Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Uso de aprendizado de máquina para identificar desigualdades sociais na base de dados do ENEM

Vinicius Alberto Alves da Silva

JUIZ DE FORA
MARÇO, 2021

Uso de aprendizado de máquina para identificar desigualdades sociais na base de dados do ENEM

Vinicius Alberto Alves da Silva

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Lorenza Leão Oliveira Moreno

Coorientador: Luciana Brugiolo Gonçalves

JUIZ DE FORA

MARÇO, 2021

Uso de aprendizado de máquina para identificar desigualdades sociais na base de dados do ENEM

Vinicius Alberto Alves da Silva

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Lorenza Leão Oliveira Moreno
Doutora em Informática - PUC-Rio

Luciana Brugiolo Gonçalves
Doutora em Computação - UFF

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação (UFRJ)

Stênio Sã Rosário Furtado Soares
Doutor em Computação - UFF

JUIZ DE FORA

10 DE MARÇO, 2021

Aos meus amigos e irmãs.

Ao meu pai, pelo apoio e sustento.

Resumo

A base de dados do Exame Nacional do Ensino médio - ENEM - fornecida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP - consiste de uma fonte rica de informações sobre o processo seletivo mais importante do Brasil. Apesar do ENEM não ser um exame da qualidade da educação, compreender os diversos fatores que interferem nos desempenhos dos alunos permite a gestores de educação elaborem de politicas públicas de acesso ao ensino superior. O presente trabalho apresenta o uso de técnicas de Mineração de Dados e Aprendizado de Máquina, com foco na discussão das desigualdades sociais que afetam o desempenho dos estudantes concluintes do ensino médio que prestaram o ENEM no ano 2019. Múltiplas técnicas foram utilizadas, como algoritmos de seleção de atributos, classificação, clusterização e mineração de regras de associação, além de uma análise dos resultados a partir de estatística descritiva. A seleção de atributos apontou que características como tipo administrativo da escola, renda familiar, escolaridade dos pais e se alunos possuem um computador no domicilio, são as mais relacionadas com o desempenho de um candidato no exame. A análise dos resultados dos algoritmos evidencia que alunos com características socioeconômicas similares tendem a ter desempenho equivalente no exame.

Palavras-chave: Mineração de Dados Educacionais, Aprendizado de Máquina, ENEM, INEP.

Abstract

The database of the National High School Exam - ENEM - provided by the National Institute of Educational Studies and Research Anísio Teixeira, consists of a rich source of information about the most important selection process in Brazil. Although ENEM is not an examination of the quality of education, understanding the various factors that interfere with student performance allows education managers to develop public policies for access to higher education. This paper presents the use of Data Mining and Machine Learning techniques, focusing on the discussion of social inequalities that affect the performance of high school graduates who completed ENEM in 2019. Multiple techniques were used, as feature selection, classification, clustering, and association rule mining. Furthermore, an analysis of the results using descriptive statistics was made. The selection of attributes expresses that characteristics such as the school's administrative type, family income, parental education, and whether students have a computer are the most related to the performance of a candidate in the exam. The analysis of the results of the algorithms indicates that students with similar socioeconomic characteristics tend to have similar performance in the exam.

Keywords: Educational Data Mining, Machine Learning, ENEM, INEP.

Agradecimentos

Ao meu Pai, Carlos, pelo incentivo ao estudo, pelo sustento, pelos puxões de orelha e pelas palavras de carinho em momento difíceis. As minhas irmãs Ana Carolina e Ana Paula pelo apoio e durante estes tantos anos de estudo a os meus sobrinhos Ingrid e Cauã por me incentivarem a ver um lado mais despreocupado da vida.

As minhas amigas Laura e Eduarda e meus amigos Guilherme, Igor, Leo, Lucas, Renan e Victor que nestes anos de curso, mesmo não estudando computação, sempre estiveram lá nos altos e baixos que fazem parte da vida universitária.

As professoras Lorenza e Luciana que orientaram, com paciência e dedicação, este trabalho e a pesquisa nele envolvida. Aos professores do departamento de ciência da Computação, Marcelo Caniato, Alessandreia, Stênio e Jairo por apoiarem a minha passagem pela universidade em diversos momentos e contextos diferentes. Ao meu professor de algoritmos no IFET durante o ensino médio integrado ao técnico, Wander Gaspar, por ser o meu maior incentivador em aprender programação.

