Trabalho de Conclusão de Curso

Impacto de fatores socioeconômicos na nota do ENADE: Uma investigação usando métodos de Aprendizagem de Máquina

de **Edvonaldo Horácio dos Santos**

orientado por
Prof. Dr. Bruno Almeida Pimentel
Prof. Msc. Lucas Benevides Viana de Amorim

Universidade Federal de Alagoas
Instituto de Computação
Maceió, Alagoas
20 de Janeiro de 2021

UNIVERSIDADE FEDERAL DE ALAGOAS
Instituto de Computação

IMPACTO DE FATORES SOCIOECONÔMICOS NA NOTA DO ENADE: UMA INVESTIGAÇÃO USANDO MÉTODOS DE APRENDIZAGEM DE MÁQUINA

Trabalho de Conclusão de Curso submetido ao Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Engenheiro de Computação.

Edvonaldo Horácio dos Santos

Orientador: Prof. Dr. Bruno Almeida Pimentel
Coorientador: Prof. Msc. Lucas Benevides Viana de Amorim

Banca Avaliadora:

Evandro de Barros Costa	Prof. Dr., UFAL
Roberta Vilhena Vieira Lopes	Prof. Dr., UFAL

Maceió, Alagoas
20 de Janeiro de 2021

UNIVERSIDADE FEDERAL DE ALAGOAS
Instituto de Computação

IMPACTO DE FATORES SOCIOECONÔMICOS NA NOTA DO ENADE: UMA INVESTIGAÇÃO USANDO MÉTODOS DE APRENDIZAGEM DE MÁQUINA

Trabalho de Conclusão de Curso submetido ao Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Engenheiro de Computação.

Aprovado em 20 de Janeiro de 2021:

Bruno Almeida Pimentel,
Prof. Dr., Orientador

Lucas Benevides Viana de Amorim,
Prof. Msc., Coorientador

Evandro de Barros Costa,
Prof. Dr., UFAL

Roberta Vilhena Vieira Lopes,
Prof. Dr., UFAL

Dedicatória

À minha mãe, pai (*in memorian*) e irmãos que tanto amo.

Agradecimentos

Agradeço a todas as pessoas que compõem o corpo do Instuto de Computação da UFAL, especialmente aos meus orientadores, os professores Dr. Bruno Almeida Pimentel e Msc. Lucas Benevides Viana de Amorim, que foram fundamentais para a realização deste trabalho. Sem suas orientações e direcionamentos, seria impossível concluí-lo.

Agradeço também à minha família, especialmente minha mãe, Maria José Paulino, e um dos meus irmãos, Edvan Horácio. Não obstante toda minha família ter sido importante, a menção dos anteriores vem representar bem todo o apoio que recebi desde a tenra infância. Sempre que puderam, não exitaram esforços para me ajudar.

Agradeço à minha noiva, Thaline, por seu amor para comigo. Aos meus amigos e amigas: Alfredo, Andressa e Leony, incríveis parceiros em projetos de disciplinas do curso; Anderson, Danilo, Ebson, Erving, Laura, Marco Antonio, Mayres, Rafael, Thaís e Wagner: amigos de outros cursos na UFAL e cujas conversas são estimulantes e uma brisa para à alma.

Aos grupos cristãos: GrACE e Núcleo da ABC^2 em Maceió. As ótimas conversas que pude ter neles, inclusive sobre computação, foram importantes para aguçar minha mente à natureza interdisciplinar das ciências.

Não poderia deixar de fora a gratidão que é participar da experiência adquirida sobre este estudo, que por meio dos algoritmos propostos, advindos de uma solidez matemática, fornecem possibilidades que apenas em sala de aula não é possível aplicar. Cada disciplina que vi ao longo do curso teve sua contribuição, mesmo que esta tenha sido tão somente em criar o hábito de pensar computacionalmente em um problema do mundo real.

Finalmente, agradeço ao Deus trino, criador e doador de todas coisas.

Se um dia pude ver mais longe, é porque subi no ombro de gigantes. Isaac Newton (adaptado).

Resumo

O Enade é um teste aplicado anualmente em graduandos habilitados. Integrando o Sineas, permite, junto com a avaliação institucional e a avaliação de cursos da graduação, medir a qualidade dos cursos no país. Os resultados podem ser consultados no site do Inep e permitem uma análise por meio do uso de técnicas de Aprendizagem de Máquina e Ciência de Dados. A proposta deste trabalho é investigar quais fatores socioeconômicos afetam a nota geral do Enade. São utilizados três algoritmos de regressão: Árvore de Decisão, Floresta Aleatória e Lasso. Para o estudo os dados são separados em dois conjuntos: classe A referente apenas aos alunos do estado de Alagoas; e classe B referente aos dos demais estados da federação (excluindo-se aqueles da classe A). Apresentaram alta influência na nota geral da prova para ambas as classes e os três algoritmos os seguintes fatores socioeconômicos: bolsa acadêmica durante a graduação (monitoria, PIBICs, PIBITs e análogos), fonte de financiamento para mensalidade (no caso do curso não ser gratuito) e a renda total familiar. Também se sobressaíram a cor do indivíduo, a modalidade do ensino médio realizada pelo aluno e o número de horas de estudo por semana.

Palavras-chave*: *Aprendizado de Máquina*;*Árvore de decisão*; *Ciência de Dados*; *Enade*; *Fatores socioeconômicos*; *Floresta aleatória*; *Lasso*; *Regressão.

Abstract

Enade is a test applied annually to qualified undergraduates and is used to measure the quality of undergraduate courses in the country. The results can be consulted on the Inep website and allow an analysis through the use of Machine Learning techniques and Data Science. The purpose of this paper is to investigate which socioeconomic factors affect Enade's overall score. Three regression algorithms are used: Decision Tree, Random Forest and Lasso. For the study the data are separated into two sets: class A referring only to students from the state of Alagoas; and class B referring to those of the other states of the federation (excluding those of class A). The following socioeconomic factors had a high influence on the overall score of the test for both classes and the three algorithms: academic scholarship during graduation (tutoring, PIBICs, PIBITs and the like), funding source for tuition (if the course is not free) and total family income. Also highlighted were the individual's color, the modality of high school held by the student and the number of hours of study per week.

***Keywords*: *Data science*; *Decision tree*; *Enade*; *Machine Learning*; *Random forest*; *Regression*; *Lasso*; *Socioeconomic factors*.**

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ENADE	Exame Nacional de Desempenho dos EstudantesEnade
EDM	Mineração de Dados Educacionais
KDD	Descoberta de Conhecimento em Base de Dados
IA	Inteligência Artificial
MAE	Erro Absoluto Médio
MSE	Erro Quadrático Médio
SINEAS	Sistema Nacional de Avaliação da Educação Superior

Sumário

Capítulo 1

Introdução

No ano de 2004 foi criado o Exame Nacional de Desempenho dos Estudantes (Enade). Sendo obrigatório para os concluintes habilitados, ele é aplicado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), e substituiu o antigo Exame Nacional de Cursos (chamado de Provão). O objetivo envolve avaliar as competências e habilidades dos estudantes com relação a formação geral e profissional, levando em conta os conteúdos programáticos em cada curso. Além disso, o exame também é importante para mensurar o nível de atualização dos graduandos de instituições superiores do país em relação às realidades nacionais e internacionais, fazendo parte do Sistema Nacional de Avaliação da Educação Superior (Sineas).

A prova é composta por questões objetivas e discursivas referentes à formação geral e aos componentes específicos de cada curso, cuja pontuação varia entre 0 (menor nota) e 100 (maior nota); e os questionários do estudante, que envolvem a percepção do discente sobre a prova, perguntas referentes a contribuições das disciplinas do curso para a formação, um questionário exclusivo às licenciaturas e um conjunto de perguntas comum a todos que formam o perfil socioeconômico do candidato. A nota geral da prova é uma média ponderada, no qual 25% do peso corresponde à formação geral e 75% a formação específica. Os resultados referentes a todos os componentes supracitados compõem uma base de dados, chamados de microdados, que pode ser consultada no site oficial do Inep.

Em virtude da disponibilidade pública de dados relacionados a educação (nacionalmente e internacionalmente) junto ao desenvolvimento computacional e da Inteligência Artificial (IA), tornou-se possível a análise e extração de informações. É nesse ambiente que se consolidou na última década a Mineração de Dados Educacionais (EDM). A aplicabilidade nesse cenário é enorme: de que forma se pode melhorar a educação em todos os níveis? Quais fatores socioeconômicos são importantes no desempenho do discente e por quê? Uma vez que os recursos (sejam eles financeiros e/ou humanos) são limitados, o uso de técnicas de aprendizagem de máquina é uma ferramenta importante para investigar as respostas nessa direção.

O processo de extração de informações envolve primeiramente a análise da

base de dados que se deseja tratar. Isso é possível por meio do uso do dicionário de variáveis que, no caso do Enade, vêm junto com os microdados coletados no site oficial do Inep correspondente ao ano desejado. A inspeção para verificação de valores ausentes e/ou inconsistentes pode trazer problemas na execução do algoritmo e obtenção dos resultados.

Levando-se em conta que a base foi previamente analisada, é preciso escolher quais algoritmos são potencialmente desejáveis para o treinamento e a extração da importância relativa de cada variável. Essa tarefa pode ser trabalhosa, pois pode envolver o ajuste de parâmetros. Tendo isso em mente, foram escolhidos os seguintes algoritmos: Árvore de Decisão, Floresta Aleatória e Lasso.

No Capítulo 2, será um breve *background* sobre desenvolvimento da Inteligência Artificial, área que tem intersecção com a Ciência de Dados. Também será mostrado como se dá o processo de Descoberta de Conhecimento em Base de Dados (KDD) e como ele é útil para o trabalho, dando suporte para a extração de informações na base utilizada. Além disso, serão abordados os fundamentos teóricos que descrevem o funcionamento dos algoritmos utilizados.

Posteriormente serão descritas quais foram as ferramentas utilizadas para o uso dos algoritmos e as configurações da máquina em que foram aplicados. Com base na fundamentação teórica correspondente ao processo de KDD, será mostrado como esses passos foram aplicados na base, desde a escolha dos dados até os parâmetros que foram usados em cada algoritmo para treiná-los.

Em seguida, serão expostos quais foram os resultados obtidos com os algoritmos usados, incluindo as métricas que exibem o desempenho. Através da exibição de gráficos, a importância de cada variável para a Árvore de Decisão, Floresta Aleatória e Lasso foi plotada, de modo que cada método pudesse ser comparado com o outro. Os resultados serão discutidos conforme a fundamentação teórica e com as ferramentas que a Ciência de Dados fornece para análise.

Por fim, na última parte, serão explicitadas as principais conclusões que o trabalho pode fornecer de levando-se em conta as ferramentas descritas anteriormente, mencionando a contribuição do que foi desenvolvido durante a investigação dos dados. Nesse sentido, como quaisquer outras áreas da ciência, serão apresentadas limitações e quais investigações mais profundas o trabalho poderia ser analisado.

Capítulo 2

Fundamentação Teórica

Ao longo deste capítulo será descrito o arcabouço teórico para trabalhar o problema. Primeiro passaremos por uma síntese histórica, falando do desenvolvimento da inteligência artificial. Logo em seguida, focaremos nas áreas de aprendizado de máquina e Ciência de Dados, finalizando com os algoritmos usados nesse trabalho.

2.1 Aprendizado de Máquina

A inteligência artificial (IA) é uma das áreas mais recentes em ciências. Seu desenvolvimento começou após a II Guerra Mundial, quando John McCarthy propôs com outros pesquisadores estudarem juntos teoria dos autômatos, redes neurais e estudo da inteligência. Ela engloba o conhecimento de diversas outras disciplinas, como por exemplo Matemática, Biologia e Psicologia. Apesar de surgir na década de 50, ela só começou a alavancar após os avanços da eletrônica (décadas de 70 e 80), que possibilitaram a criação de computadores mais potentes cuja aplicação em IA foi e ainda é fundamental.

Dentre as subáreas da IA, podemos destacar o Processamento de Linguagem Natural, Visão Computacional, Sistemas Especialistas e Aprendizado de —Máquina. Por ser tão abrangente (e fascinante), as aplicações dela incluem Reconhecimento de Fala, Problemas de Busca, Visão Computacional, Reconhecimento de Padrões em Dados, entre outros.

