Victória Medeiros Barreiros

Gradient Boosting para predição da nota do Enem através de indicadores socioeconômicos

Niterói - RJ, Brasil

13 de dezembro de 2023

Universidade Federal Fluminense

Victória Medeiros Barreiros

Gradient Boosting para predição da nota do Enem através de indicadores socioeconômicos

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Profa. Dra. Karina Yuriko Yaginuma

Niterói - RJ, Brasil

13 de dezembro de 2023

Universidade Federal Fluminense

Victória Medeiros Barreiros

Gradient Boosting para predição da nota do Enem através de indicadores socioeconômicos

Monografia de Projeto Final de Graduação sob o título *"Gradient Boosting para predição da nota do Enem através de indicadores socioeconômicos"*, defendida por Victória Medeiros Barreiros e aprovada em 13 de dezembro de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Karina Yuriko Yaginuma
Departamento de Estatística – UFF

Prof. Dr. Jaime Antonio Ultria Valdes
Departamento de Estatística – UFF

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Niterói, 13 de dezembro de 2023

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

B271g Barreiros, Victória Medeiros
Gradient Boosting para predição da nota do Enem através de indicadores socioeconômicos / Victória Medeiros Barreiros. - 2023.
68 f.

Orientador: Karina Yuriko Yaginuma.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Gradient Boosting. 2. Enem. 3. Árvores de Decisão. 4. Aprendizado de Máquinas. 5. Produção intelectual. I. Yaginuma, Karina Yuriko, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

Neste trabalho, utilizamos o método de *Gradient Boosting*, uma técnica que melhora o desempenho de modelos de Árvores de Decisão, para estimar as notas dos estudantes no Exame Nacional do Ensino Médio (ENEM) através de variáveis socioeconômicas. Um dos objetivos é identificar as variáveis mais influentes na predição das notas, além de compreender o impacto de diferentes fatores no desempenho dos alunos. Nos modelos de regressão, destinados a estimarem as notas para cada uma das áreas, observamos que o modelo de *Gradient Boosting* não apresentou resultados satisfatórios. Os coeficientes de determinação (R^2) para Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação foram, respectivamente, 0.29, 0.23, 0.29, 0.35 e 0.29. Contudo, os modelos de classificação, destinados a prever a aprovação ou reprovação em cursos específicos da UFF e UFRJ, os resultados indicam que os modelos desenvolvidos para cada área apresentaram um bom ajuste aos dados. Para o curso de Estatística na UFF, alcançamos uma acurácia de 0.72, uma sensitividade de 0.70 e uma especificidade de 0.79. Já para o modelo de Estatística na UFRJ, os resultados foram uma acurácia de 0.74, uma sensitividade de 0.74 e uma especificidade de 0.77. O modelo de regressão não apresentou boas estimativas com modelo de *Gradient Boosting*, sugerindo que somente dados socioeconômicos não são suficientes para predizer as notas. No entanto, no modelo de classificação os resultados sugerem que os modelos de *Gradient Boosting* foram capazes de fornecer boas estimativas a aprovação ou não em diferentes cursos.

Palavras-chave: Árvores de Decisão. *Gradient Boosting*. Enem. Aprendizado de Máquinas.

Dedicatória

Dedico este trabalho à minha mãe, ela sempre foi meu apoio.

Agradecimentos

Agradeço ao meu irmão por fazer o que eu pedia, mesmo que às vezes reclamando, enquanto eu estava concentrada neste trabalho.

Agradeço à minha amiga, Isabela, por sua constante presença e apoio. Que mesmo à distância, esteve ao meu lado, compartilhando alegrias e enfrentando desafios comigo.

Agradeço ao meu namorado e companheiro, Gabriel, por estar sempre ao meu lado, me apoiando e ajudando de todas as formas possíveis. Sua presença e suporte foram fundamentais, e sou profundamente grata por tê-lo em minha vida.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O Exame Nacional do Ensino Médio (ENEM) é um dos principais vestibulares do Brasil, criado em 1998 pelo Ministério da Educação (MEC) com o objetivo de avaliar a qualidade do ensino médio no país e auxiliar no acesso ao ensino superior. Também proporcionou que as universidades passassem a adotar a nota do exame como uma forma de ingresso em seus cursos, ao invés de aplicarem um vestibular próprio. Isso trouxe maior uniformidade e padronização ao processo seletivo, e possibilitar que os estudantes pudessem prestar vestibular em várias instituições com uma única prova.

E além disso, tornou a entrada na universidade mais acessível para os estudantes de baixa renda, pois estes muitas vezes não tinham condições de pagar as taxas dos vestibulares de diversas universidades, e também possibilitou o acesso ao ensino superior para pessoas que vivem em áreas remotas do país, já que é realizado em muitos municípios brasileiros.

De acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), o Enem é uma prova de múltipla escolha com 180 questões que avalia habilidades e competências em quatro áreas: Linguagens, Códigos e suas Tecnologias, Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias e Matemática e suas Tecnologias. Além disso, há uma redação que avalia a capacidade de argumentação e escrita dos candidatos.

A nota é calculada a partir do desempenho do participante em cada uma das quatro áreas de conhecimento (Linguagens, Matemática, Ciências Humanas e Ciências da Natureza) e na redação. Cada área tem peso diferente, e cada instituição pode adotar um método diferente para a utilização da nota. Algumas universidades, por exemplo, adotam uma nota mínima para cada área de conhecimento, enquanto outras utilizam a nota total do exame. Além disso, a nota de corte para um mesmo curso pode variar de uma instituição para outra, dependendo da concorrência e da oferta de vagas. Na Figura 1 abaixo é possível visualizar como a Universidade Federal Fluminenses e Universidade

Federal do Rio de Janeiro utilizam pesos diferentes para a entrada no curso de estatística através da nota do Enem.

Áreas de conhecimento	UFF	UFRJ
Redação	2	3
Ciências da Natureza e suas Tecnologias	1	3
Ciências Humanas e suas Tecnologias	1	1
Linguagens, Códigos e suas Tecnologias	2	2
Matemática e suas Tecnologias	4	5

Figura 1: Peso notas do Enem para o curso de Estatística na Universidade Federal Fluminenses e Universidade Federal do Rio de Janeiro

O Enem foi desenvolvido com base em estudos e pesquisas pedagógicas, sendo aprimorado ao longo dos anos para se tornar uma ferramenta cada vez mais eficaz para medir o desempenho dos estudantes e fornecer informações relevantes para o MEC. O exame é aplicado anualmente, com a participação de milhões de estudantes em todo o país.

A importância do Enem vai além da seleção para o ensino superior. Segundo o MEC, o exame também é utilizado como ferramenta de avaliação e melhoria da qualidade do ensino médio, com base nas informações obtidas a partir das notas e desempenhos dos estudantes. Além disso, o Enem também é usado para a obtenção de bolsas de estudos em instituições de ensino superior, como o Programa Universidade para Todos (ProUni), e para ingresso em programas governamentais de financiamento estudantil, como o Fundo de Financiamento Estudantil (FIES).

Um estudo realizado pelo INEP em 2020 apontou que o Enem é uma das principais formas de acesso ao ensino superior no Brasil e além disso é um critério para a seleção de estudantes em instituições públicas de ensino superior em todo o país. De acordo com o estudo, 92% dos estudantes que ingressaram em cursos de graduação em universidades públicas em 2018 foram selecionados por meio do Enem. Além disso, o exame também tem sido utilizado como critério para seleção em programas de intercâmbio estudantil e para obtenção de certificação de conclusão do ensino médio.

As informações trazidas pelo Enem são bastante úteis para o MEC e para as instituições de ensino superior. A partir das notas e desempenhos dos estudantes, é possível avaliar a qualidade do ensino médio e identificar as áreas em que é preciso investir mais

para melhorar a educação no país. Além disso, as notas dos estudantes também são utilizadas para o cálculo de índices de qualidade de ensino, como o Índice de Desenvolvimento da Educação Básica (IDEB).

Em resumo, o Enem é uma ferramenta crucial para avaliar a qualidade do ensino médio no Brasil e para proporcionar o acesso ao ensino superior. Além disso, as informações coletadas a partir do exame são úteis para melhorar a educação no país e desenvolver políticas públicas na área. Uma das principais contribuições do Enem foi centralizar o processo seletivo nas universidades brasileiras, tornando-o mais democrático e acessível, garantindo ainda maior padronização e transparência na seleção.

No Capítulo 2, de Materiais e Métodos, serão apresentados os conceitos dos modelos e métodos usados neste trabalho, bem como uma descrição da base e do conjunto de dados utilizado. Já no Capítulo 3, de Análise dos Resultados, serão apresentadas todas as análises e resultados obtidos e cada um dos modelos. Por fim, no Capítulo 4 de Conclusão, serão reforçados todos os resultados encontrados e como o objetivo central deste trabalho foi alcançado.

1.1 Motivação

Segundo os autores (SILVEIRA; BARBOSA; SILVA, 2015), a análise do nível socioeconômico pode contribuir para a compreensão das desigualdades educacionais no país e para o desenvolvimento de políticas públicas mais efetivas no âmbito da educação.

Na Figura 2 é retratada uma Charge de Luiz Fernando Cazo, que mostra um menino, morador da periferia, tendo dificuldades para estudar por não ter acesso à internet. Ilustrando o quanto o nível socioecônomico pode impactar no acesso à educação.

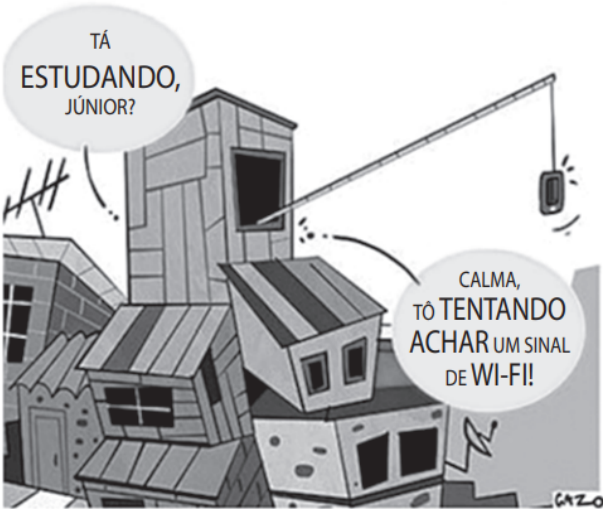

Figura 2: CAZO. Retirado da prova do Enem 2022.

Estudos comprovam que diferenças sociais têm sido associadas a diferenças de desempenho em testes educacionais de larga escala (JALOTO; PRIMI, 2021). Visto que esses fatores podem influenciar significativamente o desempenho dos alunos na prova. O objetivo deste trabalho é desenvolver um modelo de predição para a nota do Enem com base em indicadores socioeconômicos usando *Gradient Boosting*. Para isso, serão utilizados dados disponibilizados pelo INEP, referentes a alunos que prestaram o Enem no ano de 2021. A partir desses dados, serão identificadas as variáveis socioeconômicas mais relevantes para a predição da nota do Enem, e serão desenvolvidos modelos preditivos utilizando técnicas de *Gradient Boosting*.(HAN MICHELINE KAMBER, 2012) Os dados utilizados foram coletados através dos Microdados do Enem 2021, que são disponibilizados pelo INEP e incluem informações específicas sobre as provas, gabaritos, itens, notas e o questionário socioeconômico respondido pelos inscritos no Enem. Para realizar a análise dos dados, foi necessário realizar um pré-processamento, onde foram selecionadas somente as colunas de interesse referentes aos dados dos participantes, dados da escola, dados da prova objetiva, dados da prova de redação e os dados do questionário socioeconômico. Somente alunos que estudaram no estado do Rio de Janeiro e responderam todas as perguntas socioeconômicas foram considerados, reduzindo o número de observações para 19.835.

1.2 Objetivos

1.2.1 Objetivo Principal

O principal objetivo deste trabalho é desenvolver um modelo de predição que seja capaz de estimar a nota dos estudantes no Exame Nacional do Ensino Médio (Enem) e desenvolver um modelo de classificação para a predição da aprovação ou não de estudantes em três cursos específicos - Jornalismo, Enfermagem e Estatística - na Universidade Federal Fluminense e na Universidade do Rio de Janeiro, baseado nas variáveis socioeconômicas, usando o modelo de apendizado de máquinas *Gradient Boosting*.

1.2.2 Objetivo Específico

Como objetivo específico, destaca-se o estudo aprofundado do método de *Gradient Boosting*, tanto para tarefas de regressão quanto de classificação. Além disso, realizar análise dos dados disponíveis, buscando compreender de que maneira diferentes fatores podem influenciar o desempenho dos alunos no Enem. Aspectos como a instituição de ensino frequentada pelo aluno, seu perfil pessoal e o contexto socioeconômico serão investigados. Através dessa análise, o objetivo é identificar as variáveis mais relevantes para a predição da nota, ou seja, aquelas que exercem maior impacto na determinação do desempenho dos estudantes.

2 Materiais e Métodos

Neste Capítulo, serão apresentados a base de dados utilizada na análise, detalhes sobre o pré-processamento dos dados e as metodologias adotadas neste trabalho.

2.1 Base de Dados

Para a realização da análise, foram utilizados os microdados do Enem[1], disponibilizados para o público anualmente. Os microdados são o menor nível de desagregação de dados recolhidos por meio do exame. Eles atendem a demanda por informações específicas ao disponibilizar as provas, os gabaritos, as informações sobre os itens, as notas e o questionário respondido pelos inscritos. Os dados estão por participante e nunca constaram nos dados divulgados quaisquer variáveis que permitissem a identificação direta do participante.