Aos muitos amigas e amigos que foram feitos nas minhas passagens pelo CGCO, GET, RECMEM e comissão InterPET/GET por mostrarem que a universidade é muito além da sala de aula.

Aos meus companheiros que cursam ciência da computação e engenharia computacional. Juntos, passamos por vários momentos divertido, seja estudando para provas ou apenas passando o tempo conversando pelo Campus.

Aos professores e funcionários do Instituto de Ciências Exatas pelos seus ensinamentos, que durante esses anos, contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

“O ser humano é aquilo que a educação faz dele.“

(Immanuel Kant)

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ENEM	Exame Nacional do Ensino Médio
ProUni	Programa Universidade para Todos
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
Ideb	Índice de Desenvolvimento da Educação Básica
MDE	Mineração de Dados Educacionais
ENADE	Exame Nacional de Desempenho de Estudantes
SISU	Sistema de Seleção Unificada

1 Introdução

Os processos de descoberta de conhecimento em bases de dados, permitem, na contemporaneidade, que gestores empreguem decisões com a ajuda de resultados processados por algoritmos. A partir de técnicas de Mineração de Dados e Aprendizado de Máquina conhecimentos que estavam inexplorados em bases de dados podem ser levados em consideração para processos de tomada de decisão.

Na área de educação não é diferente. A Mineração de Dados Educacionais - MDE - aparece como uma maneira de apoiar processos de ensino e aprendizagem (COSTA et al., 2013). Ademais, os resultados obtidos pela mineração de dados abertos podem ser usados por governos e instituições a fim de serem desenvolvidas políticas públicas (GOMES; GOUVEIA; BATISTA, 2017).

O Exame Nacional do Ensino Médio, ENEM, é desde 2009, o principal mecanismo de ingresso no ensino superior no Brasil. Seja em instituições federais com o Sistema de Seleção Unificada, SISU, ou a partir de bolsas de estudos em faculdades privadas com o Programa Universidade para Todos - ProUni. Ao se inscreverem, os participantes respondem um questionário socioeconômico. Na edição de 2019, o questionário foi composto por 25 questões. Além destas perguntas, quando o participante se identifica como concluinte do ensino médio, lhe é solicitado o preenchimento de informações sobre sua escola. Outros dados do candidato como cidade natal, cidade de residência atual, raça autodeclarada, nacionalidade, idade e sexo são solicitados. Também consiste de dados sobre o local de realização da prova, gabarito da prova, respostas dos candidatos e notas obtidas. Estas informações descritas acima constituem a base de dados do ENEM 2019 (INEP, 2020).

Apesar dos grandes potenciais de descoberta possíveis a partir dos dados do ENEM 2019, o estudo de uma base de dados tão complexa e sensível exige um olhar crítico (JR, 2019). Os cientistas de dados devem reconhecer as limitações dos algoritmos e sempre considerar as particularidades das bases de dados de forma a não perpetuar preconceitos. Por exemplo, o INEP atualmente é desaforável ao ranqueamento de escolas baseados no desempenho dos alunos no ENEM (INEP, 2019), devido a essas classificações

não levarem em conta o contexto de cada escola. Portanto, ao se trabalhar com análise de dados educacionais abertos é importante ter a consciência da característica interdisciplinar deste tópico a fim de se produzir um conhecimento que não promova desigualdade.

A avaliação da educação no Brasil tem muitas frentes de pesquisa, dada a urgência em melhorias no setor, principalmente para os necessitados de uma educação gratuita e de qualidade. Desta forma, o problema consiste em desenvolver uma solução de mineração de dados que auxilie gestores na criação de políticas públicas.

O INEP, responsável por desenvolver e aplicar o ENEM, disponibiliza os microdados do exame e incentiva a pesquisa e descoberta de conhecimento (INEP, 2020). Desta forma, mapear na base de dados do ENEM as desigualdades socioeconômicas determinantes no desempenho dos alunos é uma abordagem para fundamentar a construção de abordagens político-pedagógicas que tenham o intuito de melhorar a qualidade de ensino. Tendo em mente que o ENEM não é um exame de avaliação da qualidade da educação, a análise dos dados do exame permite auxiliar gestores a promoverem políticas públicas de acesso ao ensino superior (FILHO; ADEODATO, 2019).

