	1,24654	0,38000	0,704093	
as.factor(q4)C	0,79285	1,29143	0,61400	0,539296	
as.factor(q4)D	0,55419	1,25829	0,44000	0,65965	
as.factor(q4)E	1,72244	1,34860	1,27700	0,201607	
as.factor(q4)F	1,28043	1,46982	0,87100	0,383727	
as.factor(q5)B	0,80178	1,59408	0,50300	0,615012	
as.factor(q5)C	-0,46653	1,63787	-0,28500	0,775785	
as.factor(q5)D	-0,19411	1,60968	-0,12100	0,904022	
as.factor(q5)E	0,13640	1,67869	0,08100	0,935244	
as.factor(q5)F	0,46830	1,71933	0,27200	0,785351	
as.factor(q6)B	1,55506	1,15186	1,35000	0,177083	
as.factor(q6)C	0,02307	1,35745	0,01700	0,986442	
as.factor(q6)D	1,01273	0,91169	1,11100	0,266711	
as.factor(q6)E	1,72645	4,88469	0,35300	0,723778	
as.factor(q6)F	2,11909	2,12138	0,99900	0,317894	
as.factor(q7)B	0,38460	1,05570	0,36400	0,715646	
as.factor(q7)C	-0,45441	1,09769	-0,41400	0,67892	
as.factor(q7)D	-1,13404	1,10348	-1,02800	0,304155	
as.factor(q7)E	-1,71577	1,16472	-1,47300	0,140801	
as.factor(q7)F	-3,05939	1,34752	-2,27000	0,023239	*
as.factor(q7)G	-1,40278	1,81465	-0,77300	0,439551	
as.factor(q7)H	-2,19254	2,19872	-0,99700	0,318735	
as.factor(q8)B	1,10781	0,67096	1,65100	0,0988	.
as.factor(q8)C	1,62538	0,71359	2,27800	0,022795	*
as.factor(q8)D	3,28426	0,81319	4,03900	5,48E-05	***
as.factor(q8)E	4,38342	0,83097	5,27500	1,40E-07	***
as.factor(q8)F	5,20281	0,96061	5,41600	6,46E-08	***
as.factor(q8)G	7,66768	1,63928	4,67700	3,00E-06	***
as.factor(q9)B	1,23454	0,91297	1,35200	0,176384	
as.factor(q9)C	3,10504	0,93656	3,31500	0,000924	***
as.factor(q9)D	2,99179	1,07626	2,78000	0,005465	**
as.factor(q9)E	0,92481	1,05644	0,87500	0,381413	
as.factor(q9)F	2,43470	1,18026	2,06300	0,039193	*
as.factor(q10)B	-0,72803	0,85540	-0,85100	0,394769	
as.factor(q10)C	-1,16321	0,99958	-1,16400	0,244617	
as.factor(q10)D	-0,68409	0,73041	-0,93700	0,349031	
as.factor(q10)E	-2,41751	0,64257	-3,76200	0,000171	***

Coeficientes	Estimado	Std. Error	t value	Pr(>|t|)	Significância
as.factor(q11)B	-0,83185	1,87853	-0,44300	0,657921	
as.factor(q11)C	5,43514	1,97974	2,74500	0,006072	**
as.factor(q11)D	4,87270	2,27139	2,14500	0,031995	*
as.factor(q11)E	-0,26978	1,88668	-0,14300	0,886304	
as.factor(q11)F	-1,16016	2,29971	-0,50400	0,613952	
as.factor(q11)G	-0,23654	2,27300	-0,10400	0,917123	
as.factor(q11)H	1,32024	1,80850	0,73000	0,46542	
as.factor(q11)I	0,82386	2,27574	0,36200	0,717359	
as.factor(q11)J	1,24918	2,43109	0,51400	0,607396	
as.factor(q11)K	-3,08284	3,15423	-0,97700	0,328449	
as.factor(q12)B	-1,27091	2,37664	-0,53500	0,592854	
as.factor(q12)C	1,90964	1,32866	1,43700	0,150721	
as.factor(q12)D	3,37055	1,62371	2,07600	0,037975	*