O processo de aprendizado de máquina envolve a aplicação de abstrações por meio de representações matemáticas que visam transmitir quando um agente está aprendendo, isto é, quando seu desempenho em tarefas futuras é melhorado. Ele pode ser dividido basicamente em três tipos:

1. **Aprendizado Supervisionado**
2. **Aprendizado Não-Supervisionado**
3. **Aprendizado por Reforço**

Aprendizado Supervisionado

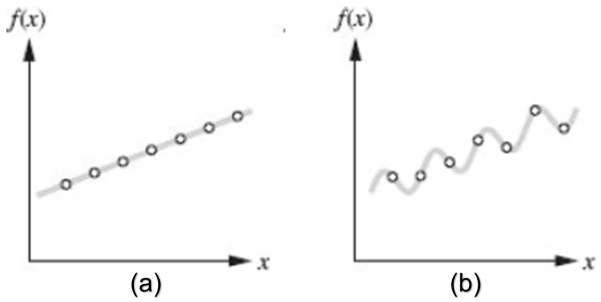
Nesse tipo de aprendizagem, o agente observa alguns exemplos de pares de entrada e saída para descobrir qual função realiza o mapeamento correto. Nela, os valores de saídas estão disponíveis diretamente para o agente. [Russell and Norvig, 2004] assim conceitua o mapeamento de acordo com a Definição (2.1.1).

Definição 2.1.1. Dado um **conjunto de treinamento** de N pares de exemplos de entrada e saída $(x_1, y_1, x_2, y_2, ..., x_n, y_n)$, com x_i, $y_1 \in \mathbb{R}^n$, onde cada y_i foi gerado por uma função desconhecida y = f(x_i), há uma função h (chamada de hipótese) que mais se aproxima do comportamento de f.

Para mensurar a qualidade da a hipótese, é necessário fornecer um conjunto de testes de exemplos que são distintos do conjunto de treinamento. Nesse aspecto, quando a saída y é um conjunto finito de valores (por exemplo: sim, não; ótimo, bom, regular; classe A, classe B, classe C), o problema é chamado de classificação. Se a saída y for um número (por exemplo: temperatura, notas de uma prova), o problema é de regressão. Vamos dar atenção ao processo de aprendizagem por regressão, descrevendo os três algoritmos usados no trabalho.

Um exemplo clássico de regressão é encontrar uma determinada função f que se ajuste aos pontos $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, onde $x_i, y_i \in \mathbb{R}$. Pode-se encontrar alguns exemplos de funções na Figura (2.1). Note-se que, não raro, várias são as funções que podem se ajustar ao conjunto de dados.

Figura 2.1: Exemplos de Regressão.



Fonte: Russell e Norvig (2013) adaptado.

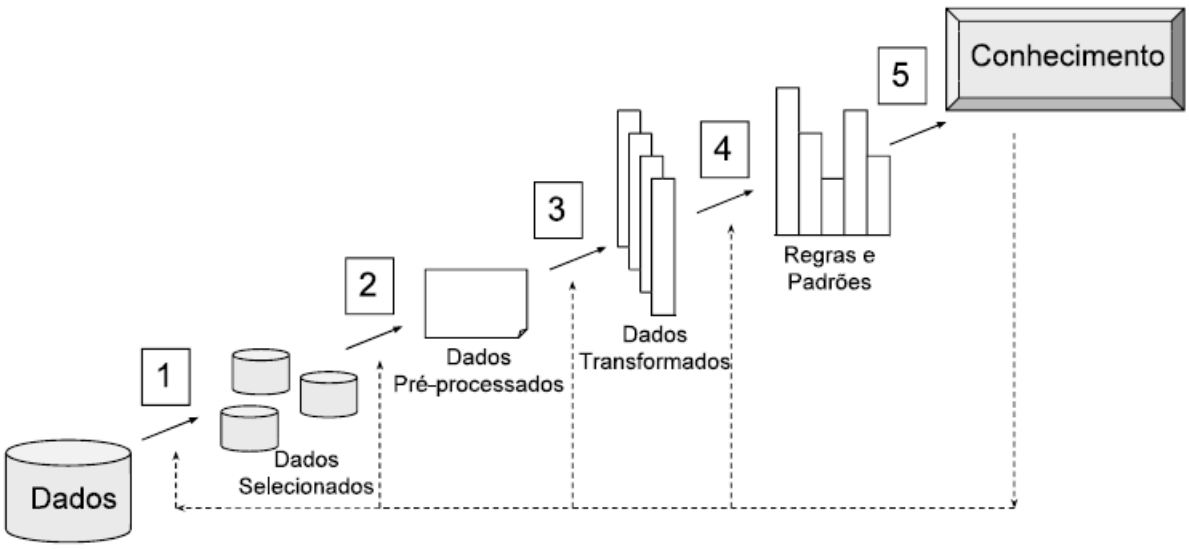
Naturalmente, para obter essas representações que descrevem os dados, bem como realizar análises sobre o problema, são necessárias algumas etapas. Elas serão abordadas na Seção 2.1.1 que trata sobre **Processo de Aprendizagem.**

2.1.1 Processo de Aprendizado

A aprendizagem de máquina envolve a descoberta de padrões e/ou regras após o agente realizar observações sobre o mundo. Nessa abordagem, as extrações são realizadas por meio do uso de um banco de dados seguindo alguns passos bem consolidados. Apesar de haver alguma divergência quanto ao número, os tipos são bem definidos. [Fayyad et al., 1996] define os 5 passos no aprendizado, que também é chamado de Mineração de Dados ou KDD. A Figura (2.2) exibe cada um deles, os quais são:

1. **Seleção de Dados**
2. **Pré-processamento (Limpeza e Enriquecimento)**
3. **Transformação**
4. **Mineração**
5. **Interpretação e Avaliação**

Figura 2.2: Processo de Aprendizagem (KDD).



Fonte: U.M. Fayyday et. al. adaptado (1996).

Seleção de Dados

Essa etapa diz respeito a escolha dos dados a serem usados. Eles podem ser nos mais diversos formatos (.txt, .sql, entre outros), sendo o csv (comma-separated values) um dos mais usados devido à facilidade na manipulação, que auxilia na próximas etapas do aprendizado. Além disso, é importante definir bem a pergunta que desejamos responder,

pois ela vai nortear onde queremos chegar, bem como os melhores algoritmos e técnicas nos próximos passos.

Pré-processamento (Limpeza e Enriquecimento)

Os dados provindos no processo de escolha geralmente possuem campos não desejados e/ou valores ausentes. Essa etapa busca resolver esses problemas.

1. **Limpeza:** são removidos da base de dados aquelas informações não desejadas, como erros nas informações (campos com números em formatos não desejados, categorias inexistentes entre outras).

2. **Enriquecimento:** é comum em grandes quantidades de dados, haver ausências de valores. O enriquecimento tem o objetivo de preencher esses campos para o processamento nas próximas etapas.

Transformação

No processo de aprendizado, além de variáveis numéricas é comum o uso daquelas categóricas, de texto e grupos de valores que podem ser reduzidos sem perda de informação importante. Diante disso, como realizar essa ligação entre o que a máquina consegue processar e as informações originadas da base de dados? A resposta a essa questão estão no uso de transformações na base de dados, tornando-a compreensível à máquina e facilitando na etapa de mineração. Para isso, são aplicadas técnicas de transformação que fazem essa ponte. As mais comuns são a normalização, a discretização de atributos quantitativos e a transformação de atributos qualitativos em quantitativos. Dentro delas, há diversos algoritmos, como one-hot encoding que foi usado nesse trabalho.

Em diversas situações envolvendo aprendizagem de máquina, os dados a serem utilizados da base são variáveis que representam categorias, como por exemplo, homem e mulher; Sim e Não; Ruim, Regular e Bom; e assim sucessivamente. Contudo, o algoritmo de aprendizagem não consegue usar essas categorias. É preciso um meio de quantificá-las, tornando-as utilizáveis na aprendizagem. Nessa problemática o one-hot encoding é uma técnica que transforma as categorias uma matriz, na qual cada coluna é uma categoria, e as linhas (em codificação binária) contêm o valor sinalizando que a categoria está presente ou não.

Como exemplo, considere a seguinte situação (vamos chamá-la de situação Pessoa-Renda): Temos em pequeno conjunto de dados em .csv onde há 3 pessoas (chamaremos de Pessoa 1, Pessoa 2 e Pessoa 3) e usando o one-hot encoding vamos codificar os dados de acordo com sua faixa de renda bruta familiar, sendo previamente definidas 3 categorias, as quais são:

1. Categoria A: Até 1 salário mínimo;
2. Categoria B: De 1 salário mínimo até 2 salários mínimos;
3. Categoria C: De 2 salários mínimos até 3 salários mínimos.

Sabe-se que a Pessoa 1 está na Categoria B, a Pessoa 2 está na Categoria C e a Pessoa 3 está na Categoria A. Antes da codificação, podemos representar essa situação de acordo com a Tablela (2.1).

Tabela 2.1: Representação da Situação Pessoa-Renda **antes** do One-Hot Encoding

Pessoa	**Categoria**
1	B
2	C
3	A

Fonte: Autor.

Tabela 2.2: Representação da Situação Pessoa-Renda **após** o One-Hot Encoding

	A	**B**	**C**
1	0	1	0
2	0	0	1
3	1	0	0

Fonte: Autor.

Aplicando-se o one-hot encoding a matriz se dá conforme a Tabela (3.4). Conforme se pode observar, aplicar essa técnica facilita bastante no processo de aprendizagem, uma vez que os dados ficam bem representados e não há perda de informação.

Mineração

É a etapa mais importante no processo de aprendizado. Costuma gastar grande quantidade de tempo, especialmente quando se trabalha com grandes volumes de dados. Ela envolve a aplicação de algoritmos que buscam encontrar os padrões e regras do problema.

Para realizar o treinamento adequado dos algoritmos com os dados pré-processados, é preciso dividi-los em dois: conjunto de treino e conjunto de teste. Para isso, são aplicados algoritmos que realizam a divisão, como o holdout, subamostragem aleatória e a validação cruzada. Como foi usado o este último para realizar o ajuste de parâmetros no conjunto de treino, e devido a sua maior complexidade, ele é descrito a seguir e representado na Figura (2.3).

1. A base é particionada aleatoriamente em k partes com aproximadamente o mesmo tamanho (em condições ideais, são do mesmo tamanho).

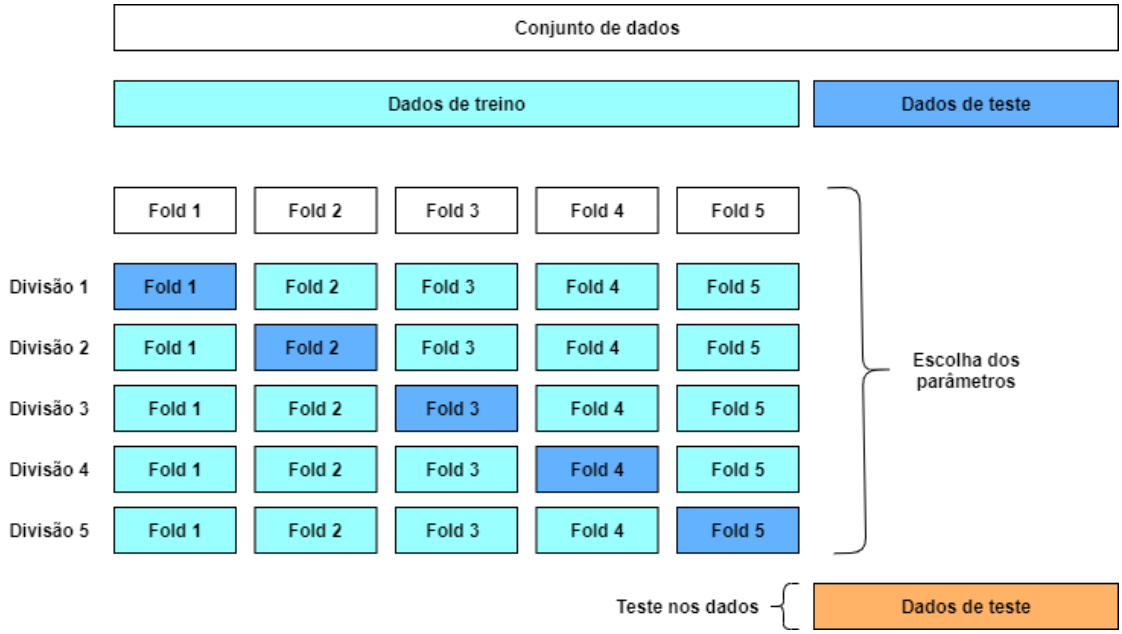
2. Os conjuntos de treinamento e teste são executados k vezes em cada execução:

   i) 1 partição de teste.

   ii) *k-1* partições de treinamento.