1.1 Objetivos

Este trabalho de conclusão de curso é uma continuação da pesquisa realizada em (SILVA et al., 2020), naquela pesquisa, foi utilizada mineração de dados para identificação de desigualdades sociais nos dados do ENEM com foco em alunos do Estado de Minas Gerais.

O objetivo principal deste trabalho é expandir esta pesquisa, utilizando toda a base de dados do Enem 2019 e aplicar novas técnicas de mineração de dados para alcançar novos resultados ou constatar que os resultados em Minas Gerais são representativos para o Brasil.

Os objetivos secundários deste trabalho consistem em:

- Identificar padrões nas bases de dados
- Entender o comportamento dos dados relacionando desempenho no ENEM e características socioeconômicas dos candidatos.
- Estudar e aplicar as técnicas de Aprendizado de Máquina na literatura.

- Estudar características das bibliotecas de Python usadas em mineração de dados, como *Sklearn* e *Pandas*.

- Escrever códigos em Python para os processos de descoberta de conhecimento.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte maneira: o capítulo 2 destaca trabalhos da literatura que utilizam de mineração de dados para explorar a base de dados do Enem. O Capítulo 3 apresenta técnicas de processamento de dados, algoritmos de aprendizado de máquina e fundamenta métricas de avaliação que são usadas na seção de resultados. Já o capítulo 4 traz detalhes do processo de tratamento de dados e da aplicação dos algoritmos. O capítulo 5 consiste dos resultados da árvore de decisão para classificação. O resultado dos algoritmos não supervisionados estão relatados no capítulo 6. Já o capítulo apresenta um debate sobre a desigualdade social no desempenho do ENEM guiado pelos resultados apresentados anteriormente e estatística descritiva. O último capítulo conclui o estudo e apresenta propostas de pesquisa futuras.

2 Trabalhos Relacionados

A base de dados do ENEM fornecida pelo INEP é riquíssima e permite realizar pesquisas com os mais diversos tipos de abordagens. Esta seção, no geral, abrange trabalhos que de alguma forma apresentaram uma metodologia de exploração da base e construção de conhecimento baseada em mineração de dados.

Uma Revisão Sistemática é apresentada sobre análise de dados do ENEM e do ENADE até 2016 em (LIMA et al., 2019). A revisão busca responder quais são os objetivos das análises, quais suas motivações e que tipo de análise foi feita. Quatro tópicos de pesquisa são destacados com relação ao ENEM. 1) Questões pedagógicas e relacionadas com o conteúdo da prova. 2) Pesquisas relacionadas ao ingresso no ensino superior. 3) Análises envolvendo desempenho e rendimento do aluno no exame. 4) Desenvolvimento de sistemas e aplicações com os dados do exame. No texto, também é destacado que novas pesquisas utilizando mineração de dados devem ser feitas, pois vários trabalhos eram limitados apenas a estatística descritiva.

No trabalho (Silva Filho; Adeodato, 2019) é apresentada uma solução de descoberta de conhecimento em bases de dados que é capaz de predizer o desempenho de estudantes de Institutos Federais no ENEM 2016. Os resultados apresentaram que indicadores socioeconômicos são os principais fatores, porém outros fatores como "grau de escolaridade do professor", "escolha da língua inglesa na prova de ensino estrangeira" e "desejo de entrar no ensino superior" também tem impacto na predição. Já em (ADEODATO; FILHO, 2020) é apresentado uma análise sobre as bases do ENEM e Enem por Escola de 2018 que afirma que fatores como educação dos pais e professores têm maior influência no desempenho do aluno do que fatores relacionados a infraestrutura da escola. Nestas pesquisas o modelo classificador é avaliado utilizando de gráficos ROC (PRATI et al., 2008) e métrica AUC.

Definir o que é sucesso ou fracasso de estudantes é um assunto debatido na literatura e nos trabalhos (ADEODATO, 2016; ADEODATO; FILHO, 2020; Silva Filho; Adeodato, 2019) é estabelecido um limiar sobre a performance dos estudantes para definir

se o desempenho destes estudantes é bom ou não. Nestas pesquisas, o quartil superior foi usado para dicotomizar a base de dados, desta forma, alunos com notas neste quartil são aqueles que possuem um bom desempenho.