as.factor(q12)E	2,02455	1,42150	1,42400	0,154458	
as.factor(q12)F	0,68746	1,60695	0,42800	0,668818	
as.factor(q13)B	2,83306	0,71106	3,98400	6,89E-05	***
as.factor(q13)C	-0,34738	1,34826	-0,25800	0,79669	
as.factor(q13)D	2,56087	0,91607	2,79600	0,005207	**
as.factor(q13)E	1,55748	2,52799	0,61600	0,537869	
as.factor(q13)F	-0,89344	0,83057	-1,07600	0,282132	
as.factor(q14)B	-0,39638	1,57106	-0,25200	0,800822	
as.factor(q14)C	0,03345	2,67249	0,01300	0,990014	
as.factor(q14)D	-2,57771	8,55248	-0,30100	0,763126	
as.factor(q14)E	0,07447	1,59941	0,04700	0,962867	
as.factor(q14)F	-1,95299	1,36713	-1,42900	0,153219	
as.factor(q15)B	-1,47990	1,64026	-0,90200	0,366988	
as.factor(q15)C	1,12258	0,82733	1,35700	0,174904	
as.factor(q15)D	1,68424	0,70393	2,39300	0,016776	*
as.factor(q15)E	2,58688	0,90527	2,85800	0,004292	**
as.factor(q15)F	-1,25701	1,70446	-0,73700	0,460873	
as.factor(q16)12	5,87196	12,61019	0,46600	0,641491	
as.factor(q16)13	-4,16800	5,15466	-0,80900	0,418802	
as.factor(q16)14	2,88203	9,30210	0,31000	0,75671	
as.factor(q16)15	2,30213	4,48509	0,51300	0,607781	
as.factor(q16)16	6,27544	7,34089	0,85500	0,392681	
as.factor(q16)17	-2,66936	5,35088	-0,49900	0,617903	
as.factor(q16)21	2,93510	4,72606	0,62100	0,534605	
as.factor(q16)22	7,61547	5,31694	1,43200	0,152137	
as.factor(q16)23	-0,32522	4,92307	-0,06600	0,947333	
as.factor(q16)24	1,31437	5,02502	0,26200	0,79367	
as.factor(q16)25	2,17127	4,86802	0,44600	0,655603	
as.factor(q16)26	1,38490	4,55469	0,30400	0,761098	
as.factor(q16)27	-5,75878	6,48536	-0,88800	0,374614	
as.factor(q16)28	-7,24040	5,15527	-1,40400	0,160261	
as.factor(q16)29	1,63382	4,14259	0,39400	0,69331	
as.factor(q16)31	2,71019	3,95403	0,68500	0,493117	
as.factor(q16)32	6,34667	4,39991	1,44200	0,149255	

Coeficientes	Estimado	Std. Error	t value	Pr(>|t|)	Significância
as.factor(q16)33	2,61867	4,16075	0,62900	0,52914	
as.factor(q16)35	3,34230	3,85507	0,86700	0,386002	
as.factor(q16)41	3,65918	3,95483	0,92500	0,354897	
as.factor(q16)42	4,99264	4,26454	1,17100	0,241777	
as.factor(q16)43	5,76361	4,31045	1,33700	0,18126	
as.factor(q16)50	2,77898	4,84850	0,57300	0,566567	
as.factor(q16)51	0,82926	4,61235	0,18000	0,857326	
as.factor(q16)52	3,22889	4,20728	0,76700	0,44286	
as.factor(q16)53	7,34044	4,63514	1,58400	0,113355	
as.factor(q16)99	0,55045	4,96880	0,11100	0,911795	
as.factor(q17)B	2,82393	0,53911	5,23800	1,71E-07	***
as.factor(q17)C	8,70087	8,80291	0,98800	0,323014	
as.factor(q17)D	0,10275	1,01825	0,10100	0,919628	
as.factor(q17)E	1,28986	0,97638	1,32100	0,186559	
as.factor(q17)F	6,09608	3,15487	1,93200	0,053398	.