3. Todas as partições são usadas em algum momento para teste.

Figura 2.3: Processo K-Fold de Validação Cruzada para escolha dos parâmetros.



Fonte: Página da biblioteca scitkit-learn, adaptado.

Para treino e teste foi usado o holdout, que consiste em dividir a base em duas: uma para treino e outra para teste, normalmente usando a proporção de 2/3 para treino e 1/3 para teste. Essa técnica também é representado pela Figura (2.3), no passo anterior ao teste com os parâmetros.

Conforme já abordado no início do capítulo, primeiro devemos verificar qual tipo de aprendizagem. Na nossa aplicação, de acordo com a definição 2.2.1, é um problema que se categoriza pelo tipo de aprendizagem supervisionada que usa regressão. Assim, nossa abordagem envolve a aplicação de algoritmos de regressão que buscam encontrar a função hipótese que melhor descreve o comportamento dos dados, mais precisamente os seguintes:

1. **Árvore de Decisão**
2. **Floresta Aleatória**
3. **Lasso**

A partir deles, vamos usar métricas apropriadas de avaliação, que devem medir a **acurácia** de cada um, além de avaliar quais são as variáveis mais importantes para predição das notas.

Interpretação e Avaliação

Após a etapa de mineração, é preciso avaliar como foi o desempenho de cada algoritmo, buscando calcular as métricas e verificando se é necessário realizar mudanças em parâmetros ou mesmo nos algoritmos escolhidos para o problema.

Para cada tipo de algoritmo (classificação, regressão) há métricas específicas que mensuram o desempenho. Em aprendizagem supervisionada usando regressão, destacam-se o **MAE (erro médio absoluto)**, o **MSE (erro quadrático médio)** e o R **(r ao quadrado)**.

Sejam N o número de amostras, $\hat{y}_i$ o valor predito, y_i o valor real, $\bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$, onde i é a i-ésima amostra, então:

1. O MAE é definido pela Equação (2.1).

$$MAE(y, \hat{y}) = \frac{1}{N} \sum_{N}^{i=1} |y_i - \hat{y}_i| \tag{2.1}$$

2. O MSE é dado pela Equação (2.2).

$$MSE(y, \hat{y}) = \frac{1}{N} \sum_{N}^{i=1} (y_i - \hat{y}_i)^2 \tag{2.2}$$

3. O R^2 score é definido conforme a Equação (2.3).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2.3}$$

Para o MAE e o MSE, quanto mais próximos de 0 os valores, melhor a predição do modelo. No caso do R^2 score podemos ter três possibilidades:

1. Se $R^2 < 0$, então o modelo é ruim;
2. Se $R^2 = 0$, então o modelo está prevendo valores dentro de uma variância esperada, isto é, compatível com os dados, apesar de possuir pouco poder explicativo;
3. Se $R^2 > 0$, então o modelo, além de prever valores dentro da variância dos dados, possui poder explicativo das variáveis dependentes; o melhor caso possível é quando $R^2 = 1$, mas que por si só pode indicar um sobreajuste (overfiting) e por isso deve ser analisado com outras métricas de descrição dos dados.

2.2 Ciência de Dados

Durante as décadas de 70 e 80, empresas e centros de pesquisas começaram a acumular grande quantidade de dados e, a partir da década de 90, esse conglomerado de informações, junto ao desenvolvimento da IA com o poder computacional, possibilitou a extração de novas informações. É nesse contexto que nasce a Ciência de Dados. Assim [Igual and Seguí, 2017] mostra que esta é uma área interdisciplinar com foco no estudo e na análise de dados econômicos e sociais, sejam eles estruturados ou não, extraindo conhecimento a partir do reconhecimento de padrões não triviais e a obtenção de *insights* para a tomada de decisão. Além disso, [Igual and Seguí, 2017] destaca 4 estratégias para exploração dos dados:

1. Exame da realidade.
2. Descoberta de padrões.
3. Predição de eventos futuros.
4. Compreensão das pessoas e do mundo.

Na área da educação, a mineração de dados é uma área em grande desenvolvimento. Ela é conhecida como **Mineração de Dados Educacionais (EDM)**. De acordo com [Romero and Ventura, 2013] o intuito é descobrir como estudantes aprendem e que configurações ajudam no bom desempenho do docente. Nesse contexto, o trabalho apresentado visa extrair quais fatores socioeconômicos impactam na nota geral do Enade, exame realizado anualmente pelo Inep.

Semelhante aos demais problemas do mundo real, a área da educação tem suas particularidades e características especiais [Romero and Ventura, 2013]. Isso significa que apesar dos algoritmos serem genéricos, isto é, idealizados para aplicações em diversas áreas, é preciso escolher os recursos mais apropriados para a análise.

2.3 Regressão usando Árvore de Decisão

Árvore de Decisão (DT) é um algoritmo de aprendizado de máquina supervisionado e não paramétrico, sendo aplicada tanto para problemas de classificação quanto de regressão. Seu objetivo geral é criar um modelo que consiga predizer uma variável a partir de regras de decisão inferidas a partir do conjunto de testes. Ou, de forma mais técnica e usando a Definição (2.1.1), encontrar a função h que melhor se ajuste aos dados.

Para alcançar a decisão, ela executa uma sequência de testes. Cada nó (seja a raíz seja um nó interno) corresponde a um teste do valor dos atributos da entrada, e as arestas dos nós são as classificações de acordo com os possíveis valores do atributo. Assim

Figura 2.4: Árvore de Decisão com 3 níveis

Fonte: Autor.

a condição é formada pelo nó e pela aresta. Cada caminho da raiz até a folha representa uma regra, que são conjunções percorridas que geram um valor (número, no caso de uma árvore de decisão por regressão). A Figura (2.4) descreve a forma genérica de uma árvore de decisão com 3 níveis.

Dado que a entrada constitui de diversos atributos (categóricos, numéricos ou ambos), para escolher o melhor critério são usados diversos algoritmos, os quais merecem citar: ID3 (iterative dichotomiser 3), C4.5, C5.0 e o CART (arvores de regressão e classificação). Visto que o aprendizado no problema deste trabalho parte de variáveis categóricas nominais para prever variáveis numéricas, o algoritmo usado foi o CART.

Algoritmo CART (Classification And Regression Trees)

Baseado no trabalho de [Breiman et al., 1984] sobre arvores de classificação e regressão, o algoritmo CART é uma árvore binária de decisão construída dividindo-se o nó em dois outros nós (filhos) repetidamente, iniciando da raiz da árvore, que contém os critérios de aprendizado. Para problemas de regressão o erro quadrático médio (MSE) é usado no algoritmo CART para mensurar o critério de impureza e realizar a divisão, que é feita em 3 passos:

1. Encontre a feature que proporciona a melhor divisão;
2. Encontre o melhor nó para a divisão: Entre as divisões do passo anterior, escolhe-se àquela que maximiza o critério de divisão (maior valor do MSE);
3. Realize a divisão usando o nó encontrado no passo 2 até que o critério de parada seja satisfeito (MSE baixo ou homogêneo).

Sendo Y = $\{y_1, y_2, ..., y_i\}$ o conjunto que designa a variável dependente, com i = 1, 2, 3, ..., n onde n $\in$ N, isto é, o número total de variáveis dependentes, o MSE é dado por:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{2.4}$$

onde $\overline{y}$ é a média do conjunto Y. Define-se ainda que quanto maior é o valor do MSE para um nó t, maior é seu grau de impureza; caso contrário, o grau de pureza deste nó é maior. Uma das grandes utilidades dessa métrica para a árvore de regressão, é medir a importância de cada variável de entrada para o modelo.

Importância de Variável Para Árvore de Decisão

Haja vista que deseja-se mensurar o impacto de cada variável dependente na predição da nota, o recurso mais apropriado para isso é fornecido pelas features importances da árvore. Baseado no trabalho de [Breiman et al., 1984], [Grömping, 2009] define a importância de uma variável (VImp) x_i para uma árvore T como:

$$VImp(x_i, T) = \sum_{t \in T} \Delta(x_i, t) \tag{2.5}$$

onde $\Delta(x_i, t)$ é a redução da impureza por causa de uma real ou potencial divisão na variável x_i para o nó t. Como a impureza é dada pelo MSE, de modo geral quanto maior é a redução da impureza, maior é a importância daquela variável; igualmente, quanto menor a redução de impureza, menor é sua importância.

Para fins de utilização e comparação, a importância de cada variável é normalizada, de modo que:

$$\sum_{i=1}^{N} VImp(x_i^{normalizado}) = 1 \tag{2.6}$$

Algumas das vantagens e desvantagens de uma Árvore de Decisão, de acordo com [James et al., 2013], se encontram nas Tabelas (2.3) e (2.4).

Tabela 2.3: Vantagens de uma Árvore de Decisão

Vantagens
Simples entendimento e fácil interpretação.
Requer uma preparação de dados pouco exaustiva.
O custo no uso da árvore é logarítmico.
Uso de variáveis categóricas e numéricas.
Validação dos testes usando testes estatísticos.
Apresenta um bom desempenho.

Fonte: [James et al., 2013].

Tabela 2.4: Desvantagens de uma Árvore de Decisão

Desvantagens
O processo de aprendizado pode ser complexo e não representar bem os dados.
Podem ser instáveis.
O algoritmo é guloso e não garante a solução ótima.
O aprendizado pode criar um viés nas variáveis dominantes.

Fonte: [James et al., 2013].

2.4 Regressão usando Floresta Aleatória

Uma Floresta Aleatória (Random Forest) é formada por um número elevado de árvores de decisão. Esse número de árvores é especialmente importante quando é necessário analisar as importâncias das variáveis de entrada. A aleatoriedade no algoritmo pode ser vista de duas formas possíveis:

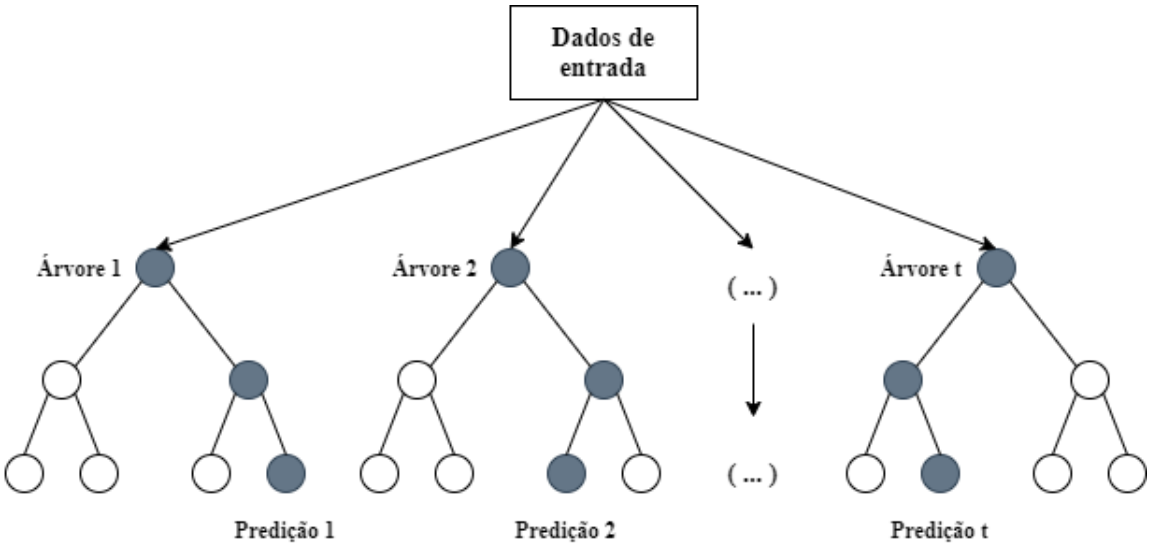
i) Cada árvore é baseada é um subconjunto aleatório de observações;

ii) Cada divisão em cada árvore é criada a partir de um conjunto aleatório de variáveis candidatas.