Para concluir a inscrição no Enem, é necessário preencher o questionário socioeconômico. Por meio das informações coletadas, o MEC busca conhecer melhor o perfil do participante, o que pode auxiliar a aprimorar o exame e traçar o perfil dos jovens que estão concluindo a educação básica. O edital do Enem especifica a obrigatoriedade de preencher corretamente o questionário, conforme trecho abaixo:

> "O participante é responsável por preencher corretamente as informações prestadas no sistema de inscrição, inclusive as relacionadas ao Questionário Socioeconômico, inserir os documentos solicitados e verificar se a inscrição foi concluída com sucesso. Os dados informados no Questionário Socioeconômico e os referentes à situação do ensino médio não poderão ser alterados."[2]

As perguntas do questionário socioeconômico visam alcançar três objetivos principais:

[1] Disponíveis em: ⟨https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem⟩

[2] ⟨https://www.gov.br/pt-br/noticias/educacao-e-pesquisa/2022/04/publicado-edital-do-enem-2022⟩

conhecer os dados socioeconômicos e profissionais do estudante e de sua família; avaliar a opinião do estudante sobre seus estudos no ensino médio; e conhecer algumas de suas opiniões sobre assuntos gerais, interesses e planos para o futuro.

Este estudo utilizou os Microdados do Enem 2021, uma base inicialmente composta por informações de 3.389.832 estudantes e 76 variáveis. Para tentar garantir um grupo mais homogêneo, os dados foram filtrados para considerar apenas estudantes de escolas localizadas no município do Rio de Janeiro/RJ, e que não fizeram a prova com intuito de apenas treinar seus conhecimentos, resultando em 24.163 indivíduos. Adicionalmente, algumas variáveis foram descartadas como cor de prova, informações sobre o local de prova, presença do estudante nas provas e informações sobre localidade. Por fim 45 variáveis de interesse foram mantidas para análise, sendo 25 váriáveis do questionário socioeconômico. Após obter somente as váriaveis de interesse foram retirados todas as linhas que continham NA's restando por fim uma amostra de 19.835 estudantes.

No Apêndice 1 contém as Tabelas com todas os dados disponibilizados pelo Inep, as variáveis descartadas possuem fundo vermelho. Na Tabela 16 podem ser encontradas as variáveis relacionadas ao estudante, na Tabela 17 as variáveis relacionadas a escola, na Tabela 21 podem ser encontradas as váriaveis relacionadas as perguntas socieconômicas e na Tabela 19 e Tabela 20 podem ser encontradas as váriaveis que queremos estimar.

2.2 Metodologia

Com o crescente volume de dados, a necessidade de métodos automatizados para análise de dados aumenta cada vez mais. O objetivo do Aprendizado de Máquinas é desenvolver métodos que possam detectar automaticamente padrões nos dados e, em seguida, usar os padrões descobertos para prever dados futuros ou outros resultados de interesse.(MURPHY, 2012)

Toda a análise será feita no contexto de Aprendizado de Máquinas Supervisionado, ou seja, os dados de treinamento são rotulados com as respostas corretas. Uma prática comum no aprendizado de máquinas é dividir os dados em uma base de treinamento e uma base de teste. A base de treinamento é utilizada para realizar análise e modelagem dos dados, enquanto a base de teste é usada para avaliar quão bem as predições se aproximam dos valores observados. Essa abordagem permite verificar a capacidade do modelo de generalizar para dados não vistos durante o treinamento.

Neste trabalho, será utilizado o método de Aprendizado de Máquinas de *Gradient*

Boosting, uma técnica que visa melhorar o desempenho de modelos de Árvores de Decisão. Portanto, também será apresentada a Metodologia de Árvores de Decisão, pois é a base para o método do *Gradient Boosting*.

Exploraremos o uso do *Gradient Boosting* e Árvores de Decisão para resolver o problema de regressão[3] ao estimar as notas dos estudantes no Exame Nacional do Ensino Médio (ENEM). Exploraremos o uso do *Gradient Boosting* e Árvores de Decisão para resolver o problema de classificação[4] ao desenvolver um modelo de classificação para a predição da aprovação ou não de estudantes em alguns cursos.

2.2.1 Árvores de Decisão

Árvore de decisão é um modelo de aprendizado de máquina que utiliza uma estrutura em forma de árvore para tomar decisões com base em regras lógicas e condições sobre as características dos dados. Na construção de uma árvore de decisão as váriáveis são separadas apenas em "Sim" e "Não".

O método de Ávore de decisão envolve dividir os preditores em regiões distintas com base em critérios específicos. Buscando assim a homogenização dos dados em cada folha da árvore.(PROVOST, 2018)

Árvores de Decisão seguem uma estrutura hierárquica que são compostas por:

- **Raiz**: É o primeiro nó, ou seja, é a primeira divisão dos dados, e a partir da raiz ocorrem sucessivas divisões que direcionam as observações para diferentes ramos, até que as observações alcancem as folhas.
- **Nós**: Representam uma decisão ou divisão com base em um critério específico.
- **Ramos**: Ao fazer uma divisão em um nó, é criado um ramo para cada resultado possível do teste realizado no nó. Ou seja, os ramos são os caminhos resultantes das divisões feitas a partir dos nós.
- **Folhas**: São as saídas ou resultados da árvore, as folhas armazenam os resultados obtidos durante o treinamento do modelo. As folhas representam as classificações ou valores preditos finais para as observações com base nas divisões e critérios de decisão ao longo da árvore.

[3]Problema de Regressão: quando os valores da Variável de Interesse são valores contínuos, o algoritimo criado será chamado de Regressor.

[4]Problema de Classificação: quando os valores da Variável de Interesse são categorias, o objetivo é atribuir uma categoria aos dados com base em suas características.

Neste trabalho será usada Árvore de Regressão e a Árvore de Classificação, ou seja, as saídas (Folhas) da árvore de decisão de regressão são valores numéricos contínuos enquanto que as saídas (Folhas) da árvore de decisão de classificação são probabilidades. A Figura 3 é a ilustração de uma árvore de regressão.

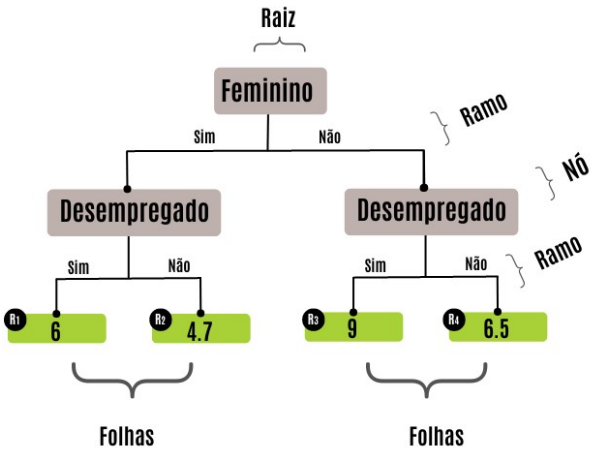

Figura 3: Representação Árvore de Decisão.

Para construir uma árvore de decisão, o algoritmo busca encontrar a sequência de variáveis que minimiza a impureza ou maximize o ganho de informação. Essa medida de impureza ajuda a determinar quais atributos e valores de corte são mais eficazes para separar os dados e tomar as decisões. Ao definir um ponto de corte, a árvore de decisão divide os dados em diferentes regiões.

O processo de divisão de regiões é feito até que uma regra de parada seja aplicada, resultando na criação de k regiões. Para cada região $(R_1, R_2, ..., R_k)$, as constantes $(\hat{y}_1, \hat{y}_2, ..., \hat{y}_k)$ correspondem as estimativas para resposta y, em cada região R_k.

Existem diversas regras de parada, entre elas:

- **Profundidade máxima da árvore**: Define-se um limite para a profundidade da árvore, ou seja, o número máximo de níveis que a árvore pode ter. Quando a profundidade máxima é atingida, as folhas são criadas.
- **Número mínimo de exemplos**: Definir um limite mínimo para o número de exemplos que são necessários para criar um nó. Se o número de exemplos em uma região for menor do que o limite estabelecido, a divisão para essa região é interrompida e as folhas são criadas.
- **Ganho de informação não significativo**: A criação de nós também pode ser interrompida quando o ganho de informação (ou redução do erro, no caso da re-

gressão) não é significativo. Isso significa que, se a divisão de um nó não resultar em uma melhoria significativa na qualidade da árvore (como medido pelo SQR na Equação (2.1) ou outra métrica relevante), a criação de nós é interrompida e as folhas são criadas.

As regras de parada ajudam a evitar o sobreajuste (*overfitting*), garantindo que a árvore não se ajuste excessivamente aos dados de treinamento, mas sim consiga generalizar para novos dados.

Com o objetivo de encontrar o melhor $\hat{y}_k$ (estimativa para o valor da variavél resposta) em sua respectiva região R_k minimiza-se a soma dos resíduos ao quadrado (SQR), definido na Equação (2.1). Para decidir qual variável será a raiz da árvore e os demais nós, são calculados os resíduos.

$$SQR_k = \sum_{i=1}^{N} (y_i - \hat{y}_k)^2, \tag{2.1}$$

Sendo y_i o valor observado e $\hat{y}_i$ o valor previsto.

Obtem-se que a estimativa para $\hat{y}_k$ é a média dos y_i presentes na região R_k, Equação (2.2).

$$\hat{y}_k = \frac{1}{n_k} \sum_{y_i \in R_k} y_i, \tag{2.2}$$

Sendo y_i o valor observado e n_k o número de observações na região k.

Cada novo nó é criado com o objetivo de minimizar a soma dos resíduos ao quadrado, que mede a diferença entre os valores previstos e os valores observados. A criação de novos nós é realizada até que o critério de parada seja atingido.

Com o objetivo de auxiliar na compreesão da construção de uma árvore de decisão será utilizado um exemplo prático.

2.2.1.1 Exemplo de aplicação de Árvores de Decisão para Regressão

Um professor de aprendizado de máquinas visa estimar a nota de seus alunos na disciplina, para isso são utilizados informações dos alunos de sua turma anterior como, Sexo, Tipo de Emprego, e as Notas, os dados estão representados na Tabela 1.

Tabela 1: Informações dos alunos.

Índice	Sexo	Tipo de Emprego	Nota
1	Masculino	CLT	6
2	Feminino	Estágio	4
3	Feminino	Estágio	5
4	Feminino	Desempregado	7
5	Feminino	Desempregado	5
6	Masculino	Estágio	7
7	Masculino	Desempregado	9
8	Feminino	CLT	5

- Critério de Parada

  Usaremos a profundidade máxima da árvore igual a 1 como critério de parada, quando o nível 1 for alcançado teremos as predições finais. A raiz da árvore está no nível 0, e os nós filhos diretos da raiz estão no nível 1.

- Escolhendo a Raiz do Modelo

  Para escolher a Raíz do Modelo é necessário calcular a soma dos resíduos ao quadrado, a váriavel com o menor resíduo tem melhor qualidade de previsão. Na Figura 4, podemos observar como é escolhida a raiz.

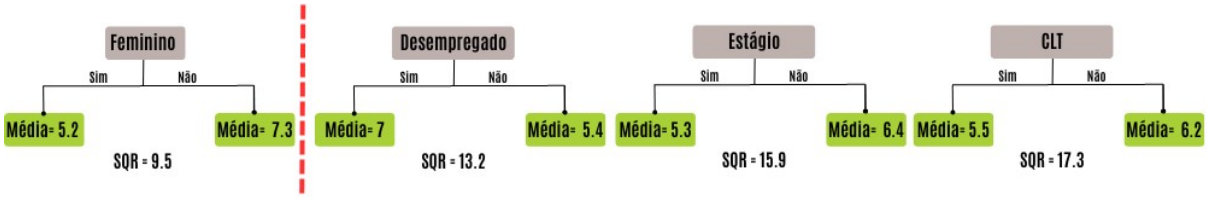

Figura 4: Exemplo da partição do espaço das preditoras para a Raiz.

A média das notas por sexo e por variações do Tipo de Emprego foi calculada e comparada com os valores observados. A variável em que a média mais se aproximou dos valores observados resultou no menor SQR, o que indica melhores previsões. Portanto, com base nessa análise, o sexo seria a melhor escolha para ser a raiz da árvore de decisão, pois apresentou um melhor desempenho na previsão das notas.

- Criação de um Novo nó.

A Figura 5 mostra, as divisões possíveis da váriável Tipo de Emprego e seus respectivos SQR, podemos observar que o menor SQR é entre Desempregado e Não Desempregado, portanto essa é a váriavel do último nó. Apresentando as folhas com as Predições Finais.

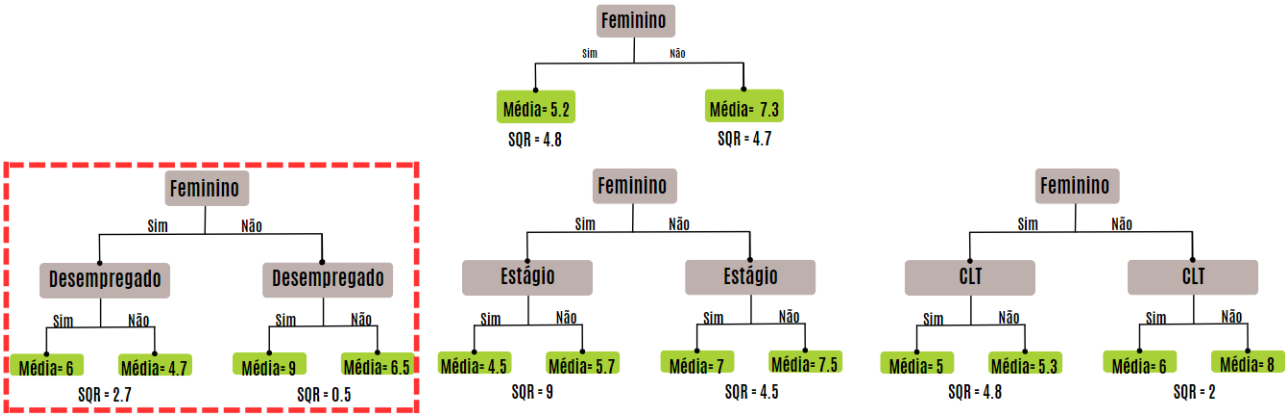

Figura 5: Exemplo da partição do espaço das preditoras para o Nó.