as.factor(q18)B	1,53488	0,55038	2,78900	0,005317	**
as.factor(q18)C	-1,04256	3,37547	-0,30900	0,757441	
as.factor(q18)D	0,11370	1,26100	0,09000	0,928162	
as.factor(q18)E	-4,12811	2,57031	-1,60600	0,108338	
as.factor(q19)B	-1,22004	0,61045	-1,99900	0,045721	*
as.factor(q19)C	-2,13128	0,93943	-2,26900	0,023342	*
as.factor(q19)D	3,43244	1,36117	2,52200	0,011719	*
as.factor(q19)E	-6,32247	5,00708	-1,26300	0,206771	
as.factor(q19)F	-2,04965	1,20000	-1,70800	0,087709	.
as.factor(q19)G	-0,19215	1,31941	-0,14600	0,884217	
as.factor(q20)B	-0,90191	0,90784	-0,99300	0,320542	
as.factor(q20)C	-0,07736	0,50902	-0,15200	0,879211	
as.factor(q20)D	0,14626	1,58732	0,09200	0,926592	
as.factor(q20)E	-2,45258	1,40135	-1,75000	0,080171	.
as.factor(q20)F	-2,27940	2,77518	-0,82100	0,411497	
as.factor(q20)G	1,30207	0,70774	1,84000	0,065878	.
as.factor(q20)H	-0,84046	1,02914	-0,81700	0,414167	
as.factor(q20)I	-0,95185	2,97622	-0,32000	0,749124	
as.factor(q20)J	1,84766	2,00871	0,92000	0,357721	
as.factor(q20)K	0,11565	0,94064	0,12300	0,902158	
as.factor(q21)B	-0,63237	0,49358	-1,28100	0,200206	
as.factor(q22)B	-1,44976	0,53238	-2,72300	0,006495	**
as.factor(q22)C	-0,92651	0,58742	-1,57700	0,114817	
as.factor(q22)D	-0,88188	0,93430	-0,94400	0,345284	
as.factor(q22)E	-0,87754	0,83864	-1,04600	0,295449	
as.factor(q23)B	0,19094	0,93922	0,20300	0,838914	
as.factor(q23)C	0,56903	0,96600	0,58900	0,555853	
as.factor(q23)D	1,79423	1,04506	1,71700	0,086083	.
as.factor(q23)E	2,15547	1,06954	2,01500	0,04394	*
as.factor(q24)B	-1,90279	1,80277	-1,05500	0,291271	
as.factor(q24)C	0,27428	0,86166	0,31800	0,750265	
as.factor(q24)D	1,19513	0,85043	1,40500	0,160003	

Coeficientes	Estimado	Std. Error	t value	Pr(>|t|)	Significância
as.factor(q24)E	-0,14124	0,48829	-0,28900	0,772402	
as.factor(q25)B	-1,10215	0,80561	-1,36800	0,171363	
as.factor(q25)C	-0,91334	0,55549	-1,64400	0,100217	
as.factor(q25)D	-2,13755	1,62578	-1,31500	0,18866	
as.factor(q25)E	1,06553	0,50444	2,11200	0,034725	*
as.factor(q25)F	3,30675	3,85593	0,85800	0,39118	
as.factor(q25)G	-5,13364	3,18554	-1,61200	0,107142	
as.factor(q25)H	-0,44903	0,63769	-0,70400	0,481385	
as.factor(q26)B	0,10682	1,08563	0,09800	0,921626	
as.factor(q26)C	2,29045	0,81784	2,80100	0,005126	**
as.factor(q26)D	-0,87014	1,54862	-0,56200	0,574232	
as.factor(q26)E	-1,08274	1,10223	-0,98200	0,326003	
as.factor(q26)F	2,63548	0,70134	3,75800	0,000174	***
as.factor(q26)G	-0,14497	1,50047	-0,09700	0,923034	
as.factor(q26)H	1,47700	1,01953	1,44900	0,147502	
as.factor(q26)I	1,00420	0,95500	1,05200	0,293085	