Devido a essa aleatoriedade característica, as árvores individuais podem ser bastante instáveis, criando predições bem distintas em cada árvore. A predição nessa técnica é dada pela predição média do conjunto formado pelas árvores individuais. Podemos visualizar uma Floresta Aleatória genérica, com número **t** de árvores, na figura 2.4.

A Floresta Aleatória é construída usando em cada árvore o algoritmo CART da Árvore de Decisão, ideia esta proposta por [Breiman, 2002]. Após selecionar os nós de divisão de forma aleatória, uma média de 1/3 dos dados não são usadas em nenhuma árvore individual. Eles são "**out of bag (OOB)**". Levando isso em consideração e com muita semelhança para a Árvore de Decisão, a acurácia de predição para a Floresta Aleatória é dada por:

Figura 2.5: Floresta Aleatória genérica



Fonte: Autor.

$$OOB - MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{\hat{y}}_{iOOB})^2 \tag{2.7}$$

onde $\overline{\hat{y}}_{iOOB}$ é predição para a i-ésima observação de todas as árvores para as quais essa observação foi OOB.

As principais vantagens e desvantagens mencionadas por [Buskirk, 2018] a respeito da floresta aleatória estão nas Tabelas (2.5) e (**??**), respectivamente.

Tabela 2.5: Vantagens da Floresta Aleatória

Vantagens
Grande eficiência em grandes bancos de dados.
Reduz o problema de overfitting e a variância nas árvores de decisão.
Funciona bem com variáveis categóricas e contínuas.
Florestas aleatórias são eficazes na predição de funções complexas com grande interações ou de funções não lineares.

Fonte: [Buskirk, 2018].

Tabela 2.6: Desvantagens da Floresta Aleatória

Desvantagens
A Floresta Aleatória pode ser computacionalmente cara.
A importância de variável pode ser enviesada caso as variáveis estejam correlacionadas.
Os dados ausentes devem ser tratados antes da execução do algoritmo.

Fonte: [Buskirk, 2018].

Importância de Variável Para Floresta Aleatória

Para mensurar as importâncias das variáveis em uma floresta aleatória usando regressão através do algoritmo CART usa-se a redução média das impurezas. Contudo, quando são efetuadas as seleções das variáveis há um viés na escolha delas a partir dessa métrica (os recursos baseados em valores numéricos são vistos como mais importantes), cujo corolário natural é tornar as métricas de importância de variável também tendenciosas. Diante desse problema, para escolha adequada do valor de feature importance, [Breiman, 2002] propõe a redução de MSE (MSE reduction), usado no MSE com permutação de uma variável. A redução de MSE é explicado conforme as Definições (2.4.1) e (**??**):

Definição 2.4.1. Para uma árvore **t** qualquer, o MSE-OOB para sua respectiva predição é:

$$OOB - MSE_t = \frac{1}{N_{OOB,t}} \sum_{i=1;i\in OOB_t}^{N} (y_i - \hat{y}_{i,t})^2 \tag{2.8}$$

onde $\hat{y}_{i,t}$ indica a predição i da árvore t, OOB_t = {observação i que é OOB da árvore **t**}, $N_{OOB,t}$ é o número de observações OOB na árvore **t**.

Definição 2.4.2. Se o X_j não tem valor preditivo para a resposta, não há diferença caso esses valores sejam aleatoriamente permutados nas observações OOB antes das predições serem geradas. Equacionando, têm-se:

$$OOB - MSE_t(X_j permutado) = \frac{1}{N_{OOB,t}} \sum_{i=1;i\in OOB_t}^{N} (y_i - \hat{y}_{i,t}(X_j permutado))^2 \tag{2.9}$$

Por fim, usando o conceito de redução do MSE conforme descrita acima e aplicando-a na floresta aleatória, é calculada a média com o intuito de obter o valor final

para mensurar a importância daquela variável. Para fins de comparação a importância de cada variável foi normalizada conforme a Equação (2.6).

2.5 Regressão Lasso

A regressão Lasso (operador de seleção e redução mínima absoluta) é um método de redução (shrinkage methods) por regressão linear proposto originalmente em 1986 cuja redescoberta (independente) e popularização foi feita em 1996 por Robert Tibshirani, baseada na normalização L_1. Antes de explicitar como funciona o método, é necessário um *background* suscinto sobre a famosa regressão linear.

A **Regressão Linear** tem sido até hoje uma das mais importantes ferramentas para a realização de predição de diversos fenômenos. Dada uma entrada X = $\{x_1, x_2, ..., x_N\}$, onde x_i é um vetor coluna com p dimensões, e Y = $\{y_1, y_2, ..., y_N\}$ é a saída, com y_i sendo um escalar e N o número de amostras, queremos encontrar o conjunto de escalares W = $\{w_1, w_2, ..., w_N\}$ tais que:

$$f(X) = Y = w_0 + \sum_{i=1}^{N} w_i x_i \tag{2.10}$$

e os valores w_i sejam tais que:

$$W_{regressão-linear} = min_w ||Xw - y||^2 \tag{2.11}$$

O termo w_0 é conhecido como *bias* (**viés** em uma tradução livre) na área de Aprendizado de Máquina. A regressão linear encontra os valores do conjunto W usando o **método dos mínimos quadrados**. A regressão Lasso é semelhante à regressão linear, com a diferença de que há um termo de penalidade, adicionando um certo viés no modelo. Ele pode ser muito útil para fazer com que o modelo se ajuste bem tanto para os dados de treino quanto para os dados de teste.

Matematicamente, baseado no trabalho de [Tibshirani, 1996], a regressão Lasso busca encontrar os valores do conjunto W tais que:

$$W_{lasso} = min_w \frac{1}{2N} ||Xw - y||^2 + \alpha ||w||_1 \tag{2.12}$$

onde α é um parâmetro ajustado manualmente, e $||w||_1$ é a norma L_1, que é dada por:

$$L_1 = \sum_{i=1}^{N} |w_i| \tag{2.13}$$

Esse tipo de regressão pode levar algum coeficiente w_i ser nulo, reduzindo o excesso de ajuste, a seleção de features para avaliação e o combate no viés do modelo por regressão linear. Além disso, em relação ao parâmetro α, [James et al., 2013] declara de

forma implícita os seguintes casos:

1. Se $\alpha = 0$, nenhum parâmetro w_i é eliminado. Então, é uma Regressão Linear.
2. Se α é incrementado positivamente, teoricamente mais coeficientes w_i tornam-se nulos e portanto são eliminados.
3. Se α aumenta, então o viés aumenta.
4. Se α diminui, então a variação aumenta.

As principais vantagens e desvantagens da regressão usando Lasso, de acordo com [Signorino and Kirchner, 2018] estão presentes nas Tabelas (2.7) e (2.8), respectivamente.

Tabela 2.7: Vantagens da regressão Lasso

Vantagens
Reduz o para 0 um bom número de coeficientes.
Fornece variáveis implícitas em cenários com dados esparsos.
Modelo único.

Fonte: [Signorino and Kirchner, 2018].

Tabela 2.8: Desvantagens da regressão Lasso

Desvantagens
Não fornece estimativas de certeza quando executado apenas uma vez.
Em algumas situações pode não ser consistente com os dados.
Dificuldade em diferenciar preditores relevantes e irrelevantes quando as variáveis são altamente correlacionadas.

Fonte: [Signorino and Kirchner, 2018].

Importância de Variável Para Regressão Lasso

Na regressão Lasso chega-se a um conjunto W = $\{w_1, w_2, ..., w_N\}$ de pesos que se ajustam aos dados de treino. Esse conjunto pode ser usado para analisar a importância das variáveis de modo que:

$$IV(w_i) \propto w_i \tag{2.14}$$

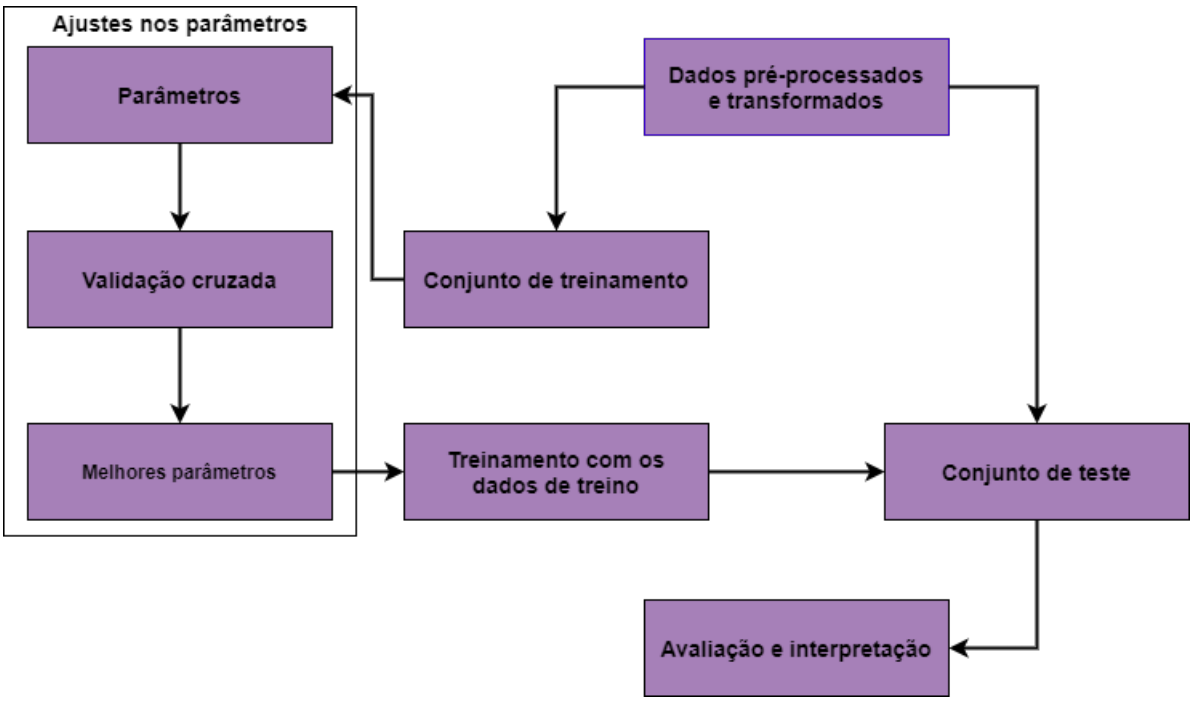
Assim, quanto maior o valor do peso, mais importante é àquela variável para o modelo. Para fins práticos, cada valor do conjunto W foi normalizado de modo que pudesse ser mensurada sua contribuição percentual, semelhante a Equação (2.6).

2.6 Validação Cruzada

Para obter uma boa acurácia nos algoritmos floresta aleatória e Lasso é necessário realizar um ajuste nos parâmetros de número de árvores (ntree ou n_componets na scikit-learn) e coeficiente de ajuste α, respectivamente. Essa etapa é anterior à etapa IV (mineração) e posterior à etapa III (transformação). Além disso, é possível melhorar o problema de overfiting.

Para encontrar os melhores parâmetros, usa-se a validação cruzada. A Figura (2.6) representa o funcionamento desse processo.

Figura 2.6: Processo de escolha dos parâmetros usando Validação Cruzada.



Fonte: Página da biblioteca scikit-learn, adaptado.

O processo descrito na Figura (2.6) pode ser descrito nos seguintes passos:

1. Os parâmetros de ajustes são impostas as condições de início e fim.
2. Usando a validação cruzada (nos k folds) em cada um dos valores, esses parâmetros são testados.
3. Para cada um dos parâmetros usados no passo anterior, são avaliadas as acurácias e é escolhida aquela que possui maior valor.
4. Após encontrar os valores de melhor acurácia, eles são aplicados nos dados após o pré-processamento e a transformação.

Capítulo 3

Metodologia

Nas próximas seções será descrito como foi o procedimento para obtenção das **variáveis de importância**. Todo processamento foi feito usando uma máquina cujas configurações estão presentes na Tabela (3.1). A ordem seguida para análise dos resultados foi a mesma descrita na Seção 2.1.1, com a variação de que entre as etapas de Pré-processamento e Mineração foi realizado o ajuste de parâmetros que possibilitaram as melhores métricas.