Na Figura 6, temos a árvore com as predições finais em suas determinadas regiões, com critério de parada de 1 nível.

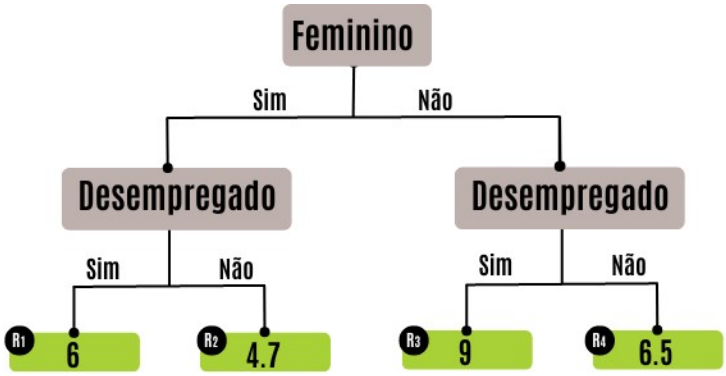

Figura 6: Predições Finais

2.2.2 Gradient Boosting

O *Gradient Boosting* é uma técnica de Apendizado de Máquinas que busca melhorar o desempenho de modelos de Árvores de Decisão, construindo as árvores sequencialmente e utilizando informações das árvores construídas anteriormente para melhorar a estimativa do modelo seguinte. Ou seja, no *Gradient Boosting*, a otimização é realizada ajustando os modelos sucessivos para reduzir o erro residual (a diferença entre as previsões do modelo

e os valores observados). O gradiente é calculado a partir das derivadas da função de perda em relação às previsões atuais do modelo. O processo de ajustar o próximo modelo é guiado pela direção e magnitude desse gradiente, de modo a reduzir o erro residual em cada iteração.

A diferença entre o *Gradient Boosting* e a Árvore de Decisão é que o Boosting não ajusta execessivamente uma única árvore, em vez de ajustar a árvore diretamente nos resultados previstos, ela é ajustada aos resíduos do modelo atual. E assim o modelo é gradualmente aprimorado nas áreas em que não tem uma boa predição. É comum definir a taxa de aprendizado (*learning rate*), de modo que a contribuição de cada árvore seja proporcionalmente ponderada.

O *Gradient Boosting* apresenta um ótimo desempenho e lida bem com o *overfitting*. No entanto, é importante notar que seu processo de treinamento é mais demorado e suas interpretações podem ser complexas devido à combinação de vários modelos e às interações entre eles.

2.2.2.1 Pseudo Algoritimo do Gradient Boosting para Regressão

O *Gradient Boosting* é um algoritmo de aprendizado de máquinas que tem como objetivo aprender um modelo preditivo $F(x)$ capaz de minimizar uma função de perda $L(y, F(x))$. A seguir é mostrado quais são as etapas do método de *Gradient Boosting* em formato de pseudo algoritimo[5].

- Inicialização do modelo

  A inicialização do modelo é dada pela Equação (2.3), responsável por minimizar a função de perda, resultando na constante necessária para iniciar o modelo.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma), \tag{2.3}$$

  sendo,

  y_i = valor observado;

  γ = valor previsto;

  $L(y_i, \gamma)$ = função de perda;

  $\sum_{i=1}^{N} L(y_i, \gamma)$ = soma da função perda para cada valor observado;

[5]Pseudo Algoritimo: Forma de representar o Algoritmo de forma simplificada.

$argmin_\gamma$ = valor previsto que minimiza a soma da função perda.

A otimização do γ é realizada como parte do processo iterativo, em que cada iteração ajusta o modelo atual adicionando um valor corretivo γ_{im} para melhorar as previsões do modelo anterior.

Para problemas de regressão, a função perda é definida pela Equação seguinte.

$$L(y_i, \gamma) = \frac{1}{2}(y_i - \gamma)^2. \tag{2.4}$$

A média dos $y_i's$ é o valor que minimiza a soma da função perda. Como mostrado na Equação (2.5)

$$F_0(x) = \frac{\sum_{i=1}^{N} y_i}{N}. \tag{2.5}$$

– Iterações de $m = 1$ até M

M é o número total de árvores definido pelo usuário e m representa a árvore atual. Geralmente o valor de M é definido como 100.

– Cálculo dos Resíduos

De $i = 1$ até N, sendo N o número de observações, é calculado o resíduo r_{im}, que é a derivada parcial da função de perda L em relação ao modelo $F(x)$ no ponto x_i onde $F(x) = F_{m-1}(x)$.

$$r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\bigg|_{F(x)=F_{m-1}(x)}. \tag{2.6}$$

Como

$$L(y_i, F(x_i)) = \frac{1}{2}(y_i - F(x_i))^2. \tag{2.7}$$

Então

$$r_{im} = -\frac{\partial}{\partial F(x_i)}(\frac{1}{2}(y_i - F(x_i))^2)\bigg|_{F(x)=F_{m-1}(x)}. \tag{2.8}$$

Portanto

$$r_{im} = y_i - F_{m-1}(x_i). \tag{2.9}$$

Essa etapa mede a discrepância entre as previsões do modelo atual e das observações usadas para o treinamento. Resíduos maiores indicam que o modelo precisa melhorar. O sinal negativo indica a direção em que o modelo precisa ser ajustado para reduzir o resíduo.

– Árvore de regressão aos resíduos

Calculando os resíduos, uma nova árvore de regressão é construída a partir dos resíduos. Determinando as regiões terminais R_{jm}, para $j = 1, 2, ..., J_m$.

Sendo m o índice da Árvore com resíduos e j o índice para cada folha da Árvore.

É feita uma árvore de regressão aos resíduos, com o objetivo de encontrar a estrutura da árvore que melhor captura os padrões presentes nos resíduos. A árvore resultante é usada para melhorar o modelo atual.

– Novo valor da função de perda

Para cada região terminal R_{jm}, calculamos o valor ótimo de γ_{jm}, que minimiza a função de perda L quando adicionado ao modelo atual $F_{m-1}(x)$.

A Equação (2.10) fornece um valor corretivo específico para cada região terminal, que será adicionado ao modelo atual.

$$\gamma_{jm} = argmin_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma). \tag{2.10}$$

Para obter γ_{jm}, basta derivar o somatório da função perda com relação a γ e igualar a zero, como apresentado na Equação (2.11)

$$\frac{\partial}{\partial \gamma_{jm}} \left(\sum_{x_i \in R_{jm}} \frac{1}{2} (y_i - (F_{m-1}(x_i) + \gamma_{jm}))^2 \right) = 0. \tag{2.11}$$

– Modelo atualizado

$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} \cdot I(x \in R_{jm}), \tag{2.12}$$

sendo,

$F_m(x)$ = modelo atualizado até a iteração m;

$F_{m-1}(x)$ = modelo da iteração anterior;

γ_{jm} = valor corretivo para a região terminal R_{jm};

v = Taxa de aprendizado, que tem um valor entre 0 e 1, serve para reduzir o efeito que cada árvore tem na previsão final e melhorar a previsão a longo prazo;

$I(x \in R_{jm})$ = função indicadora que retorna 1 se x pertencer a R_{jm} e 0 caso contrário.

Enquanto não forem feitas as M iterações o modelo recomeça pelo Cálculo de Resíduos.

- Predição Final

  A predição final é dada pela soma de todas as predições parciais do modelo:

$$F(x) = \sum_{m=1}^{M} F_m(x), \tag{2.13}$$

  sendo, $F(x)$ a predição final do modelo.

2.2.2.2 Exemplo de aplicação do Gradient Boosting para Regressão

Para auxiliar a compreesão da construção do método de *Gradient Boosting* será retomado o exemplo já apresentado na explicação de Árvores de Decisão. Os dados estão representados na Tabela 1.

- Inicialização do modelo

  O modelo vai ser inicializado por uma constante, dada pela Equação (2.5). Usaremos a média de y_i.

  A primeira estimativa é 6, sendo assim:

$$F_0(x) = 6$$

Tabela 2: Primeira Estimativa.

Índice	Sexo	Tipo de Emprego	Nota	Estimativa $F_0(x)$
1	Masculino	CLT	6	6
2	Feminino	Estágio	4	6
3	Feminino	Estágio	5	6
4	Feminino	Desempregado	7	6
5	Feminino	Desempregado	5	6
6	Masculino	Estágio	7	6
7	Masculino	Desempregado	9	6
8	Feminino	CLT	5	6

- Iterações de m = 1 até M

– Cálculo dos Resíduos

A próxima árvore é constuída a partir dos erros da árvore anterior. Dada pela Equação (2.14), que calcula a diferença entre a nota observada e a predição atual do modelo, ou seja, o resíduo indica o erro de previsão para cada um dos dados.

$$r_{i1} = y_i - F_0(x_i). \tag{2.14}$$

Sendo r_{i1} o resíduo para o i-ésimo aluno na iteração atual, y_i a nota observada do i-ésimo aluno e $F_0(x_i)$ a estimativa atual do modelo para o i-ésimo aluno.

Tabela 3: Resíduos na primeira iteração.

Índice	Sexo	Tipo de Emprego	Nota	Estimativa $F_0(x)$	r_{i1}
1	Masculino	CLT	6	6	0
2	Feminino	Estágio	4	6	-2
3	Feminino	Estágio	5	6	-1
4	Feminino	Desempregado	7	6	1
5	Feminino	Desempregado	5	6	-1
6	Masculino	Estágio	7	6	1
7	Masculino	Desempregado	9	6	3
8	Feminino	CLT	5	6	-1

– Árvore de regressão aos resíduos

Com os resíduos do passo anterior calculados, é construída uma árvore de regressão com o objetivo de estimar os resíduos como mostrado na Figura7.

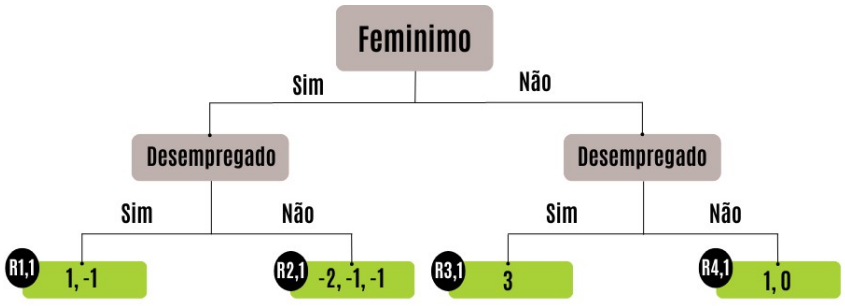

Figura 7: Exemplo da partição do espaço das preditoras.

É ajustada uma nova árvore de regressão aos resíduos atualizados, considerando novamente as variáveis preditoras. É importante ressaltar que as árvores não precisam ter a mesma raiz ou nó em toda iteração do modelo.

– Novo valor da função de perda

A função de perda é utilizada para avaliar a qualidade da previsão. E a função de perda usada popularmente para regressão é dada pela Equação (2.4)

Retomando a Equação (2.10)

O cálculo de $\gamma_{1,1}$ é dado por:

$$\gamma_{11} = argmin_{\gamma} \left[\frac{1}{2}((7 - (6 + \gamma))^2 + (5 - (6 + \gamma))^2) \right]$$

Para encontrar γ_{11}, vamos derivar a função e igualar a a zero.

$$\frac{\partial}{\partial \gamma_{11}} \left[\frac{1}{2}(1 - \gamma_{11})^2 + \frac{1}{2}(-1 - \gamma_{11})^2 \right] = 0$$

E assim obtemos

$$(1 + \gamma_{11}) + (-1 + \gamma_{11})] = 0$$

Sendo $\gamma_{11} = 0$

O cálculo de $\gamma_{2,1}$ é dado por:

$$\gamma_{21} = argmin_{\gamma} \left[\frac{1}{2}((4 - (6 + \gamma))^2 + (5 - (6 + \gamma))^2 + (5 - (6 + \gamma))^2) \right]$$

Seguindo a lógica demonstrada anteriormente obtemos $\gamma_{21} = -1.3$

O cálculo de $\gamma_{3,1}$ é dado por:

$$\gamma_{31} = argmin_{\gamma} \left[\frac{1}{2}((9 - (6 + \gamma))^2) \right]$$

Seguindo a lógica demonstrada anteriormente obtemos $\gamma_{31} = 3$

O cálculo de $\gamma_{4,1}$ é dado por:

$$\gamma_{41} = argmin_{\gamma} \left[\frac{1}{2}((6 - (6 + \gamma))^2 + (7 - (6 + \gamma))^2) \right]$$

Seguindo a lógica demonstrada anteriormente obtemos $\gamma_{41} = 0.5$

Na Figura 8 podemos observar os valores de γ_{i1} que são os valores de saída (R_{i1}) para cada folha.