Tabela 3.1: Configurações da Máquina usada no Processamento

Configurações
Memória RAM: 8GB
Disco rígido: HDD 1 TB
Processador: Intel Core i5-7200U
Placa de vídeo: AMD Readon R7 M445 2GB GDDR5
Sistema Operacional: Windows 10 PRO

Fonte: Autor.

Devido a versatilidade e facilidade de uso, além da abundante documentação disponível, a linguagem usada foi o Python versão 3.7. Para o Pré-processamento, a Transformação e especialmente os algoritmos de regressão, foi utilizada a seguinte biblioteca do Python: **scikit-learn** [**Buitinck et al., 2013**].

3.1 Seleção dos Dados

O portal do Inep contêm diversas informações sobre a Educação no país, que vão do nível básico ao superior. Nesse contexto, destacam-se os Microdados, que constituem um conjunto de dados recolhidos por pesquisas, avaliações e exames. No caso do Enade, eles são disponbilizados de forma gratuita no site `http://portal.inep.gov.br/web/guest/microdados`, e contém diversas informações, tais como as notas na prova do Enade e um questionário socioeconômico do aluno.

O perfil socioeconômico até o ano de 2013 continha 25 variáveis categóricas, enquanto que os do ano de 2014 a 2018 continham 26. Além disso, a ordem das perguntas até o ano de 2013 diferia naquelas dos anos de 2014 a 2018. Logo, a escolha com melhor *tradeoff* (ordem das perguntas, questionário socioeconômico, atualização dos dados) eram os Microdados de 2014 a 2018.

Para uma análise comparativa dos resultados, também decidiu-se separar em dois grupos:

1. Microdados de Alagoas: contém os dados dos alunos das IES **apenas** do estado de Alagoas que prestaram exame do Enade no intervalo supracitado.

2. Microdados do Brasil: contém os dados dos alunos das IES do demais estados do Brasil, excluindo àqueles que são do item anterior.

Assim, quando for dito que os dados são de Alagoas, deve-se ter em mente a descrição do ponto 1 acima. Do mesmo modo, quando for dito que a análise é feita sobre os dados do Brasil, a referência é a descrição do ponto 2 citado anteriormente.

O arquivo com os Microdados possuem diversas variáveis (numéricas e categóricas). As de interesse são a àquelas que correspondem à nota geral do Enade, nomeada no banco de dados de **NT_GER**, que é um valor que pode variar de 0 (nota mínima) até 100 (nota máxima); e as variáveis que dizem respeito ao questionário socioeconômico que cada estudante deve responder ao realizar o exame. Essas variáveis estão descritas na Tabela (3.2). Para fins de facilidade na elaboração do gráfico com as importâncias de cada variável, escolheu-se algumas palavras-chave para representar a respectiva variável categórica.

Tabela 3.2: Variáveis socioeconômicas e suas respectivas descrições e palavras-chave

Nome	**Descrição**	**Palavras-chave**
QE_01	Estado civil do estudante (5 alternativas)	Estado civil
QE_02	Qual a cor do estudante (6 alternativas)	Cor; raca
QE_03	Nacionalidade do estudante (2 alternativas)	Nacionaldiade
QE_04	Máxima escolaridade do pai (6 alternativas)	Escolarização; pai
QE_05	Máxima escolaridade da mãe (6 alternativas)	Escolarização; mãe
QE_06	Tipo de residência e com quem mora (6 alternativas)	Onde;com quem;mora
QE_07	Quantas pessoas da família moram com o estudante (7 alternativas)	Quem;mora;comigo
QE_08	Renda bruta total da família incluindo rendimentos do estudante (7 alternativas)	Renda;total;família

QE_09	Situação financeira atual do estudante (4 alternativas)	Situação; financeira; atual
QE_10	Atual situação de trabalho do estudante (5 alternativas)	Situação; atual; trabalho
QE_11	Tipo de bolsa de estudos ou financiamento foi recebido para custeio das mensalidades (11 alternativas)	Fonte; bolsa; mensalidade
QE_12	Auxílio permanência recebido ao longo da graduação (6 alternativas)	Aux; permanência
QE_13	Bolsa acadêmica recebida ao longo da graduação (6 alternativas)	Bolsa; acadêmica; graduação
QE_14	Participação de programas e/ou atividades curriculares no exterior (5 alternativas)	Atividade; exterior
QE_15	Se o ingresso do estudante se deu através de alguma política de ação afirmativa ou inclusão social (6 alternativas)	Ingresso; cota
QE_16	Unidade da Federação em que concluiu o Ensino Médio (28 alternativas)	UF; medio
QE_17	Tipo de escola que o estudante cursou o Ensino Médio (6 alternativas)	Tipo; escola; medio
QE_18	Tipo da modalidade do Ensino Médio (5 alternativas)	Modalidade; medio
QE_19	Pessoa que lhe deu maior incentivo concluir a graduação (7 alternativas)	Quem; incentivo; curso
QE_20	Grupo de pessoas determinante para superar as dificuldades da graduação (11 alternativas)	Grupo; força; curso
QE_21	Se pelo menos uma pessoa da família possui curso superior (2 alternativas)	Quem; família; superior
QE_22	Quantos livros foram lidos no ano do Enade, exceto os da bibliografia do curso (5 alternativas)	Quantos; livros; ano
QE_23	Horas de estudo por semana, excluindo as horas de aula (5 alternativas)	Horas; estudo; semana
QE_24	Se o estudante teve oportunidade de aprender um idioma estrangeiro na IES (5 alternativas)	Oportunidade; idioma; estrangeiro
QE_25	Principal motivo para escolha do curso (8 alternativas)	Porque; curso
QE_26	Principal razão para escolha da IES (9 alternativas)	Porque; IES

Fonte: Autor.

A quantidade de dados para cada ano em Alagoas e o Brasil nos anos de 2014 a 2018 estão na Tabela (3.3). Em primeiro lugar foram escolhidos as quantidades do estado de Alagoas para logo depois escolher as demais UFs, pressupondo (de forma razoável e após verificar na base) que o número discentes que realizaram o Enade na região alagoana é bem menor que a soma daqueles que realizaram o exame na época correspondente nos outros estados do país. Nesse contexto, para aqueles que não são da região acadêmica do estado, a coleta foi aleatória, levando em conta sempre a quantidade de alunos presente na base naquele mesmo ano. Então, como em 2014, o número de alunos foi de 7 257, a coleta para as demais as outras UFs foi uma quantidade aleatória de 7257 alunos; e assim sucessivamente até o ano de 2018. Desta forma foram colhidos os dados para cada ano de modo que a comparação entre ambos seja a mais coerente possível.

Tabela 3.3: Quantidade de dados selecionados em Alagoas e no Brasil

Ano	**Quantidade de alunos**
2014	7 257 alunos
2015	5 469 alunos
2016	3 559 alunos
2017	7 084 alunos
2018	5 482 alunos
Total: 28 851 alunos	

Fonte: Autor.

3.2 Pré-processamento

A base escolhida possuía o seguinte formato:

- Colunas: dicionário de variáveis na forma abreviada contendo números e categorias, cujo número para cada ano variou de 137 a 154.
- Linhas: os valores presentes em cada uma das linhas correspondia ao do aluno em cada variável do dicionário.

A variação na quantidade de variáveis do dicionário dizem respeito a incrementações e melhorias que podem ser feitas anualmente, além do que em cada ano também são avaliados um grupo de cursos distintos. Apesar disso, as variáveis analisadas no presente trabalho (conferir Tabela 3.3) sempre estiveram presentes nos anos coletados. Assim, essas colunas foram as escolhidas.

Os campos de avaliação socieconômicos (QE_I01 a QE_I26) não apresentaram valores vazios, ao contrário de alguns da nota geral (campo NT_GER). Este último, que em Python são representados por *Nan*, foram substituídos pela média aritmética.

3.3 Transformação

Como as variáveis independentes são formadas apenas por alternativas, as quais podem ser apenas escolhidas uma por resposta do aluno. Nesse contexto aplicou-se o one-hot encoding. A quantidade de colunas resultante após o uso do algoritmo para os dados em Alagoas e no Brasil estão explicitadas na Tabela (3.4).

Tabela 3.4: Dados antes e depois da aplicação do One-Hot Encoding

	Alagoas	**Brasil**
Antes do One-Hot Encoding	26 colunas	26 colunas
Após o One-Hot Encoding	157 colunas	184 colunas

Fonte: Autor.

A diferença no número de colunas após o uso do one-hot enconding se dá porque nos dados referente a Alagoas, apenas este estado consta como UF, enquanto que nos dados Brasil tem-se os demais estados da UF adicionado de uma categoria **Outro**, caso o candidato não tenha concluído seu ensino médio em qualquer estado do Brasil.

3.4 Escolha dos parâmetros

Com a base de dados dividida em 2/3 para treino e 1/3 para teste, foi usada a validação cruzada por meio da do algoritmo cross_val_score da scikit-learn, aplicados na Árvore de Decisão e na Regressão Lasso. Experimentos similares de ajustes foram testados na Floresta Aleatória, cujos resultados apresentaram melhoria significativa ao avaliar o algoritmo com as métricas e no tempo de execução, razão pela qual foi ajustado. Vale destacar que os mesmos valores foram usados para os dados para Alagoas e o Brasil.

Dois parâmetros em comum foram escolhidos nessa etapa e na de mineração: o random_state (scikit-learn) e o número de folds (cv na scikit-learn). O primeiro foi estabelecido em 42 e o segundo foi fixado em 10.

No algoritmo de Árvore de Decisão [Mantovani et al., 2018] aponta valores nos quais o parâmetro *min_samples_split* varia entre 1 e 20, enquanto que *min_samples_leaf* varia entre 1 e 40. As combinações de valores máximos citados anteriormente foram testados, mas os melhores resultados para as métricas de avaliação foram obtidas com os valores de min_samples_split = 320 e min_samples_leaf = 200. Esses parâmetros correspondem ao número de amostras necessárias para dividir um nó interno (min_samples_split)

e ao número mínimo de amostras necessárias para estarem em uma folha da árvore (min_samples_leaf).

Na Floresta Aleatória foi útil apenas focar na escolha da quantidade de árvores para a construção da floresta. Dado que o objetivo é extrair informações sobre a importância de cada variável categórica, um alto número deve ser escolhido. Nesse contexto, [Grömping, 2009] aponta valores na casa das milhares. Então, a escolha foi fixada para o número de 1000 árvores (parâmetro **n_estimators** na scikit-learn). Enquanto isso, os valores dos parâmetros min_samples_split e min_samples_leaf forma ajustados para 40 e 20, respectivamente.

Para o algoritmo de regressão Lasso, o ajuste do parâmetro α ficou em 0.005. Esse valor trouxe os melhores resultados. Além disso, para facilitar a análise da importância de cada variável, os coeficientes (parâmetro **positive** na scikit-learn) foi ajustado para True, resultando apenas em valores positivos. Todas os ajustes estão reunidos na Tabela (3.5).

Tabela 3.5: Parâmetros dos Algoritmos de Mineração

Parâmetro	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
min_samples_split	320	40	Não possui
min_samples_leaf	200	20	Não possui
n_components	Não possui	1000	Não possui
α	Não possui	Não possui	0.005
random_state	42	42	42

Fonte: Autor.

3.5 Aplicação dos Algoritmos Regressores

Com os parâmetros ajustados pela validação cruzada k-fold (com k = 10), os dados no conjunto de treinamento foram usados nos algoritmos de regressão usando Árvore de Decisão, regressão usando Floresta Aleatória e regressão usando Lasso. Computados os tempos de treino de cada um, os modelos foram testados nos dados de teste, a partir do qual foram extraídas as métricas (MAE, MSE e R^2) e as importâncias de cada variável, sendo comparadas em um gráfico. Tais informações serão abordadas com maiores detalhes no Capítulo 4.

Além disso, para fins de acompanhamento na etapa de escolha dos parâmetros, bem como na de treino e teste são criados arquivos do tipo .csv para registrar as principais informações sobre os algoritmos. Para a validação cruzada k-fold, os arquivos têm os seguintes nomes: RF_CV_AL.csv, DT_CV_AL.csv e LS_CV_AL.csv para os dados de Alagoas; e RF_CV_AL.csv, DT_CV_AL.csv e LS_CV_AL.csv para os dados do Brasil. Para as

técnicas de mineração, as informações são colocadas nos arquivos RF_AL.csv, DT_AL.csv, LS_AL.csv (Alagoas); RF_BR.csv, DT_BR.csv e LS_BR.csv (Brasil).

Capítulo 4

Resultados e Discussão

As próximas seções abordarão os resultados obtidos usando-se os algoritmos de aprendizagem já mencionados anteriormente. O foco será a análise de importância das variáveis, que corresponde a proposta estabelecida inicialmente, isto é, de analisar quais fatores socioeconômicos influenciam na obtenção da nota geral do Enade. A análise será feita com base na resposta de cada algoritmo, realizando as comparações necessárias entre eles a partir dos gráficos e valores dos conjuntos de dados em Alagoas e no Brasil e entre eles, isto é, Alagoas e os demais estados da federação.

4.1 Métricas de Desempenho

Com o modelo treinado a partir dos parâmetros presentes na Tabela (3.5), foram obtidas as métricas cujos valores estão presentes na Tabela (4.1) para Alagoas e na Tabela (4.2) para o Brasil, seguido do respectivo tempo de treinamento.

Tabela 4.1: Métricas em Alagoas

	MAE	**MSE**	R^2	**Tempo de treinamento**
Árvore de Decisão	5.38	76.09	0.0572	0.3 s
Floresta Aleatória	5.21	73.58	0.0883	3 min 59.35 s
Lasso	5.57	75.09	0.0681	0.18 s

Fonte: Autor.

Para os dados em Alagoas, as métricas de desempenho relacionadas ao erro e erro quadrático (MAE e MSE, respectivamente) mostraram valores compatíveis com os valores das notas gerais do Enade, especialmente com a média da classe (uma pontuação equivalente a cerca de 39 pontos). O que apresentou um resultado ligeiramente melhor foi o algoritmo de Floresta Aleatória. Enquanto a Árvore de Decisão e o Lasso tiveram treinamentos dentro de segundos, o terceiro algoritmo precisou de um tempo maior para

Tabela 4.2: Métricas no Brasil (excluindo Alagoas)

	MAE	**MSE**	R^2	**Tempo de treinamento**
Árvore de Decisão	5.20	76.18	0.0431	0.62 s
Floresta Aleatória	5.1	73.06	0.0698	5 min 19.41 s
Lasso	5.32	75.64	0.0499	0.18 s

Fonte: Autor.

treinamento. Isso mostra que o custo de uma Floresta Aleatória em termos de tempo de treinamento foi bem maior sem melhoria significativa para o problema.

Ainda para os dados no estado de Alagoas, o R^2, o MAE e o MSE mostraram que os três algoritmos fazem boas predições compatíveis considerando a base de dados, uma vez que seus valores no conjunto de teste foram maiores que 0. O que apresentou melhor resultado foi o algoritmo de Floresta Aleatória, sendo maior em relação à Árvore de Decisão ao Lasso. Apesar disso, são valores próximos, considerando que ficaram abaixo de 0.1 para o R^2.

Comparando ainda as 3 métricas, é possível notar que tiveram mais similaridade nos resultados os algoritmos da Árvore de Decisão e Lasso. Além do R^2 mostrar que todos eles tiveram os erros em harmonia com a variabilidade dos dados.

Para os dados no Brasil, MAE e MSE tiveram resultados mais próximos entre si e inferiores aos de Alagoas, mas dentro do valor médio para esse conjunto (aproximadamente 45 pontos). O tempo de execução foi similar ao dos dados em Alagoas e como esperado a Floresta Aleatória demandou bem mais tempo para o seu treinamento.

O R^2 score mostrou que todos os algoritmos que a dispersão do erro nas duas classes são semelhantes, especialmente a Árvore de Decisão e o Lasso. A Floresta Aleatória mostrou ser ligeiramente melhor que os outros dois algoritmos, mas a um alto custo no seu treinamento.

As métricas de avaliação em ambas as classes mostram que, conforme [Romero and Ventura, 2013], o uso de algoritmos com dados relacionados a educação é complexo. Isso é evidenciado pelos ajustes nos parâmetros de cada algoritmo que precisaram ser feitos para a obtenção de resultados que garantissem métricas compatíveis com ambos os conjuntos, e isso é importante para que ao extrair a importância relativa das variáveis seja minimamente confiável.

Deve-se notar que todo o corpo de variáveis dependentes são variáveis categóricas binárias (corresponde aos fatores socioeconômicos do aluno) quando se aplica o one-hot encoding, cada alternativa de cada categoria pode assumir valor 1 ou 0. Isso difere muito de outros problemas de regressão, nos quais o corpo de variáveis dependentes podem ser numéricas e categóricas ou apenas numéricas.

4.2 Análise da Importância de Variável

A importância de variável para os fatores socioeconômicos foi calculada de acordo com as técnicas apresentadas no parágrafo da Seção (2.3) para a Árvore de Decisão, no parágrafo da Seção (2.4) para a Floresta Aleatória e no parágrafo da Seção (2.5) para o algoritmo Lasso. Como deseja-se comparar as importância de cada variável, todas elas estão compreendidas na mesma escala, com valores entre 0 e 1, nos quais são multiplicadas por 100 a fim de que seja exibida o seu percentual.

Para a comparação foram plotados dois gráficos de barra, no qual cada um deles exibem os valores de importância relativa. Um deles corresponde aos dados de Alagoas (Figuras 4.1); e o outro corresponde ao dado do Brasil (Figuras 4.2). Ao final do capítulo, na Seção (4.2.4), também estão os valores, em %, correspondente ao impacto de cada um dos 26 fatores socioeconômicos na nota geral do Enade das classes A e B (Tabelas 4.13 e 4.14 respectivamente) para os três algoritmos usados.

4.2.1 Dados em Alagoas

No gráfico da Figura (4.1) destacaram-se nos três algoritmos os fatores referentes a cor do indivíduo, ao tipo de bolsa de estudos ou financiamento para custear (totalmente ou parcialmente) as mensalidades na instituição (no caso do curso não ser gratuito) e a bolsa acadêmica recebida pela IES (PIBICs, PIBITs, Monitorias e semelhantes). A Tabela (4.3) exibe os valores (em %) para estes fatores.

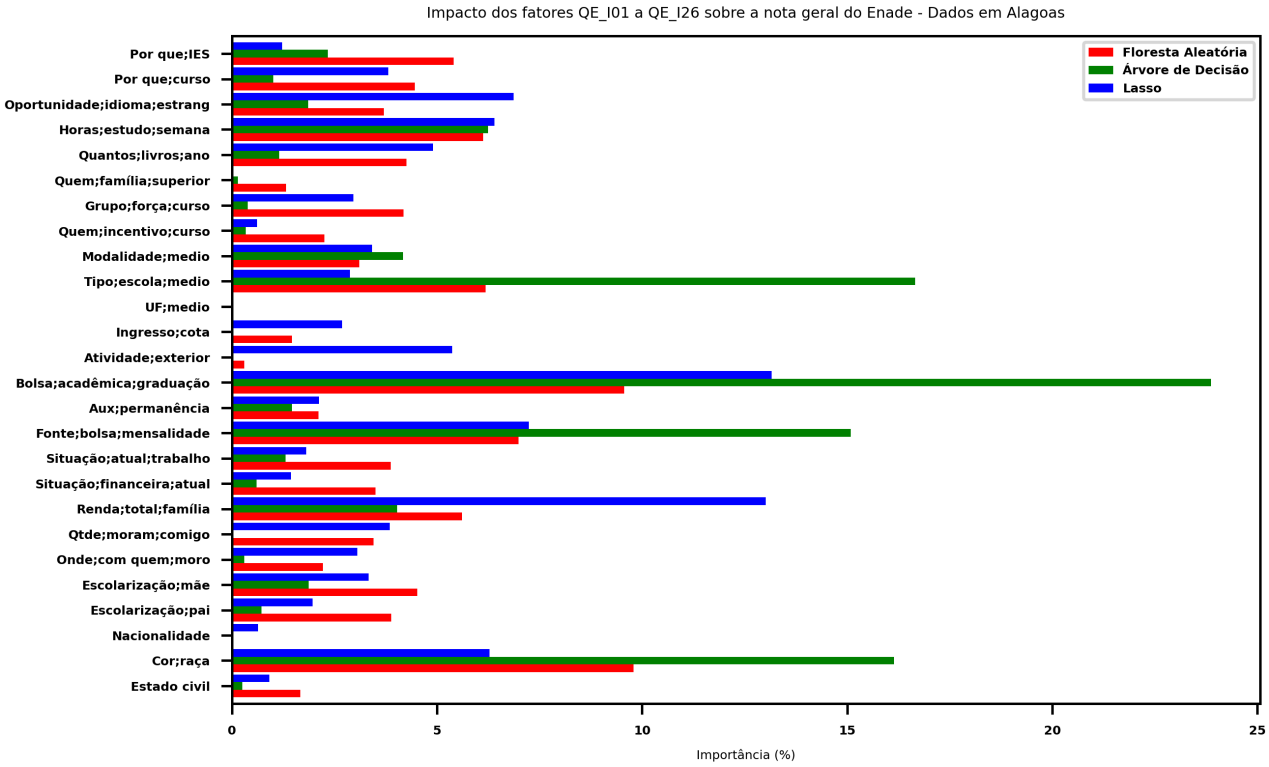
Tabela 4.3: Fatores com alta influência na nota do Enade nos 3 algoritmos: categorias QE_01 a QE_13 de Alagoas

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Qual a cor/raça do estudante (QE_02)	16.14%	9.79%	6.28%
Renda bruta total da família incluindo rendimentos do estudante (QE_08)	4.03%	5.61%	13.01%
Atual situação de trabalho do estudante (QE_11)	15.09%	6.98%	7.24%
Bolsa acadêmica recebida ao longo da graduação (QE_13)	23.87%	9.56%	13.16%

Fonte: Autor.

Também predominou, apesar de apresentar maior diferença na importância relativa para o algoritmo Lasso em relação à Árvore de Decisão e a Floresta Aleatória a

Figura 4.1: Impacto de variável na nota geral do Enade: categorias QE_01 a QE_26 nos dados de Alagoas



Fonte: Autor.

renda bruta familiar do discente (Tabela 4.3). É interessante notar que destes fatores, os três últimos estão diretamente atrelados à questões financeiras do discente ao longo da graduação.

Por outro lado, a influência para todos os três algoritmos nos demais fatores, foram menores ou iguais a cerca de 5%, destacando-se o baixo peso da nacionalidade, que pode ser brasileira, estrangeira ou brasileira naturalizada estrangeira. Também nesse quadro o estado civil teve pouca interferência em todas as três técnicas utilizadas, e também o tipo de auxílio permanência que o indivíduo recebeu ao longo dos seus estudos (Tabela 4.5).

Para as categorias QE_14 a QE_26, exibidas na Figura (4.1), tiveram maior peso em pelo menos dois dos algoritmos, o tipo de escola que a pessoa cursou o ensino médio (todo em escola privada, todo em escola pública, todo no exterior, a maior parte em escola privada, a maior parte em escola pública e parte no Brasil e parte no exterior); o tipo de modalidade do ensino médio (tradicional, técnico profissionalizante, entre outros);e a quantidade de horas de estudo por semana (excluindo as aulas). Para os detalhes de valores, verifique a Tabela (4.4).

Tabela 4.4: Fatores com alta influência na nota do Enade em pelo menos 2 dos 3 algoritmos: categorias QE_14 a QE_26 do dados no Brasil (excluindo Alagoas)

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Tipo de escola que o estudante cursou o Ensino Médio (QE_17)	16.65%	6.19%	2.88%
Tipo da modalidade do Ensino Médio (QE_18)	4.18%	3.11%	3.42%
Horas de estudo por semana, excluindo as horas de aula (QE_23)	6.24%	6.12%	6.40%

Fonte: Autor.