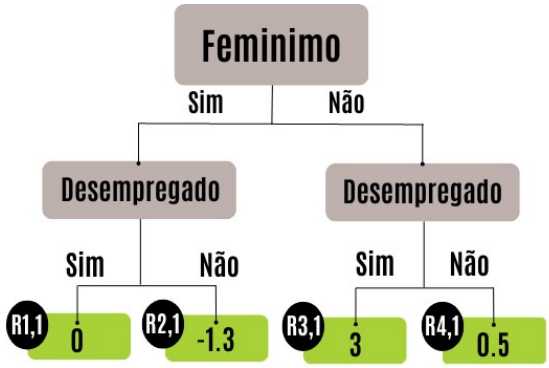

Figura 8: Valores de γ_{i1}

– Modelo atualizado

Retomando a Equação (2.12) e substituindo pelos valores do nosso exemplo: Usaremos $F_1(x)$ para fazer novas previsões.

$$F_1(x) = F_0(x) + v \sum_{j=1}^{J_M} \gamma_{j1} \cdot I(x \in R_{j1})$$

Definiremos v, nossa taxa de aprendizado, como 0.1.

$$F_1(x) = 6 + 0.1 \sum_{j=1}^{J_M} \gamma_{j1} \cdot I(x \in R_{j1})$$

As predições serão atualizadas como ilustrado na Figura 9 resultando na Tabela 4

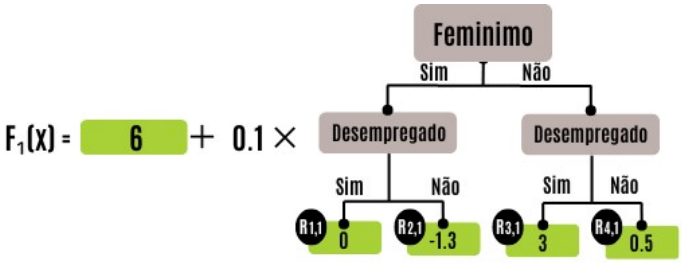

Figura 9: Predições da Árvore 1

Tabela 4: Estimativa $F_1(x)$.

Índice	Sexo	Tipo de Emprego	Nota	Estimativa $F_0(x)$	r_{i1}	Estimativa $F_1(x)$
1	Masculino	CLT	6	6	0	6.1
2	Feminino	Estágio	4	6	-2	5.9
3	Feminino	Estágio	5	6	-1	5.9
4	Feminino	Desempregado	7	6	1	6
5	Feminino	Desempregado	5	6	-1	6
6	Masculino	Estágio	7	6	1	6.1
7	Masculino	Desempregado	9	6	3	6.3
8	Feminino	CLT	5	6	-1	5.9

- Predição Final

  A predição final é composta pela soma dos $F_i(x)$ de 1 até M, como mostrado na Equação (2.13) e é ilustrado na Figura 10

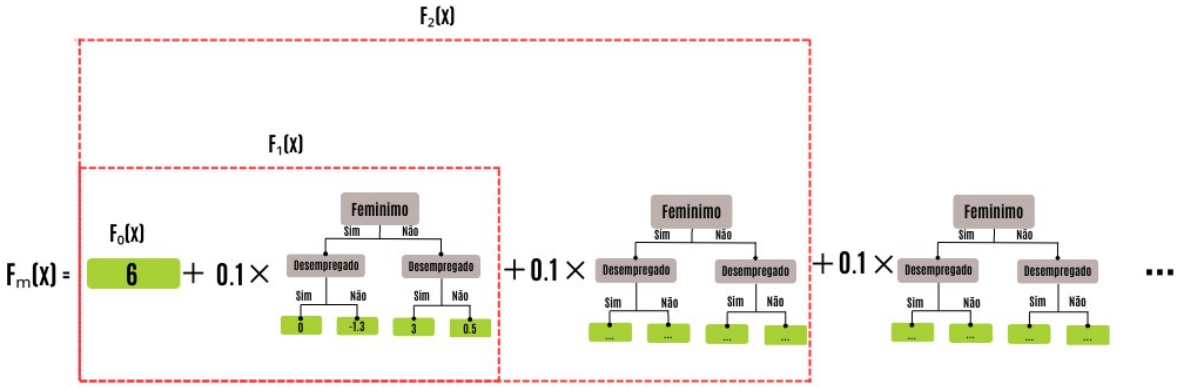

Figura 10: Predições

2.2.2.3 Pseudo Algoritimo do Gradient Boosting para Classificação

Assim como o *Gradient Boosting* de Regressão o de Classificação também tem o objetivo de construir um modelo preditivo $F(x)$ capaz de minimizar uma função de perda $L(y, F(x))$. As etapas dos modelos de *Gradient Boosting* para Classificação e Regressão são as mesmas o que muda são as equações que resultam o valor do modelo, porquê diferente do modelo de regressão que a saída é um resultado numérico, na classificação a saída é a probabilidade de ocorrência do elemento apresentar a caracterıstica de interesse.

O processo de construção do modelo classificação é análogo ao de regressão o que muda

é que a saída do modelo é uma probabilidade (valor entre 0 e 1), e pela saída do modelo ser uma probabilidade a função de inicialização e a função perda também são distintas do modelo de regressão. As equações explicadas abaixo foram baseadas em (SPOLADOR, 2021).

No modelo de classificação a inicialização do modelo é dada pela Equação (2.15), responsável por minimizar a função de perda, resultando na constante necessária para iniciar o modelo.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma), \tag{2.15}$$

sendo,

y_i = valor observado;

$$y_i = \begin{cases} 1, & \text{se o } i\text{-ésimo elemento apresenta a característica de interesse.} \\ 0, & \text{caso contrário.} \end{cases}$$

γ = valor previsto.

E para problemas de classificação, a função perda é definida pela Equação (2.16).

$$L(y_i, \gamma) = -y_i \times \gamma + \ln(1 + e^\gamma)). \tag{2.16}$$

Sendo, $\gamma = \ln(\text{chance}) \leftrightarrow e^\gamma = \text{chance}$.

A chance de ocorrência do evento é definida como:

$$\text{chance} = \frac{p}{1-p}. \tag{2.17}$$

Sendo p o valor previsto, a probabilidade de determinado elemento apresentar a característica de interesse.

Respectivamente, p pode ser escrito como:

$$\text{chance} = \frac{p}{1-p} \quad \Rightarrow \quad p = \frac{\text{chance}}{1+\text{chance}} \quad \Rightarrow \quad p = \frac{e^\gamma}{1+e^\gamma}.$$

A proporção $y_i's$ igual a 1 é o valor que minimiza a soma da função perda. Como

mostrado na Equação (2.18). Obtém-se assim a predição inicial para o ln(chance), que será a folha inicial,

$$F_0(x) = \ln(chance) = \ln\left(\frac{p}{1-p}\right) = \frac{\sum_{i=1}^{N} y_i}{N - \sum_{i=1}^{N} y_i}. \tag{2.18}$$

2.2.3 Métodos de Avaliação do Modelo de Regressão

Serão apresentadas três das principais medidas utilizadas para avaliar a precisão de um modelo de regressão.

2.2.3.1 MSE

O *MSE (Mean Squared Error)*, em português Erro Médio Quadrático, calcula a média dos quadrados das diferenças entre os valores previstos pelo modelo $\hat{y}$ e os valores observados y do conjunto de dados, conforme mostrado na Equação (2.19),

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{2.19}$$

2.2.3.2 RMSE

O *RMSE (Root Mean Squared Error)*, em português Erro Médio Quadrático da Raiz, é obtido a partir do *MSE*, conforme mostrado na Equação (2.20). No RMSE, é calculada a raiz quadrada após ser feita a média dos quadrados das diferenças entre os valores previstos pelo modelo $\hat{y}$ e os valores observados y do conjunto de dados. Uma vantagem do RMSE é que ele mantém a mesma unidade de medida dos valores da variável resposta original, o que facilita a interpretação do erro em relação à escala dos dados originais,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{2.20}$$

2.2.3.3 R^2

O R^2 *(Coefficient of Determination)*, em português Coeficiente de determinação, é uma métrica que varia de 0 a 1 e representa a proporção da variabilidade na variável

dependente que é explicada pelas variáveis independentes. Um R^2 mais próximo de 1 indica que o modelo explica uma grande parte da variação nas notas, enquanto um valor mais próximo de 0 significa que o modelo tem uma explicação limitada para a variação.

No R^2, é feito a soma dos quadrados dos resíduos dividida pela soma total dos quadrados,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}. \tag{2.21}$$

Sendo a soma do quadrados dos resíduos a diferença entre os valores observados y do conjunto de dados e os valores previstos pelo modelo $\hat{y}$,

$$SS_{res} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{2.22}$$

E a soma total dos quadrados a diferença entre a média dos valores observados $\bar{y}$ do conjunto de dados e os valores observados y do conjunto de dados,

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \bar{y}_i)^2. \tag{2.23}$$

2.2.4 Métodos de Avaliação do Modelo de Classificação

Serão apresentadas três das principais métricas utilizadas para avaliar a precisão de um modelo de classificação, a Acurácia, Sensibilidade e Especificidade. Essas métricas variam de 0 a 1 e representam a proporção de previsões corretas do modelo.

Para Avaliar a capacidade preditiva de um Modelo de Classificação é utilizada a matriz de confusão. Ela mostra a relação entre as previsões do modelo e os valores observados nos dados.

Tabela 5: Matriz de Confusão

	Classe Prevista	
Classe Real	Classe Positiva	Classe Negativa
Classe Positiva	VP	FP
Classe Negativa	FN	VN

Sendo,

- Verdadeiro Positivo (VP): Número de observações corretamente classificadas como

positivo;

- Verdadeiro Negativo (VN): Número de observações corretamente classificadas como negativo;
- Falso Positivo (FP): Número de observações erroneamente classificadas como positivo;
- Falso Negativo (FN): Número de observações erroneamente classificadas como negativo.

No modelo a ser apresentado, positivo é caso o estudante tenha sido aprovado no curso ou seja caso a nota de corte seja menor ou igual a nota alcançada pelo aluno, e o resultado negativo é caso o estudante não tenha sido aprovado no curso ou seja caso a nota de corte seja maior que a nota alcançada pelo estudante.

2.2.4.1 Acurácia

A Acurácia, fornece uma medida geral da precisão do modelo em prever corretamente as classes,

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.24}$$

2.2.4.2 Sensitividade

A Sensitividade, avalia a capacidade do modelo em identificar corretamente os casos positivos,

$$Sensitividade = \frac{VP}{VP + FN}. \tag{2.25}$$

2.2.4.3 Especificidade

A Especificidade, mede a capacidade de prever corretamente os casos negativos,

$$Especificidade = \frac{VN}{VN + FP}. \tag{2.26}$$

2.2.4.4 Curva ROC

Os modelos de classificação geram valores no intervalo (0,1), indicando a probabilidade de uma determinada estimativa pertencer à classe de referência. A determinação do ponto de corte ideal para as probabilidades previstas é realizada por meio da curva ROC. Essa curva é uma métrica que permite avaliar o desempenho do modelo em diferentes pontos de corte, permitindo a seleção daquele que melhor atende aos objetivos de classificação.

A Curva ROC (Receiver Operating Characteristic) é usada para comparar testes diagnósticos. Representa graficamente a relação entre a taxa verdadeira positiva e a taxa falsa positiva, como mostrado na Figura 11.

A Curva ROC avalia o desempenho de modelos, permitindo a escolha de pontos de corte ideais para classificação.

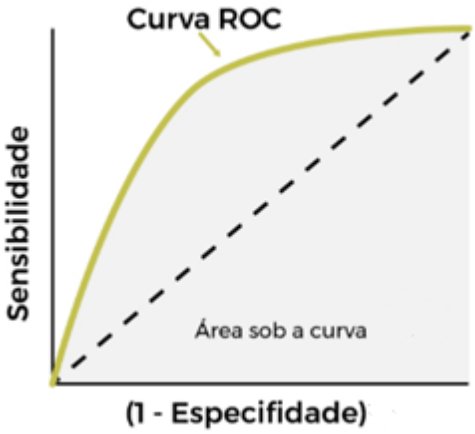

Figura 11: Representação da Curva Roc

É importante frisar que, uma diminuição na sensibilidade geralmente resulta em um aumento na especificidade, e vice-versa. E Quanto mais próximo o gráfico estiver das bordas superior e esquerda, mais preciso é o modelo. A área sob a curva (AUROC) indica a precisão, sendo 1 para um modelo perfeito e 0,5 para um modelo com resultados ruins.

2.2.5 Validação cruzada

Na área de Aprendizado de Máquinas, é crucial avaliar a capacidade de generalização de um modelo, ou seja, sua habilidade de prever corretamente dados não utilizados no treinamento. Para isso, é comum separar os dados em conjuntos de treino e teste. Uma das abordagens mais simples é a divisão 70/30, em que 70% dos dados são usados para treinar

o modelo e os `30%` restantes são utilizados para avaliar o desempenho e a capacidade de generalização do modelo.

No entanto, a validação cruzada fornece uma avaliação mais robusta do desempenho do modelo, pois utiliza múltiplos conjuntos de treino e teste. Dessa forma, ela permite estimar de forma mais precisa a capacidade de generalização do modelo para dados não observados anteriormente. Isso é feito para detectar problemas de sobreajuste, que ocorrem quando um modelo não consegue generalizar padrões além dos dados de treinamento.

2.2.5.1 K-Fold

Existem várias técnicas de validação cruzada, sendo a validação cruzada k-fold uma das mais populares e a que será usada neste trabalho.

Nessa abordagem, os dados são divididos em k subconjuntos, sendo cada subconjunto usado como conjunto de teste uma vez, enquanto os demais subconjuntos são usados para treinamento. Essa técnica é repetida k vezes, garantindo que todos os subconjuntos tenham sido utilizados como conjunto de treino e teste em algum momento.