Tabela 4.5: Fatores com baixa influência na nota do Enade nos 3 algoritmos: categorias QE_01 a QE_13 dos dados em Alagoas

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Estado civil do estudante (QE_01)	0.25%	1.67%	0.91%
Nacionalidade do estudante (QE_03)	0.0%	0.0%	0.64%
Auxílio permanência recebido ao longo da graduação (QE_12)	1.47%	2.11%	2.12%

Fonte: Autor.

Tabela 4.6: Fatores com baixa influência na nota do Enade em pelo menos 2 dos 3 algoritmos: categorias QE_14 à QE_26 dos dados em Alagoas

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Se o ingresso do estudante se deu através de alguma política de ação afirmativa ou inclusão social (QE_15)	0.0%	1.46%	2.68%
Unidade da Federação em que concluiu o Ensino Médio (QE_16)	0.0%	0.0%	0.0%
Pessoa que lhe deu maior incentivo concluir a graduação (QE_19)	0.34%	2.26%	0.62%

Se pelo menos uma pessoa da família possui curso superior (QE_21)	0.15%	1.33%	0.02%

Fonte: Autor.

Em contrapartida, tiveram baixa influência na nota do Enade em pelo menos dois algoritmos as seguintes condições socioeconômicas: o tipo de ingresso do estudante se deu através de políticas afirmativas sociais (cotas) e o tipo de cota, em caso de resposta afirmativa; a UF em que completou o ensino médio; a pessoa que mais incentivou o aluno durante as dificuldades na graduação (que vai desde a ausência de alguém até indivíduos fora do círculo familiar); e se alguém da família havia feito algum curso superior. Os valores podem ser consultados na Tabela (4.6).

4.2.2 Dados no Brasil

De modo geral os dados no Brasil para as treze primeiras categorias, mostrados na Figura (4.2), tiveram resultados similares com os de Alagoas. Assim, semelhante Alagoas, tiveram predominância: o tipo de bolsa acadêmica durante a graduação; a fonte de bolsa para custear a mensalidade da IEs (no caso do curso não ser gratuito); e a cor do egresso. Além destes, também sobrepujou em impacto para os três algoritmos em relação a Alagoas e aos demais fatores no Brasil, a renda bruta total da família, enquanto que a importância referente à cor teve redução significativa, especialmente no algoritmo Lasso. Teve também uma certa relevância para dois algoritmos (Floresta Aleatória e Lasso) a escolarização do pai, sobressaindo em relação aos outros fatores. Para consulta os dados comparativos com a importância para esse caso estão presentes na Tabela (4.7).

Tabela 4.7: Fatores com alta influência na nota do Enade em 2 dos 3 algoritmos: categorias QE_01 a QE_13 do dados no Brasil (excluindo Alagoas)

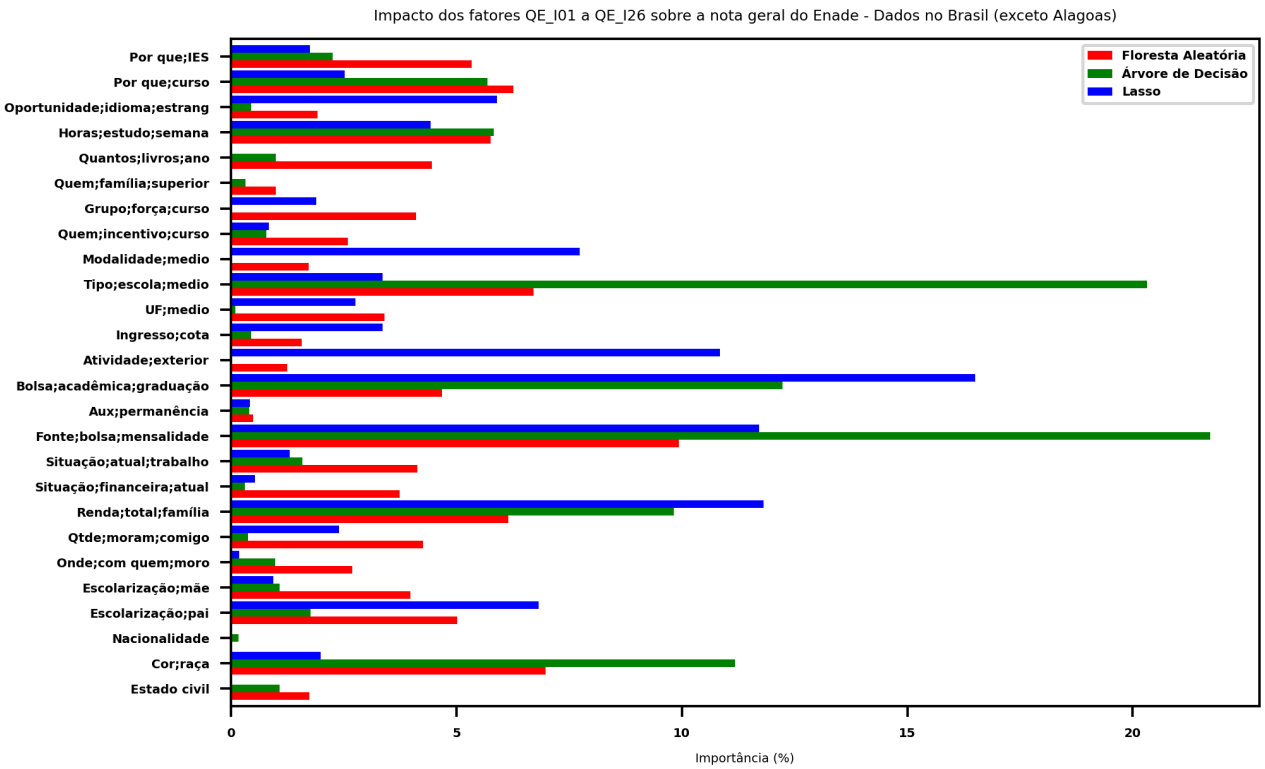
Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Qual a cor/raça do estudante (QE_02)	9.99%	6.98%	1.98%
Renda bruta total da fammília incluindo rendimentos do estudante (QE_08)	8.14%	6.14%	11.82%

Tipo de bolsa de estudos ou fnanciamento foi recebido para custeio das mensalidades (QE_11)	22.89%	9.93%	11.71%
Bolsa acadêmica recebida ao longo da graduação (QE_13)	12.31 %	4.68%	16.51%

Fonte: Autor.

Contudo, outras categorias, como estado civil, nacionalidade, onde e com quem moro além do fator auxílio permanência demonstraram pouco avulto, ficando abaixo dos 5% para os três algoritmos usados. Além destes, o auxílio permanência também se mostrou pouco relevante para obtenção da nota geral do Enade. Os valores, em percentual, das categorias 1 a 13 menor peso conforme descrito anteriormente estão presentes na Tabela (4.8).

Figura 4.2: Impacto de variável na nota geral do Enade: categorias QE_01 a QE_26 do dados no Brasil (excluindo Alagoas)



Fonte: Autor.

Tabela 4.8: Fatores com baixa influência na nota do Enade nos 3 algoritmos: categorias QE_01 a QE_13 do dados no Brasil (excluindo Alagoas)

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Estado civil do estudante (QE_01)	0.17%	1.74%	0.0%
Nacionalidade do estudante (QE_03)	0.0%	0.02%	0.0%
Tipo de residência e com quem mora (QE_06)	0.38%	2.70%	0.19%
Auxílio permanência recebido ao longo da graduação (QE_12)	0.0%	0.49%	0.42%

Fonte: Autor.

Para as categorias 14 a 26, exibidas também na Figura (4.2), os valores de influência sobre a nota ficaram mais distribuídos, com destaque de maior magnitude para o tipo de escola no ensino médio cursado pelo indivíduo e o motivo de escolha do curso. Destacou-se também em relação aos demais, o número de horas de estudo por semana (excluindo as horas em aula), sendo semelhante em importância nos três algoritmos. Note-se também que a influência não foi unânime. Por exemplo, com relação ao fator tipo de escola do médio, a regressão Lasso exibiu um valor próximo a 2.5%. Comportamento semelhante para o mesmo algoritmo ocorreu no motivo de escolha do curso. Os valores de importância para eles estão presentes na Tabela (4.9).

Tabela 4.9: Fatores com alta influência sobre a nota do Enade nos 3 algoritmos: categorias QE_14 a QE_26 do dados no Brasil (excluindo Alagoas)

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Tipo de escola que o estudante cursou o Ensino Médio (QE_17)	20.62%	6.71%	3.36%
Horas de estudo por semana, excluindo as horas de aula (QE_23)	5.35%	5.75%	4.43%
Principal motivo para escolha do curso (QE_25)	6.81%	6.26%	2.53%

Fonte: Autor.

Tabela 4.10: Fatores com baixa influência sobre a noda do Enade nos 3 algoritmos: categorias QE_14 a QE_26 do dados no Brasil (excluindo Alagoas)

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Se o ingresso do estudante se deu através de alguma política de ação afirmativa ou inclusão social (QE_15)	0.0%	1.57%	3.36%
Unidade da Federação em que concluiu o Ensino Médio (QE_16)	0.0%	3.40%	2.75%
Pessoa que lhe deu maior incentivo concluir a graduação (QE_19)	0.95%	2.55%	0.83%
Grupo de pessoas determinante para superar as dificuldades da graduação (QE_20)	0.34%	4.10%	1.89%
Se pelo menos uma pessoa da família possui curso superior (QE_21)	1.47%	1.00%	0.0%

Fonte: Autor.

Apresentaram pouca influência sobre a nota nos três algoritmos os seguintes fatores: se o egresso entrou usando políticas afirmativas e em caso positivo qual foi o tipo de cota; o estado da Federação em que concluiu o ensino médio; a pessoa que mais o incentivou a cursar a graduação; o grupo que deu maior apoio durante o curso; e se alguém da família possuía o superior completo. É possível encontrar os valores com duas casas decimais na Tabela (4.10).

Enquanto isso, apresentaram unanimidade na baixa influência em dois dos três algoritmos, especificamente na Árvore de Decisão e na Floresta Aleatória, os seguintes: se o estudante realizou algum tipo de atividade no exterior; e a modalidade do ensino médio (tradicional, profissionalizante técnico, profissionalizante magistério, EJA e outro); quantos livros leu durante o ano (excluindo os da bibliografia do curso); e o motivo pelo qual escolheu a instituição. Os valores exatos para esse caso estão presentes na Tabela (4.11).

Tabela 4.11: Fatores com baixa influência sobre a noda do Enade em 2 dos 3 algoritmos: categorias QE_14 a QE_26 do dados no Brasil (excluindo Alagoas)

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Participação de programas e/ou atividades curriculares no exterior (QE_14)	1.76%	1.24%	10.84%
Tipo da modalidade do Ensino Médio (QE_18)	0.0%	1.72%	7.72%
Quantos livros foram lidos no ano do Enade, exceto os da bibliografia do curso (QE_22)	1.34%	4.46%	0.0%
Principal razão para escolha da IES (QE_26)	1.94%	5.33%	1.75%

Fonte: Autor.

4.2.3 Disparidades entre Alagoas e Brasil

Após comparar quais fatores mostraram maior e menor impacto na nota geral do Enade, é importante destacar também quais são as discrepâncias mais significativas nesses valores para os dois conjuntos de dados (classes A e B). Logo, foram analisados e comentados quais categorias socioeconômicas tiveram notável distinção. Neste contexto, foram escolhidos aqueles cuja divergência de valores foi notada em pelo menos 2 dos três algoritmos; e que além disso apresentem diferença entre os percentuais para a mesmo fator de pelo menos 25%. Assim, a partir desses pontos de partida, foram encontradas as seguintes diferenças no impacto dos fatores socioeconômicos para os dados em Alagoas e no Brasil:

1. Qual a cor do estudante (QE_02): Em Alagoas, eles apresentaram maiores valores percentuais de importância que nas demais UFs do país em todos os três algoritmos.

2. Tipo de bolsa de estudos ou financiamento foi recebido para custeio das mensalidades (QE_11: Para os 3 algoritmos, os dados em Alagoas mostraram menor impacto na nota que no restante do país.

3. Bolsa acadêmica recebida ao longo da graduação (QE_13): Para os dados do Brasil (exceto Alagoas), os percentuais mostraram menor impacto na Árvore de Decisão e na Floresta Aleatória.

4. Participação de programas e/ou atividades curriculares no exterior (QE_14): Nos 3 algoritmos para os dados do Brasil, os valores percentuais de impacto na nota foram maiores, merecendo destaque para o algoritmo Lasso, cujo percentual foi praticamente o dobro em relação à Alagoas.

5. Unidade da Federação em que concluiu o Ensino Médio (QE_16): Mais uma vez houve disparidade nos valores das classes. Enquanto em Alagoas não houve impacto para os três algoritmos, os dados do Brasil apresentaram valores não nulos para os algoritmos Floresta Aleatória e Lasso.

6. Tipo da modalidade do Ensino Médio (QE_18): Para os dados em Alagoas, os algoritmos da Árvore de Decisão e da Floresta Aleatória mostraram maior impacto que no Brasil; por outro lado, no Lasso ocorreu o inverso, isto é, o impacto na nota para os dados no Brasil foi praticamente o dobro se comparado à Alagoas.

Tabela 4.12: Discrepâncias do impacto dos fatores socioeconômicos dos dados em Alagoas e no Brasil

Fator socioeconômico	**Árvore de Decisão**		**Floresta Aleatória**		**Lasso**	
	AL	**BR**	**AL**	**BR**	**AL**	**BR**
Qual a cor do estudante (QE_02)	16.14%	9.99%	9.79%	6.98%	6.28%	1.98%
Tipo de bolsa de estudos ou fnanciamento foi recebido para custeio das mensalidades (QE_11)	15.09%	22.89%	6.98%	9.93%	7.24%	11.71%
Bolsa acadêmica recebida ao longo da graduação (QE_13)	23.87%	12.31%	9.56%	4.68%	13.16%	16.51%
Participação de programas e/ou atividades curriculares no exterior (QE_14)	0.0%	1.76%	0.31%	1.24%	5.37%	10.85%
Unidade da Federação em que concluiu o Ensino Médio (QE_16)	0.0%	0.0%	0.0%	3.40%	0.0%	2.75%
Tipo da modalidade do Ensino Médio (QE_18)	4.18%	0.0%	3.11%	1.72%	3.42%	7.73%

Fonte: Autor.

Pode-se notar, então que apesar das similaridades no impacto da nota para as classes A e B, há diferenças consideráveis no percentual em 6 dos 26 fatores socioeconômicos. Para detalhes nos valores entre as classes A e B nos três algoritmos usados, veja a Tabela (4.12).

4.2.4 Apêndice

Tabela 4.13: Impacto, em %, dos fatores socioeconômicos sobre a nota geral do Enade para os dados em Alagoas: categorias QE_01 a QE_26

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Estado civil do estudante (QE_01)	0.25%	1.67%	0.91%
Qual a cor/raça do estudante (QE_02)	16.14%	9.79%	6.28%
Nacionalidade do estudante (QE_03)	0.0%	0.0%	0.64%
Máxima escolaridade do pai (QE_04)	0.73%	3.89%	1.97%

Máxima escolaridade da mãe (QE_05)	1.87%	4.52%	3.34%
Tipo de residência e com quem mora (QE_06)	0.30%	2.23%	3.05%
Quantas pessoas da família moram com o estudante (QE_07)	0.0%	3.45%	3.85%
Renda bruta total da família incluindo rendimentos do estudante (QE_08)	4.03%	5.61%	13.01%
Situação financeira atual do estudante (QE_09)	0.61%	3.51%	1.44%
Atual situação de trabalho do estudante (QE_10)	1.31%	3.88%	1.82%
Tipo de bolsa de estudos ou financiamento foi recebido para custeio das mensalidades (QE_11)	15.09%	6.98%	7.24%
Auxílio permanência recebido ao longo da graduação (QE_12)	1.47%	2.11%	2.12%
Bolsa acadêmica recebida ao longo da graduação (QE_13)	23.87%	9.56%	13.16%
Participação de programas e/ou atividades curriculares no exterior (QE_14)	0.0%	0.31%	5.37%
Se o ingresso do estudante se deu através de alguma política de ação afirmativa ou inclusão social (QE_15)	0.0%	1.46%	2.68%

Unidade da Federação em que concluiu o Ensino Médio (QE_16)	0.0%	0.0%	0.0%
Tipo de escola que o estudante cursou o Ensino Médio (QE_17)	16.65%	6.19%	2.88%
Tipo da modalidade do Ensino Médio (QE_18)	4.18%	3.11%	3.42%
Pessoa que lhe deu maior incentivo concluir a graduação (QE_19)	0.34%	2.26%	0.62%
Grupo de pessoas determinante para superar as dificuldades da graduação (QE_20)	0.36%	4.19%	2.96%
Se pelo menos uma pessoa da família possui curso superior (QE_21)	0.15%	1.33%	0.02%
Quantos livros foram lidos no ano do Enade, exceto os da bibliografia do curso (QE_22)	1.15%	4.26%	4.91%
Horas de estudo por semana, excluindo as horas de aula (QE_23)	6.24%	6.12%	6.40%
Se o estudante teve oportunidade de aprender um idioma estrangeiro na IES (QE_24)	1.86%	3.71%	6.86%
Principal motivo para escolha do curso (QE_25)	1.01%	4.46%	3.81%

Principal razão para escolha da IES (QE_26)	2.34%	5.41%	1.22%

Fonte: Autor.

Tabela 4.14: Impacto, em %, dos fatores socioeconômicos sobre a nota geral do Enade para os dados no Brasil: categorias QE_01 a QE_26

Fator socioeconômico	**Árvore de Decisão**	**Floresta Aleatória**	**Lasso**
Estado civil do estudante (QE_01)	0.17%	1.74%	0.0%
Qual a cor/raça do estudante (QE_02)	9.99%	6.98%	1.98%
Nacionalidade do estudante (QE_03)	0.0%	0.02%	0.00%
Máxima escolaridade do pai (QE_04)	2.47%	5.01%	6.83%
Máxima escolaridade da mãe (QE_05)	0.0%	3.98%	0.94%
Tipo de residência e com quem mora (QE_06)	2.70%	2.70%	0.19%
Quantas pessoas da família moram com o estudante (QE_07)	0.97%	4.25%	2.39%
Renda bruta total da família incluindo rendimentos do estudante (QE_08)	8.14%	6.14%	11.82%
Situação financeira atual do estudante (QE_09)	0.40%	3.74%	0.53%
Atual situação de trabalho do estudante (QE_10)	1.69%	4.14%	1.31%

Tipo de bolsa de estudos ou financiamento foi recebido para custeio das mensalidades (QE_11)	22.89%	9.93%	11.71%
Auxílio permanência recebido ao longo da graduação (QE_12)	0.0%	0.49%	0.42%
Bolsa acadêmica recebida ao longo da graduação (QE_13)	12.31%	4.68%	16.51%
Participação de programas e/ou atividades curriculares no exterior (QE_14)	1.76%	1.24%	10.85%
Se o ingresso do estudante se deu através de alguma política de ação afirmativa ou inclusão social (QE_15)	0.0%	1.57%	3.36%
Unidade da Federação em que concluiu o Ensino Médio (QE_16)	0.0%	3.47%	2.75%
Tipo de escola que o estudante cursou o Ensino Médio (QE_17)	20.62%	6.71%	3.36%
Tipo da modalidade do Ensino Médio (QE_18)	0.0%	1.72%	7.73%
Pessoa que lhe deu maior incentivo concluir a graduação (QE_19)	0.95%	2.59%	0.83%
Grupo de pessoas determinante para superar as dificuldades da graduação (QE_20)	0.34%	4.10%	1.89%

Se pelo menos uma pessoa da família possui curso superior (QE_21)	1.47%	1.00%	0.0%
Quantos livros foram lidos no ano do Enade, exceto os da bibliografia do curso (QE_22)	1.34%	4.46%	0.0%
Horas de estudo por semana, excluindo as horas de aula (QE_23)	5.35%	5.75%	4.43%
Se o estudante teve oportunidade de aprender um idioma estrangeiro na IES (QE_24)	0.0%	1.91%	5.90%
Principal motivo para escolha do curso (QE_25)	6.81%	6.26%	2.53%
Principal razão para escolha da IES (QE_26)	1.94%	5.53%	1.75%

Fonte: Autor.

Conclusão

A prova do Enade é aplicada desde 2004 e serve como suporte ao Sistema Nacional de Avaliação da Educação Superior (Sinaes), cujo intuito é avaliar o nível dos cursos de graduação no país. Os resultados são disponibilizados no site oficial do Inep, gerando dados passíveis de análise. Nesse intuito a área de Ciência de Dados e Aprendizagem de Máquina têm muito a oferecer e ajudou a consolidar no fim da última década a mineração de dados educacionais.

Diante disso, por meio do uso de três técnicas de aprendizagem, as quais foram Árvore de Decisão, Floresta Aleatória e Lasso, o presente trabalho extraiu quais fatores socioeconômicos tiveram influência na nota geral do Enade. Após aplicar o processo de KDD, os algoritmos apresentaram resultados de MAE, MSE e R^2 que possibilitam uma análise relativamente segura do modelo.

Apresentaram alto impacto na nota para os dados das classes A e B fatores relacionados à questões financeiras: renda total da família, fonte de bolsa para custear a mensalidade e bolsa acadêmica conseguida pelo aluno durante a sua graduação; condições referentes à oportunidades e escolhas do aluno: tipo de escola cursada no ensino médio, número de horas de estudo durante a semana e a possibilidade de aprender um idioma estrangeiro.

Dentre os fatores que tiveram menor peso na nota do Enade em ambas as classes, merecem destaque: o estado civil do aluno, sua nacionalidade, se ele obteve auxílio permanência cedido pela IES ao longo da graduação, se seu ingresso foi feito por cota e se alguém da família possuía algum curso superior.

É importante salientar também que os dados em Alagoas e no Brasil também tiveram discrepâncias substanciais nos valores percentuais, valendo-se o destaque para: cor, fonte de bolsa da mensalidade para custeio na graduação, bolsa acadêmica na graduação, participação em atividades no exterior, UF em que concluiu o ensino médio e a modalidade do ensino médio. Isto é importante porque aponta numa direção que há uma diferença considerável entre a educação superior em Alagoas e nos demais estados do país.

A aplicação dos algoritmos usados proporcionou uma análise com certa confiança. Contudo, ficam em abertas questões relacionadas a que outras técnicas de aprendizagem de máquina seriam interessante no modelo e que proporcionam melhores resultados. Além disso, uma investigação mais detalhada de como cada fator socioeconômico atua na

nota geral do Enade é um estudo bem interessante, especialmente levando-se em conta que o problema tem aplicações importantes no cenário estadual e nacional.

Discussões sobre quais valores de renda bruta da família são relevantes, ou como a cor do indivíduo se associa na obtenção da nota, ou ainda porque o apoio de pessoas e grupos de pessoas com quem o estudante tem vínculo resulta em baixa influência na sua nota são pertinentes nesse cenário.

Apesar dessas limitações e da complexidade que é tratar sobre a mineração de dados em educação [Romero and Ventura, 2013], é importante destacar que levando em consideração apenas o conjunto de 26 variáveis que descrevem a situação socioeconômica de cada estudante do ensino superior, pode-se notar que há uma rede de fatores que merecem ser levados em consideração para que seja mais provável um bom desempenho acadêmico do aluno e a computação possui ferramentas excelentes para o estudo desse entrelaçamento.

Bibliografia

[Breiman, 2002] Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1:58.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. belmont, ca: Wadsworth. *International Group*, 432:151–166.

[Buitinck et al., 2013] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

[Buskirk, 2018] Buskirk, T. D. (2018). Surveying the forests and sampling the trees: an overview of classification and regression trees and random forests with applications in survey research. *Survey Practice*, 11(1):1–13.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

[Grömping, 2009] Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319.

[Igual and Seguí, 2017] Igual, L. and Seguí, S. (2017). Introduction to data science. In *Introduction to Data Science*, pages 1–4. Springer.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[Mantovani et al., 2018] Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., and de Carvalho, A. C. P. d. L. F. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.

[Romero and Ventura, 2013] Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

[Russell and Norvig, 2004] Russell, S. J. and Norvig, P. (2004). *Inteligência artificial*. Elsevier.

[Signorino and Kirchner, 2018] Signorino, C. S. and Kirchner, A. (2018). Using lasso to model interactions and nonlinearities in survey data. *Survey Practice*, 11(1):2716.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.