Essa abordagem envolve a divisão aleatória das observações em k grupos, ou *folds*, de tamanho aproximadamente igual. O primeiro *fold* é tratado como um conjunto de validação, e o método é ajustado nos $k-1$ *folds* restantes. O Erro Médio Quadrático, MSE_1, é então calculado como na Equação (2.19), nas observações do *fold* de validação. Esse procedimento é repetido k vezes, cada vez tratando um grupo diferente de observações como conjunto de validação. Esse processo resulta em k estimativas do erro de teste, MSE_1, MSE_2, ..., MSE_k. (JAMES DANIELA WITTEN, 2013)

A estimativa k-fold é calculada pela média da soma MSE Equação (2.19)

$$\texttt{K-Fold}_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i, \tag{2.27}$$

Quanto menor o resultado, melhor a precisão.

3 Análise dos Resultados

Inicialmente, nosso objetivo era estimar as notas dos estudantes no Enem, abrangendo as previsões nas provas de Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação. No entanto, diante de resultados insatisfatórios, direcionamos a pesquisa para a predição da aprovação ou não de estudantes em três cursos específicos - Jornalismo, Enfermagem e Estatística - na Universidade Federal Fluminense e na Universidade do Rio de Janeiro. A escolha desses cursos foi feita porque cada um deles prioriza uma área de estudo diferente.

A base de dados inicial consistia em informações de 3.389.832 estudantes e 76 variáveis. Após a filtragem dos dados, seleção das variáveis relevantes e remoção das variáveis com variância zero ou próxima de zero, chegamos a um conjunto de variáveis significativas.

Uma das variáveis do questionário socioeconômico, Q025 (Na sua residência tem acesso à Internet?), apresentou uma variância próxima de zero, indicando baixa variabilidade nas respostas. Apesar disso, a exclusão dessa variável resultou em uma pequena piora nos modelos.

Além disso, realizamos o cálculo do coeficiente de contingência modificado para avaliar a relação entre variáveis qualitativas. Apenas as variáveis Q006 e Q018 apresentaram uma correlação superior a 0.5, com um coeficiente de 0.5059964. No entanto, a remoção de uma ou ambas as variáveis resultou em uma piora nos modelos. Levando-nos então a manter todas as variáveis socioeconômicas.

Sendo assim após a seleção de variáveis restaram 19.835 observações e 31 váriaveis. Com o objetivo de entender melhor a base e a distribuição dos dados foi realizada uma análise descritiva inicial. Serão apresentadas as análises descritivas das cinco variáveis mais relevantes para o modelo de classificação e no Apêndice 2, podem ser encontradas as análises das demais variáveis.

Os gráficos a seguir representam a média das notas nas cinco áreas, conforme definido na Equação 3.1, relacionadas a cada uma das variáveis mais influentes nos modelos.

$$\text{Média das Notas} = \frac{\text{Nota em Ciências da Natureza} + \text{Nota em Ciências Humanas}}{5} + \frac{\text{Nota em Linguagens e Códigos} + \text{Nota em Matemática}}{5} + \frac{\text{Nota na Redação}}{5}. \tag{3.1}$$

No gráfico da Questão 01 mostrado na Figura 12, podemos ver que quanto maior a formação do pai, ou o homem responsável maior é a média das notas do estudante.

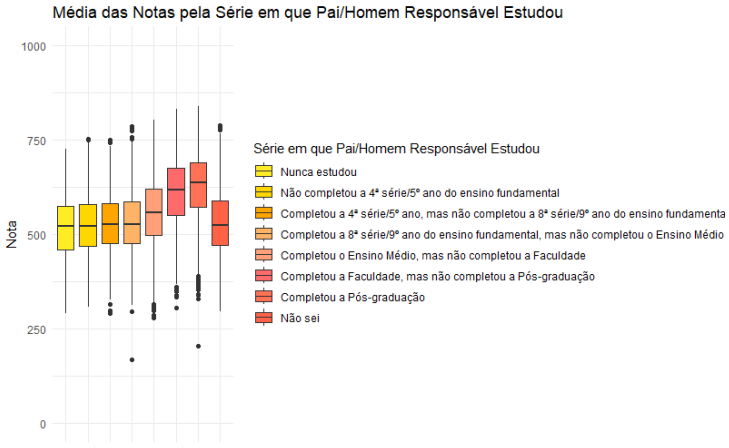

Figura 12: Média da Nota pela Q001 (Até que série seu pai, ou o homem responsável por você, estudou?)

No gráfico da Questão 02 mostrado na Figura 13, podemos ver que quanto maior a formação da mãe, ou a mulher responsável maior é e média das notas do estudante.

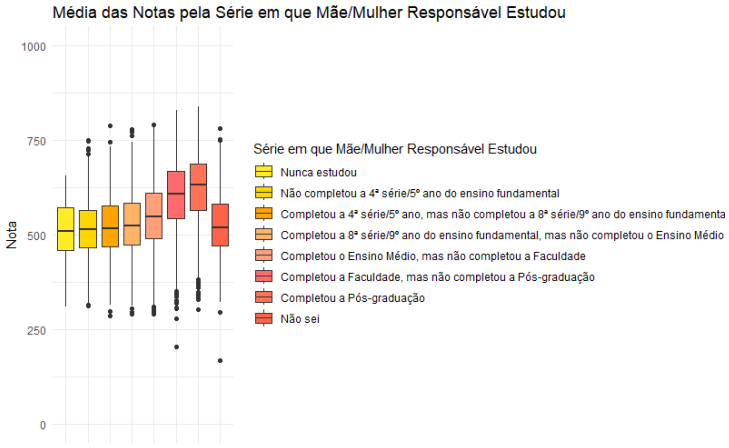

Figura 13: Média da Nota pela Q002 (Até que série sua mãe, ou a mulher responsável por você, estudou?)

No gráfico da Questão 06, apresentado na Figura 14, à medida que a renda familiar aumenta, a média das notas também tende a aumentar. Isso sugere que há uma associação positiva entre a renda familiar dos estudantes e seu desempenho médio nas notas.

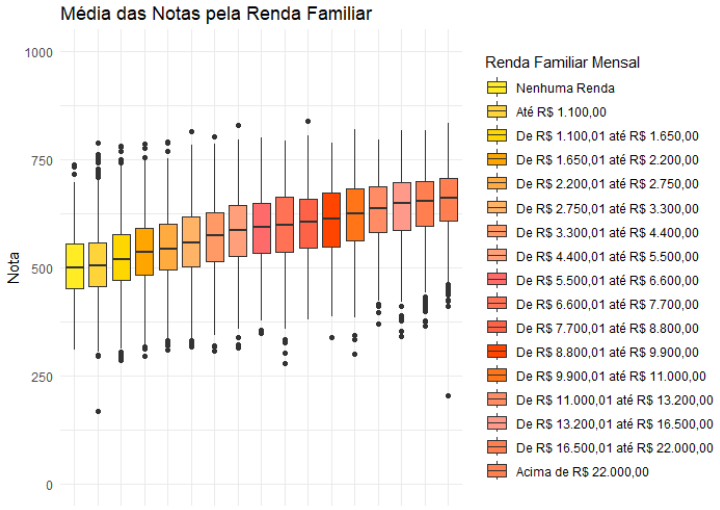

Figura 14: Média da Nota pela Q006 (Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.))

No gráfico da Questão 07, ilustrado na Figura 15, observa-se que a média das notas dos estudantes que não têm empregado doméstico trabalhando em casa é inferior à média daqueles que têm.

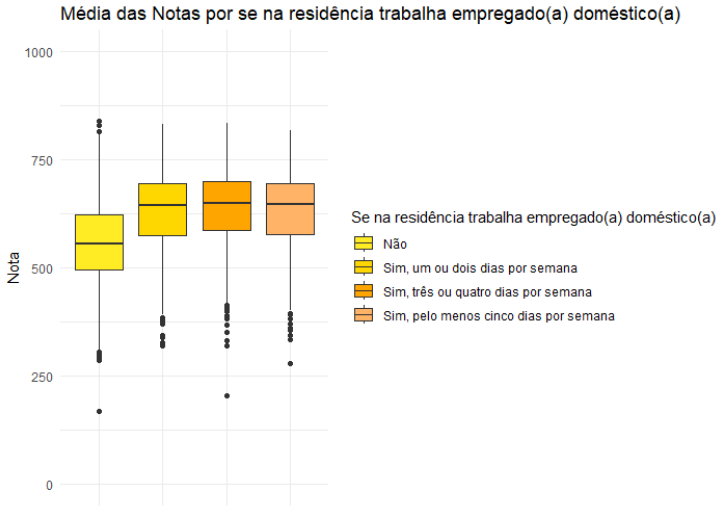

Figura 15: Média da Nota pela Q007 (Em sua residência trabalha empregado(a) doméstico(a)?)

No gráfico da Questão 24, apresentado na Figura 16, observa-se que a média das

notas dos estudantes está positivamente relacionada à quantidade de computadores na residência, indicando que aqueles com um maior número de computadores tendem a obter uma média de notas mais alta.

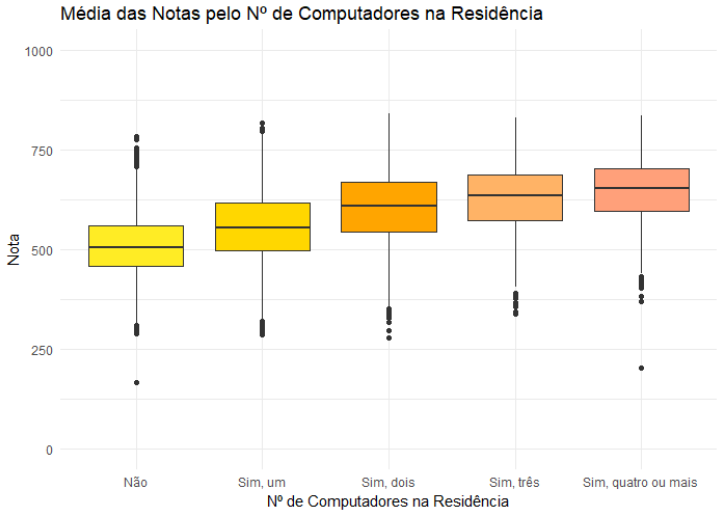

Figura 16: Média da Nota pela Q024 (Na sua residência tem computador?)

Todas as análises presentes neste trabalho foram realizadas no RStudio (versão 4.3.1), utilizando os seguintes pacotes:

- **dplyr:** Empregado na manipulação da base de dados;
- **ggplot2:** Usado para a criação dos gráficos das análises descritivas;
- **caret:** Utilizado para a exclusão de variáveis com variância zero ou próxima de zero, divisão da base em conjuntos de treino e teste, além da avaliação dos modelos;
- **gbm:** Além de implementar o algoritmo Gradient Boosting, fornece análises detalhadas sobre os modelos criados. Também foi empregado na criação das curvas ROC dos modelos de classificação.

3.1 Modelagem e análise dos modelos

Nesta seção, serão explicadas como foi feita a modelagem dos dados e analisaremos o desempenho dos modelos ajustados. A base de dados inicial continha 76 variáveis, incluindo as variáveis de interesse. Para a aplicação do modelo de *Gradient Boosting*, optamos por remover 35 variáveis para aprimorar o desempenho do modelo. A base de dados foi dividida em 70% treino e 30% teste.

No Apêndice 1, podem ser encontradas as tabelas com todos os dados disponibilizados pelo Inep, sendo que as variáveis descartadas estão marcadas com fundo vermelho. As variáveis de interesse são as notas em cada área: NU_NOTA_CN (Nota da prova de Ciências da Natureza), NU_NOTA_CH (Nota da prova de Ciências Humanas), NU_NOTA_LC (Nota da prova de Linguagens e Códigos), NU_NOTA_MT (Nota da prova de Matemática) e NU_NOTA_REDACAO (Nota da prova de redação).

No modelo de regressão, buscamos estimar a nota em cada área, enquanto no modelo de classificação, utilizamos as notas ponderadas em cada área para estimar a nota geral do aluno em relação a cada instituição e curso selecionado. Isso resultou na criação de uma nova variável que determina se o aluno seria aprovado ou não em um curso específico da universidade.

3.1.1 Modelo de Regressão

Usamos o modelo de *Gradient Boosting*, para estimar cada uma das notas dos participantes do Enem. Análises concluíram que as notas não seguem a distribuição normal ou gama. Mesmo após a normalização e a transformação dos dados para seguirem uma dessas distribuições, não obtivemos sucesso.

Para o ajuste do modelo de *Gradient Boosting*, a taxa de aprendizado, que controla a contribuição de cada árvore no modelo, foi de 0.1 (valores menores, como 0.1, geralmente são escolhidos para tornar o modelo mais robusto). Foi usada uma *bag fraction* igual a 1, ou seja, todas as amostras foram usadas em cada árvore, o que é tipicamente usado no *Gradient Boosting*. Também foram utilizados 10 folds para a validação cruzada. E inicialmente, foram utilizadas 3000 árvores para o treino dos modelos.

Após a criação de cada um dos modelos foi identificado o número ótimo de árvores para obter modelos com menores erros de validação.

A Figura 17 mostra o número ótimo de árvores que minimiza o erro quadrático da previsões do modelo de regressão do *Gradient Boosting* para a nota em Ciências da Natureza, resultando no número ideal de 689 árvores. Dado que o método utilizado no *Gradient Boosting* envolve validação cruzada, os dados são particionados em conjuntos de treinamento e teste para o treinamento do modelo. Nesse contexto, a linha verde no gráfico representa o erro quadrático no conjunto de teste, enquanto a linha preta representa o erro quadrático no conjunto de treinamento.

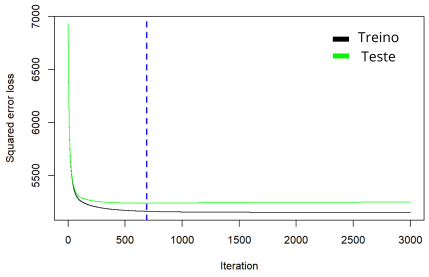

Figura 17: Número ótimo de árvores para o Modelo de Regressão da Nota em Ciências da Natureza

No Apêndice 2 são apresentadas as Figuras com o número ótimo de árvores para as notas em Ciências Humanas, Linguagens e Códigos, Matemática e Redação. Os números ideais de árvores foram encontrados como 607, 897, 593 e 593, respectivamente.

Após a determinação do número ótimo de árvores para cada modelo, os resultados alcançados estão apresentados na Tabela 6, a qual exibe os coeficientes de determinação (R^2) para cada área na amostra teste. Os resultados dos modelos foram considerados insatisfatórios, uma vez que o modelo com maior R^2 conseguiu explicar, apenas, 35% das notas, o modelo com menor $RMSE$ apresentou um número de 64.37 e com menor MAE apresentou um número de 50.28 constatando que existe uma variabilidade nas notas que não está sendo capturada pelo modelo.

Tabela 6: Desempenho do modelo na regressão.

Área	R^2	$RMSE$	MAE
Ciências da Natureza	0.29	71.12	57.36
Ciências Humanas	0.23	83.07	65.93
Linguagens e Códigos	0.24	64.37	50.28
Matemática	0.35	95.57	77.38
Redação	0.27	154.52	114.41

3.1.2 Modelo de Classificação

Usamos o modelo de *Gradient Boosting* para estimar a probabilidade de estudantes ingressarem em três cursos específicos - Jornalismo, Enfermagem e Estatística - na Universidade Federal Fluminense (UFF) e na Universidade do Rio de Janeiro (UFRJ) através do Enem. A partir da estimativa das probabilidades, prevemos a aprovação ou não nos cursos selecionados.

Primeiro foi realizado um pré-processamento nos dados com o objetivo de transformar

as notas de cada área na nota de corte para os cursos selecionados. As notas de corte usadas se referem a modalidade de ampla concorrência.[1] [2] [3]

Foram realizados os cálculos das notas para os cursos selecionados de cada estudante, considerando os pesos atribuídos a cada área do Enem, tanto na UFF quanto na UFRJ. A aplicação desses pesos, conforme especificado nas Tabelas 7, 8, resultou na obtenção das notas. Essas notas foram, então, comparadas com as notas de corte apresentadas na Tabela 9 para determinar a aprovação ou não do estudante.

Tabela 7: Pesos atribuídos a cada área do Enem para os cursos da UFF.

Curso	**Redação**	**Ciências da Natureza**	**Ciências Humanas**	**Linguagens e Códigos**	**Matemática**
Enfermangem	3.00	3.00	2.00	2.00	1.00
Estatística	2.00	1.00	1.00	2.00	4.00
Jornalismo	5.00	2.00	3.00	5.00	2.00

Tabela 8: Pesos atribuídos a cada área do Enem para os curso da UFRJ.

Curso	**Redação**	**Ciências da Natureza**	**Ciências Humanas**	**Linguagens e Códigos**	**Matemática**
Enfermangem	3.00	2.00	2.00	2.00	1.00
Estatística	3.00	3.00	1.00	2.00	5.00
Jornalismo	3.00	1.00	2.00	2.00	1.00

Os dados utilizados para a modelagem e previsão correspondem ao Enem 2021. A Tabela 9 destaca que as notas de corte adotadas referem-se ao Sisu 2021 e foram baseadas nas notas do Enem 2020. Essa escolha visa evitar possíveis viéses na análise, uma vez que as notas de corte pertencem a um ano anterior aos dados modelados.

Tabela 9: Notas de Corte (2020)

Curso	**UFF**	**UFRJ**
Enfermagem	718.42	740.31
Estatística	687.43	753.03
Jornalismo	741.99	765.18

[1] Notas de corte obtidas através do site do MEC: ⟨https://acessounico.mec.gov.br/busca⟩

[2] Cálculos de notas de corte da UFF obtidos através do site:⟨http://www.coseac.uff.br/20211/index.htm⟩

[3] Cálculos de notas de corte da UFRJ obtidos através do site:⟨https://acessograduacao.ufrj.br/2021-1/2021-1-sisu-mec.html⟩

A Tabela 10 apresenta as frequências e percentuais observados de estudantes, conforme a aprovação ou não, na amostra de treino, enquanto a Tabela 11 refere-se à amostra de teste.

Tabela 10: Tabela de Contingência do Treino.

Curso	Universidade	Frequência Aprovados	Frequência Reprovados	Percentual Aprovados	Pencentual Reprovados
Enfermangem	UFF	927	12,959	6.7	93.3
Enfermangem	UFRJ	623	13,262	4.5	95.5
Estatística	UFF	2274	11,611	16.4	83.6
Estatística	UFRJ	590	13,296	4.2	95.8
Jornalismo	UFF	569	13,317	4.1	95.9
Jornalismo	UFRJ	374	13,511	2.07	97.3

Tabela 11: Tabela de Contingência do Teste.

Curso	Universidade	Frequência Aprovados	Frequência Reprovados	Percentual Aprovados	Pencentual Reprovados
Enfermangem	UFF	396	5,553	6.7	93.3
Enfermangem	UFRJ	267	5,683	4.5	95.5
Estatística	UFF	974	4,976	16.4	83.6
Estatística	UFRJ	252	5,697	4.2	95.8
Jornalismo	UFF	243	5,706	4.1	95.9
Jornalismo	UFRJ	160	5,790	2.7	97.3

Após toda a etapa de classificação em aprovados e não aprovados, foram realizados os ajustes dos modelos para cada um dos cursos em cada uma das instituições. Aqui o resultado positivo é caso o estudante tenha sido aprovado no curso ou seja caso a nota de corte seja menor ou igual a nota alcançada pelo aluno, e o resultado negativo é caso o estudante não tenha sido aprovado no curso ou seja caso a nota de corte seja maior que a nota alcançada pelo estudante. A taxa de aprendizado, que controla a contribuição de cada árvore no modelo, foi de 0.1. Foi usada uma *bag fraction* igual a 1, também foram utilizados 10 folds para a validação cruzada, a profundidade máxima de cada árvore foi definida como dois uma vez que só temos duas classificações. E inicialmente, foram utilizadas 1000 árvores para o treino dos modelos.

Como no modelo de regressão, após a criação de cada um dos modelos foi identifi-

cado o número ótimo de árvores para obter modelos otimizados. A Figura 18 mostra o número ótimo de árvores que maximiza as previsões do modelo de classificação do *Gradient Boosting* para a nota em Estatística na UFF, resultando no número ideal de 173 árvores.

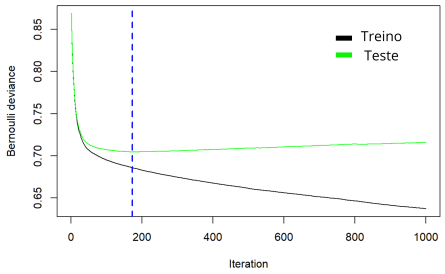

Figura 18: Número ótimo de árvores para o Modelo de Classificação de Estatística na UFF

O eixo y representa a medida de qualidade Bernoulli deviance, a qual indica quão bem o modelo se ajusta aos dados observados em comparação com um modelo nulo que prevê apenas a probabilidade média da classe positiva em todos os casos. Assim, ao longo das iterações do algoritmo de boosting, é desejável observar uma diminuição consistente na deviance, sinalizando uma melhoria contínua do modelo pois o modelo está conseguindo explicar a variabilidade nos dados de forma eficaz, minimizando a perda associada a previsões incorretas.

No Apêndice 2 são apresentadas as Figuras com o número ótimo de árvores para as notas em Enfermagem na UFF, Enfermagem na UFRJ, Estatística UFRJ, Jornalismo UFF e Jornalismo UFRJ. Os números ideais de árvores foram encontrados como 232, 140, 176, 122 e 77, respectivamente.

Após a otimização dos modelos, é conduzida a construção da curva ROC utilizando os dados de teste.A curva ROC possibilita a identificação do ponto de corte ideal para as probabilidades estimadas, onde a classificação do estudante como aprovado ou reprovado se torna mais eficaz.

A Figura 19 exibe a curva ROC do curso de Estatística na UFF, evidenciando um ponto de corte de 0.073. Este ponto de corte foi escolhido para equilibrar a classificação entre aprovação e reprovação do estudante.

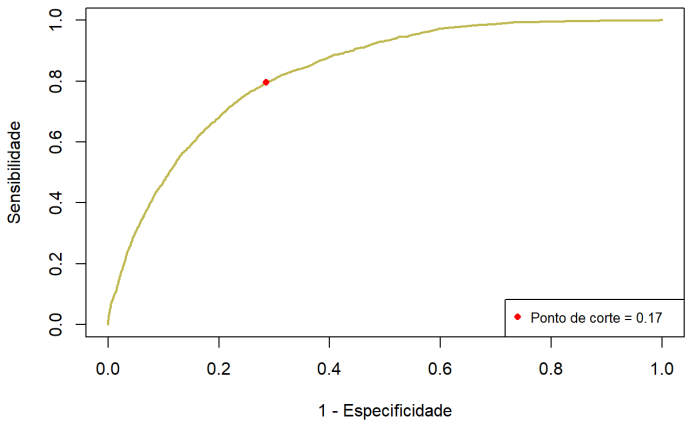

Figura 19: Curva ROC Estatística na UFF

No Apêndice 2 mostram as curvas ROCs para as notas em Enfermagem na UFF, Enfermagem na UFRJ, Estatística UFRJ, Jornalismo UFF e Jornalismo UFRJ. Os pontos de corte foram encontrados como mostrados na Tabela 12.

Tabela 12: Pontos de Corte

Curso/Universidade	Ponto de corte na Curva Roc
Enfermagem/UFF	0.07
Enfermagem/UFRJ	0.05
Estatística/UFF	0.17
Estatística/UFRJ	0.04
Jornalismo/UFF	0.04
Jornalismo/UFRJ	0.03

Após a definição do número ideal de árvores para cada modelo e a escolha do ponto de corte usando a curva ROC, os resultados obtidos são apresentados na Tabela 13, que detalha a Acurácia, Sensibilidade e Especificidade para cada um dos cursos na UFF e na UFRJ na amostra de teste. Os resultados dos modelos demonstraram alto desempenho, sendo capazes de classificar corretamente mais de 70% dos alunos, indicando se serão aprovados ou não, considerando apenas as informações socioeconômicas do estudante.

Tabela 13: Desempenho do modelo de classificação.

Curso/Universidade	Acurácia	Sensibilidade	Especificidade
Enfermagem/UFF	0.74	0.74	0.80
Enfermagem/UFRJ	0.74	0.74	0.80
Estatística/UFF	0.72	0.70	0.79
Estatística/UFRJ	0.74	0.74	0.77
Jornalismo/UFF	0.73	0.73	0.72
Jornalismo/UFRJ	0.74	0.74	0.74

3.2 Análise das Variáveis

Após a construção dos modelos, conseguimos extrair a influência das variáveis em cada um dos modelos. A função usada para calcular a influência relativa está descrito em (FRIEDMAN, 2001).

Nos modelos de regressão as variáveis com maiores influencias relativas são apresentadas na Tabela 14.

Nos modelos de classificação as variáveis com maiores influências relativas na são apresentadas na Tabela 15.

Ao analisar as variáveis mais influentes em cada modelo, é notável o impacto dos fatores socioeconômicos no desempenho dos alunos no Enem. Questões como a escolaridade dos responsáveis (Questões 01 e 02), a renda familiar (Questão 06), a presença de empregados domésticos (Questão 07), e a disponibilidade de recursos como computadores em casa (Questão 24) destacam-se como elementos significativos. A ocorrência repetida das variáveis de maior influência em modelos distintos, tanto de regressão quanto de classificação, indica a presença de padrões que impactam o desempenho e a aprovação dos estudantes.

Tabela 14: Influência relativa das variáveis nos modelos de regressão.

Área	Variável	Influência Relativa
Ciências da Natureza		
	TP_DEPENDENCIA_ADM_ESC	45.463
	Q006	21.102
	Q024	10.575
	Q001	5.657
	TP_SEXO	3.118
Ciências Humanas		
	TP_DEPENDENCIA_ADM_ESC	42.445
	Q006	22.727
	Q024	12.979
	Q001	5.199
	TP_COR_RACA	2.961
Linguagens e Códigos		
	TP_DEPENDENCIA_ADM_ESC	41.848
	Q006	19.715
	Q024	13.961
	Q001	5.104
	TP_COR_RACA	4.359
Matemática		
	TP_DEPENDENCIA_ADM_ESC	36.938
	Q006	27.482
	TP_SEXO	7.580
	Q024	7.205
	Q001	4.838
Redação		
	TP_DEPENDENCIA_ADM_ESC	53.899
	Q006	17.364
	Q024	6.194
	TP_SEXO	5.858
	Q001	2.942

Tabela 15: Influência relativa das variáveis nos modelos de Classificação.

Curso/Instituição	Variável	Influência Relativa
Enfermagem/UFF		
	Q006	45.762
	TP_DEPENDENCIA_ADM_ESC	8.172
	Q024	6.833
	Q007	4.636
	Q002	4.629
Enfermagem/UFRJ		
	Q006	44.434
	Q024	10.066
	TP_DEPENDENCIA_ADM_ESC	6.657
	TP_FAIXA_ETARIA	3.971
	Q001	3.868
Estatística/UFF		
	Q006	38.129
	TP_DEPENDENCIA_ADM_ESC	15.241
	Q024	15.240
	Q001	6.235
	Q002	5.786
Estatítica/UFRJ		
	Q006	46.542
	Q024	6.221
	TP_DEPENDENCIA_ADM_ESC	5.670
	Q001	5.406
	TP_FAIXA_ETARIA	3.888
Jornalismo/UFF		
	Q006	45.727
	Q024	7.095
	Q001	6.536
	Q007	5.666
	TP_DEPENDENCIA_ADM_ESC	4.591
Jornalismo/UFRJ		
	Q006	38.253
	Q024	9.027
	Q001	6.202
	Q007	5.514
	TP_DEPENDENCIA_ADM_ESC	4.625

4 Conclusões

Inicialmente propomos a previsão das notas do Enem utilizando modelos de *Gradient Boosting* para Regressão. Contudo, as primeiras modelagens revelaram que apenas os fatores socioeconômicos não eram capazes de predizer as notas com precisão. Diante desse cenário, adaptamos nossa abordagem para a previsão da aprovação ou reprovação do estudante em cursos específicos pré-selecionados na UFF e na UFRJ. A análise exclusiva do questionário socioeconômico para a estimativa das notas mostrou-se insuficiente para obter resultados aceitáveis. No entanto, ao focarmos na classificação da aprovação do estudante, observamos resultados bastante promissores. Esta mudança de enfoque ressalta a importância de considerar diferentes abordagens na modelagem, indicando que o mesmo conjunto de dados pode apresentar resultados distintos de acordo com a técnica escolhida.

Nos modelos de regressão, destinados a estimarem as notas para cada uma das áreas, observamos que o modelo de *Gradient Boosting* não apresentou resultados satisfatórios. Os coeficientes de determinação (R^2) para Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação foram, respectivamente, 0.29, 0.23, 0.29, 0.35 e 0.29.

Contudo, os modelos de classificação, nos quais buscamos prever a aprovação ou não em cursos específicos da UFF, os resultados indicam que os modelos desenvolvidos para cada área demonstraram bom ajuste aos dados. Um dos modelos que foi feito para o curso de Enfermagem na UFF e UFRJ obtivemos uma acurácia de 0.74, uma sensitividade de 0.74 e uma especificidade de 0.80, para o curso de Estatística na UFF, alcançamos uma acurácia de 0.72, uma sensitividade de 0.70 e uma especificidade de 0.79. Já para o modelo de Estatística na UFRJ, os resultados foram uma acurácia de 0.74, uma sensitividade de 0.74 e uma especificidade de 0.77, para o curso de Jornalismo na UFF obtivemos uma acurácia de 0.73, uma sensitividade de 0.73 e uma especificidade de 0.72 e para o curso de Jornalismo na UFRJ obtivemos uma acurácia de 0.74, uma sensitividade de 0.74 e uma especificidade de 0.74. É importante destacar a disparidade de resultados obtidos pelos modelos de regressão e classificação, ambos construídos a partir da mesma base de

dados. A compreensão da base foi crucial para a definição de novos objetivos, nos quais os modelos fossem capazes de prever de forma mais precisa, resultando em desempenhos mais favoráveis.

Na análise realizada, observou-se que as variáveis socioeconômicas que exerceram maior impacto nos modelos foram: a Questão 01, que indaga sobre o nível de escolaridade do pai ou do responsável masculino; a Questão 02, que aborda o mesmo aspecto em relação à mãe ou à responsável feminina; a Questão 06, relacionada à renda mensal familiar; a Questão 07, que investiga a presença de empregado(a) doméstico(a) na residência; e a Questão 24, referente à disponibilidade de computador no domicílio.

Neste estudo, evidenciamos que o nível socioeconômico do estudante exerce influência em seu desempenho no Exame Nacional do Ensino Médio (Enem). Para trabalhos futuros, sugere-se uma investigação mais abrangente que vá além do contexto socioeconômico. Analisar como diferentes instituições de ensino impactam a preparação dos alunos para o Enem, considerando o acesso a recursos e o ambiente escolar. A inclusão de variáveis como a qualidade do ensino oferecido pelas escolas também pode contribuir significativamente para o aprimoramento do modelo.

Referências

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189 – 1232, 2001. Disponível em: ⟨https://doi.org/10.1214/aos/1013203451⟩.

HAN MICHELINE KAMBER, J. P. J. *Data Mining: Concepts and Techniques*. 3. ed. [S.l.]: Morgan Kaufmann Publishers, 2012. (ITPro collection.; Morgan Kaufmann series in data management systems).

JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no enem. 2021. Disponível em: ⟨http://cadernosdeestudos.inep.gov.br/ojs3/index.php/emaberto/article/view/5002⟩.

JAMES DANIELA WITTEN, T. H. R. T. G. *An Introduction to Statistical Learning with Applications in R*. [S.l.]: Springer, 2013. v. 103. (Springer Texts in Statistics, v. 103).

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. 1. ed. [S.l.]: The MIT Press, 2012.

PROVOST, T. F. F. *Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. [S.l.]: Alta Books Editora, 2018.

SILVEIRA, F. L. da; BARBOSA, M. C. B.; SILVA, R. da. Exame nacional do ensino médio (enem): Uma análise crítica. 2015. Disponível em: ⟨https://www.scielo.br/j/rbef/a/TpSdTxpHR3XBgFttPmgmyPF/?lang=pt#⟩.

SPOLADOR, R. H. Aplicação do método de gradient boosting. 2021. Disponível em: ⟨https://estatistica.uff.br/tcc-2020/⟩.

APÊNDICE 1 – Variáveis Disponibilizadas

O apêndice apresenta os dados recolhidos por meio do Enem e disponibilizados pelo Inep, acompanhadas por seus respectivos nomes e descrições. Os dados contém os gabaritos, as informações sobre os itens, as notas e o questionário respondido pelos inscritos no Enem. Para mais informações como a descrição das categorias em cada variável, os dados utilizados são disponibilizados no site do Inep.[1]

As variáveis que possuem fundo vermelho foram descartadas para a modelagem.

Tabela 16: Variáveis relacionadas ao estudante

Nome da Variável	Descrição
NU_INSCRICAO	Número de inscrição
NU_ANO	Ano do Enem
TP_FAIXA_ETARIA	Faixa etária
TP_SEXO	Sexo
TP_ESTADO_CIVIL	Estado Civil
TP_COR_RACA	Cor/raça
TP_NACIONALIDADE	Nacionalidade
TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio
TP_ANO_CONCLUIU	Ano de Conclusão do Ensino Médio
TP_ESCOLA	Tipo de escola do Ensino Médio
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o Ensino Médio
IN_TREINEIRO	Indica se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos

[1]Disponíveis em: ⟨https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem⟩

Tabela 17: Variáveis relacionadas à escola

Nome da Variável	Descrição
CO_MUNICIPIO_ESC	Código do município da escola
NO_MUNICIPIO_ESC	Nome do município da escola
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)
CO_UF_ESC	Código da Unidade da Federação da escola
SG_UF_ESC	Sigla da Unidade da Federação da escola
TP_LOCALIZACAO_ESC	Localização (Escola)
TP_SIT_FUNC_ESC	Situação de funcionamento (Escola)

Tabela 18: Variáveis relacionadas ao local de aplicação da prova

Nome da Variável	Descrição
CO_MUNICIPIO_PROVA	Código do município da aplicação da prova
NO_MUNICIPIO_PROVA	Nome do município da aplicação da prova
CO_UF_PROVA	Código do município da aplicação da aplicação da prova
SG_UF_PROVA	Sigla da Unidade da Federação da aplicação da prova

Tabela 19: Variáveis relacionadas à prova objetiva

Nome da Variável	Descrição
TP_PRESENCA_CN	Presença na prova objetiva de Ciências da Natureza
TP_PRESENCA_CH	Presença na prova objetiva de Ciências Humanas
TP_PRESENCA_LC	Presença na prova objetiva de Linguagens e Códigos
TP_PRESENCA_MT	Presença na prova objetiva de Matemática
CO_PROVA_CN	Código do tipo de prova de Ciências da Natureza
CO_PROVA_CH	Código do tipo de prova de Ciências Humanas
CO_PROVA_LC	Código do tipo de prova de Linguagens e Códigos
CO_PROVA_MT	Código do tipo de prova de Matemática
NU_NOTA_CN	Nota da prova de Ciências da Natureza
NU_NOTA_CH	Nota da prova de Ciências Humanas
NU_NOTA_LC	Nota da prova de Linguagens e Códigos
NU_NOTA_MT	Nota da prova de Matemática
TX_RESPOSTAS_CN	Vetor com as respostas da parte objetiva da prova de Ciências da Natureza
TX_RESPOSTAS_CH	Vetor com as respostas da parte objetiva da prova de Ciências Humanas
TX_RESPOSTAS_LC	Vetor com as respostas da parte objetiva da prova de Linguagens e Códigos
TX_RESPOSTAS_MT	Vetor com as respostas da parte objetiva da prova de Matemática
TP_LINGUA	Língua Estrangeira
TX_GABARITO_CN	Vetor com o gabarito da parte objetiva da prova de Ciências da Natureza
TX_GABARITO_CH	Vetor com o gabarito da parte objetiva da prova de Ciências Humanas
TX_GABARITO_LC	Vetor com o gabarito da parte objetiva da prova de Linguagens e Códigos
TX_GABARITO_MT	Vetor com o gabarito da parte objetiva da prova de Matemática

Tabela 20: Variáveis relacionadas à redação

Nome da Variável	Descrição
TP_STATUS_REDACAO	Situação da redação do participante
NU_NOTA_COMP1	Nota da competência 1 - Demonstrar domínio da modalidade escrita formal da Língua Portuguesa.
NU_NOTA_COMP2	Nota da competência 2 - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
NU_NOTA_COMP3	Nota da competência 3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
NU_NOTA_COMP4	Nota da competência 4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
NU_NOTA_COMP5	Nota da competência 5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.
NU_NOTA_REDACAO	Nota da prova de redação

Tabela 21: Variáveis relacionadas ao questionário socioeconômico

Nome da Variável	Descrição
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q003	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).
Q004	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).
Q005	Incluindo você, quantas pessoas moram atualmente em sua residência?
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
Q007	Em sua residência trabalha empregado(a) doméstico(a)?
Q008	Na sua residência tem banheiro?
Q009	Na sua residência tem quartos para dormir?
Q010	Na sua residência tem carro?
Q011	Na sua residência tem motocicleta?
Q012	Na sua residência tem geladeira?
Q013	Na sua residência tem freezer (independente ou segunda porta da geladeira)?
Q014	Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)
Q015	Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?
Q016	Na sua residência tem forno micro-ondas?
Q017	Na sua residência tem máquina de lavar louça?
Q018	Na sua residência tem aspirador de pó?
Q019	Na sua residência tem televisão em cores?
Q020	Na sua residência tem aparelho de DVD?
Q021	Na sua residência tem TV por assinatura?
Q022	Na sua residência tem telefone celular?
Q023	Na sua residência tem telefone fixo?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

APÊNDICE 2 – Demais Análises

2.1 Análises Descritivas

Após o tratamento dos dados restou somente uma váriavel com os dados da escola que se refere a Depêndencia Administrativa da Escola, mostrado na Figura 20.

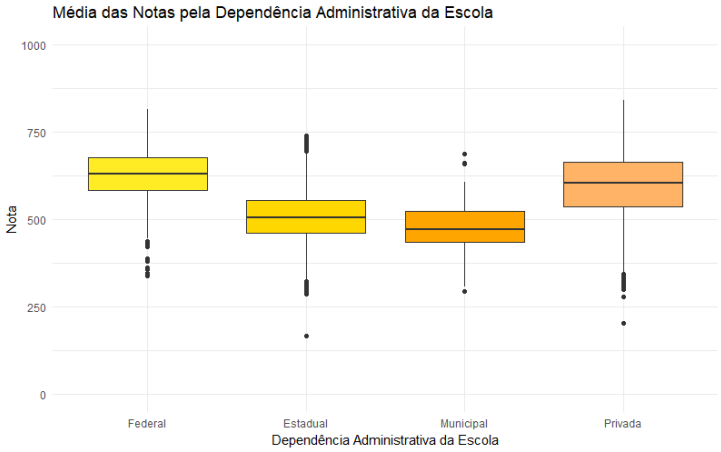

Figura 20: Média da Nota pela Dependência Administrativa da Escola

Dos dados dos alunos, foram selecionadas apenas quatro variáveis, e com elas foram construídos gráficos relacionados à Cor/Raça, conforme ilustrado na Figura 21, ao Sexo, mostrado na Figura 22, à Faixa Etária, mostrado na Figura 23,e ao Tipo de escola, mostrado na Figura 24.

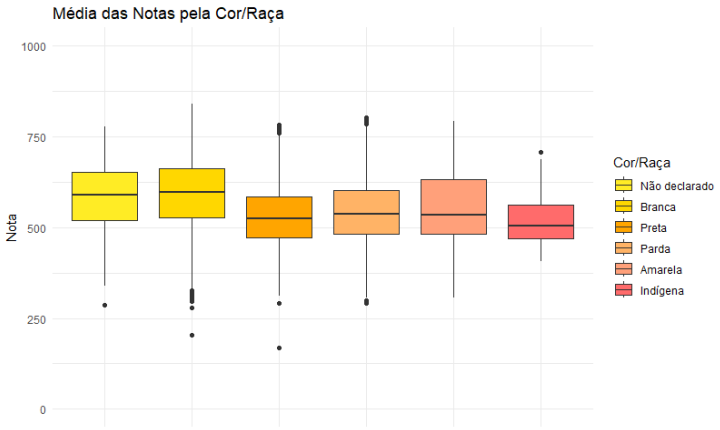

Figura 21: Média da Nota pela Cor/Raça

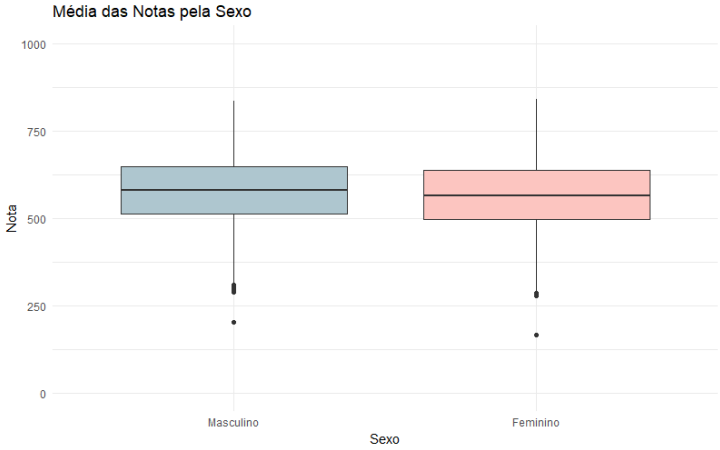

Figura 22: Média da Nota pelo sexo

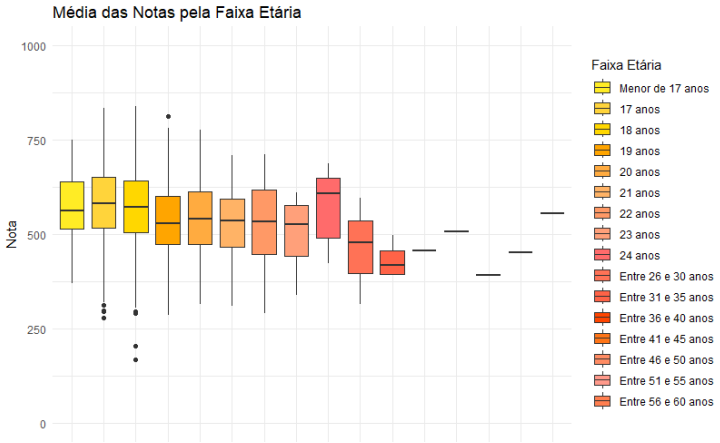

Figura 23: Média da Nota pela Faixa Etária

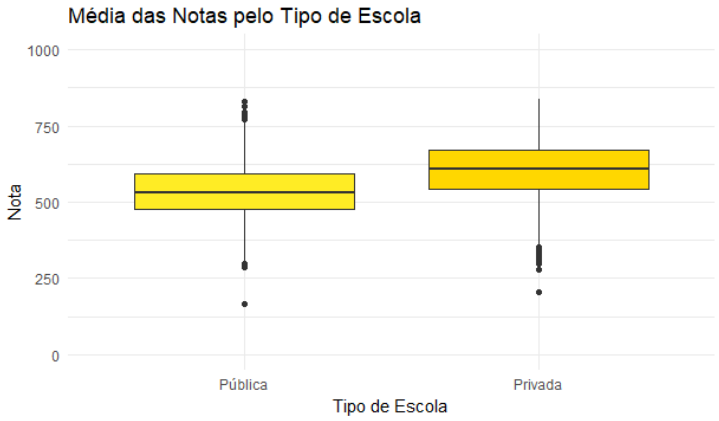

Figura 24: Média da Nota pelo Tipo de Escola

2.2 Número ótimo de Árvores no modelo de regressão

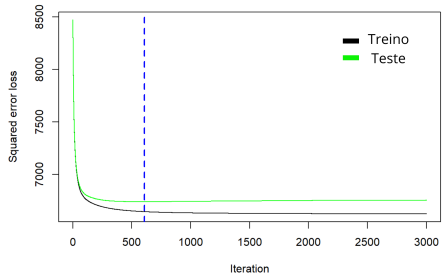

Figura 25: Número ótimo de árvores para o Modelo de Regressão da Nota em Ciências Humanas

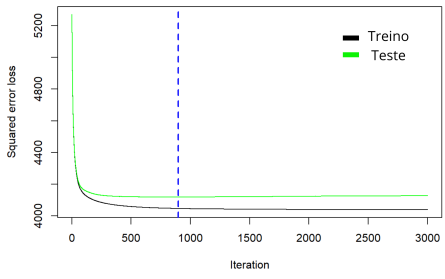

Figura 26: Número ótimo de árvores para o Modelo de Regressão da Nota em Linguagens e Códigos

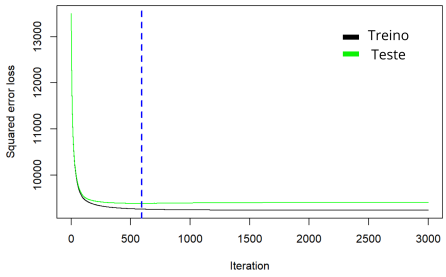

Figura 27: Número ótimo de árvores para o Modelo de Regressão da Nota em Matemática

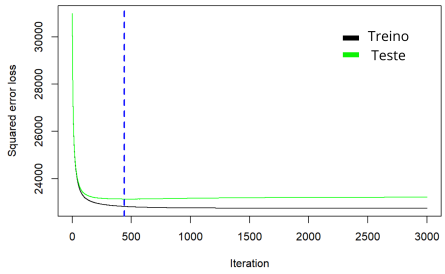

Figura 28: Número ótimo de árvores para o Modelo de Regressão da Nota em Redação

2.3 Número ótimo de Árvores no modelo de classificação

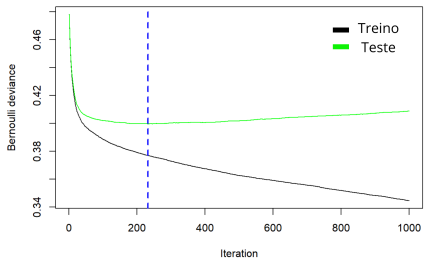

Figura 29: Número ótimo de árvores para o Modelo de Classificação para Enfermagem na UFF

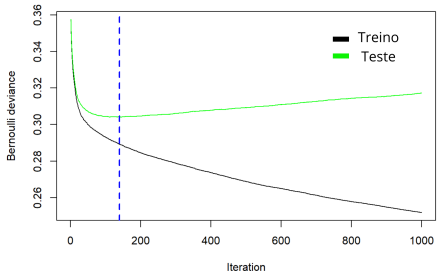

Figura 30: Número ótimo de árvores para o Modelo de Classificação para Enfermagem na UFRJ

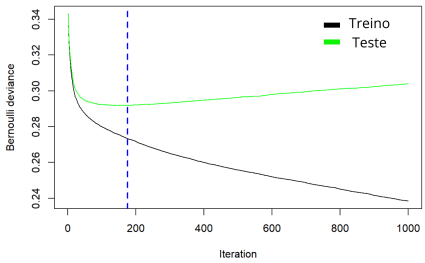

Figura 31: Número ótimo de árvores para o Modelo de Classificação para Estatística na UFRJ

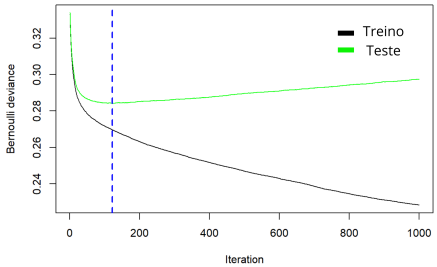

Figura 32: Número ótimo de árvores para o Modelo de Classificação para Jornalismo na UFF

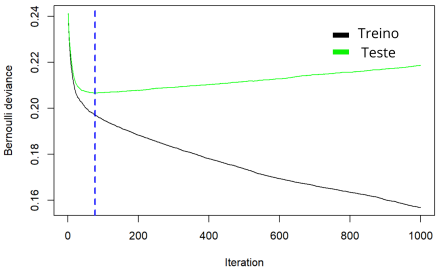

Figura 33: Número ótimo de árvores para o Modelo de Classificação para Jornalismo na UFRJ

2.4 Curva Roc

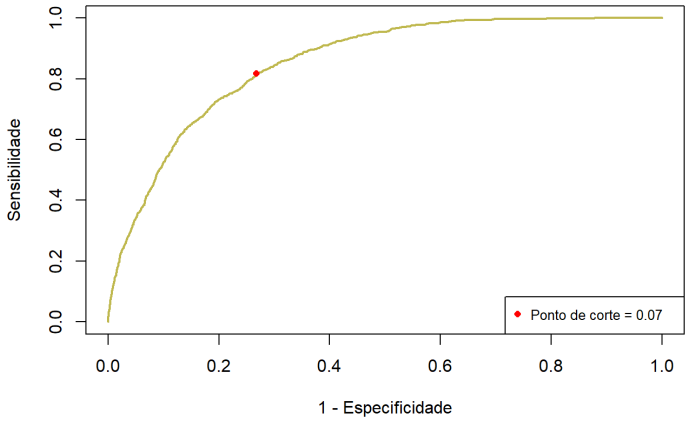

Figura 34: Curva Roc de Enfermagem na UFF

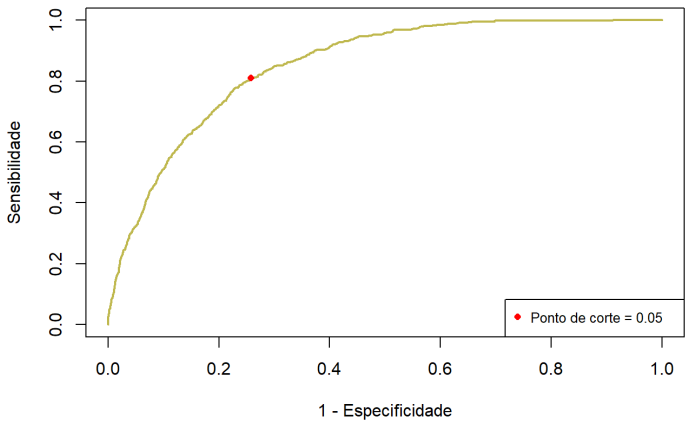

Figura 35: Curva Roc de Enfermagem na UFRJ

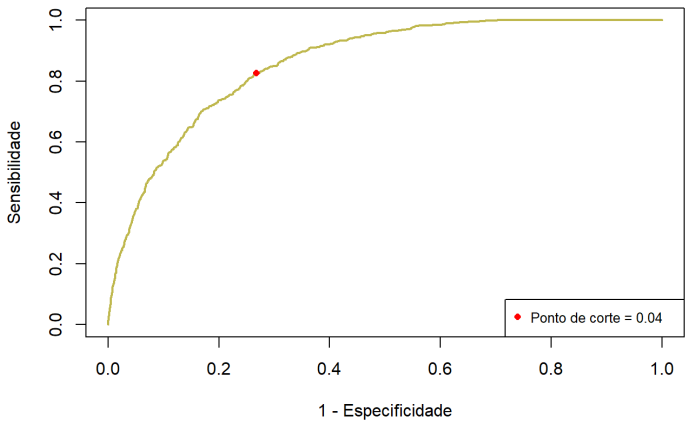

Figura 36: Curva Roc de Estatística na UFRJ

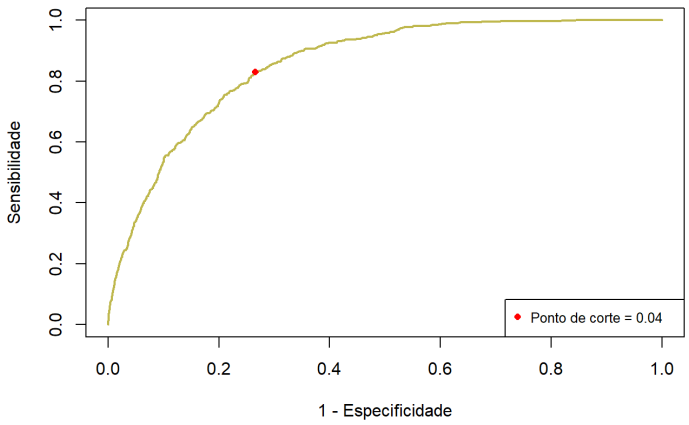

Figura 37: Curva Roc de Jornalismo na UFF

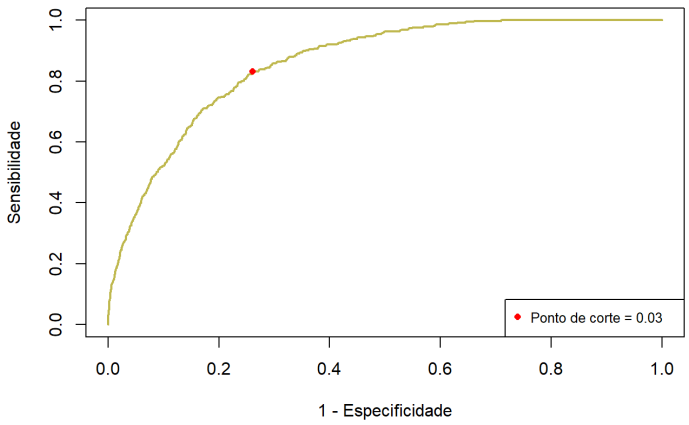

Figura 38: Curva Roc de Jornalismo na UFRJ