No trabalho (FRANCO et al., 2020), é apresentada uma proposta de seleção de 10 melhores atributos para classificação utilizando todas as bases do ENEM desde 1998. Para definir se um aluno tem desempenho baixo ou alto, foi delimitado o limiar de 600 pontos na média do exame. A técnica SMOTE (CHAWLA et al., 2002) foi aplicada para realizar *oversample* dos registros de classe de alto desempenho, já que esses são minoritários. Múltiplas ferramentas e algoritmos de seleção de atributos foram utilizados. As conclusões apontaram uma necessidade de reformulação no questionário socieconômico, pois ao longo dos anos algumas questões que foram retiradas eram extremamente relevantes para compreensão dos candidatos. Para fins de exemplo, em 2019, o modelo de classificação construído pela pesquisa conseguiu obter apenas 78.3% de acerto contra 88.9774% de 2012.

Aplicando o algoritmo *K-Means* na base do ENEM por Escola (INEP, 2019), em (LEONI; SAMPAIO, 2017) os autores agrupam as escolas em função do desempenho médio dos alunos e outros atributos relacionados. O resultado foi expresso em um agrupamento de que escolas com indicadores semelhantes tem desempenho equivalente no ENEM. No trabalho (SILVA et al., 2020) a base do ENEM 2019 é agrupada com *K-Means* a partir da nota dos alunos em cada uma das cinco provas que compõe o exame, os resultados apontaram que alunos com desempenho ruim no exame tem condições socioeconômicas similares e que alunos de escolas federais tem desempenho análogo a alunos de escolas particulares.

Utilizando mineração de regras de associação, dois trabalhos se destacam: No primeiro (SILVA; MORINO; SATO, 2014), utiliza-se o algoritmo *apriori* em uma implementação através da ferramenta *RapidMiner 5.1.* Os resultados apontaram que a renda familiar, o nível de escolaridade e o número de moradores da casa são fatores importantes no desempenho dos estudantes. Na mesma linha de pesquisa, em (GOMES; GOUVEIA; BATISTA, 2017), é desenvolvida uma relação entre renda familiar e desempenho dos estudantes no escopo da região Nordeste.

Nos trabalhos citados acima a conclusão está de acordo com o relatório do IPEA (BARROS et al., 2001) que levantou, utilizando de estatística, em 2001, a presença dos mesmos fatores com relação ao desempenho dos alunos. Expondo que, mesmo com uma diferença de mais de 10 anos entre os trabalhos, a condição socioeconômica é o principal fator no desempenho do aluno.

Na literatura de Ciência da Computação sobre o tema vemos que os trabalhos tem foco nas metodologias desenvolvidas e métricas de avaliação dos algoritmos. Contudo, na literatura faltam interpretações das respostas dos algoritmos, muitas vezes nos resultados são apresentado apenas tabelas ou valores com poucas discussões sobre como o desempenho do aluno pode ser afetado. Este trabalho se destaca pois tem como objetivo entender como as características sociais dos candidatos interferem na prova através dos resultados de algoritmos. Para isso, metodologias de classificação como apresentado em (ADEODATO, 2016) são aplicadas. Também são usadas técnicas de mineração de regras de associação seguindo os moldes de (GOMES; GOUVEIA; BATISTA, 2017). Na base do ENEM por aluno, não se encontrou outro trabalho aplicando clusterização.

3 Fundamentação Teórica

3.1 Descoberta de conhecimento em bases de dados

A descoberta de conhecimento em bases de dados é chamada na literatura de *Knowledge Discovery in Databases* (KDD), geralmente constituída pelas fases de seleção de dados, limpeza, enriquecimento, transformação, mineração de dados e construção de conhecimento (SIMON; CAZELLA, 2017; GOMES; GOUVEIA; BATISTA, 2017). Nesta seção, as etapas do processo de KDD são descritas. É importante destacar que é um processo iterativo, sendo habitual voltar em algumas etapas para que se possa conseguir uma melhor estrutura das informações obtidas. Na figura 3.1 é ilustrado um possível processo de KDD para a base de dados do ENEM.

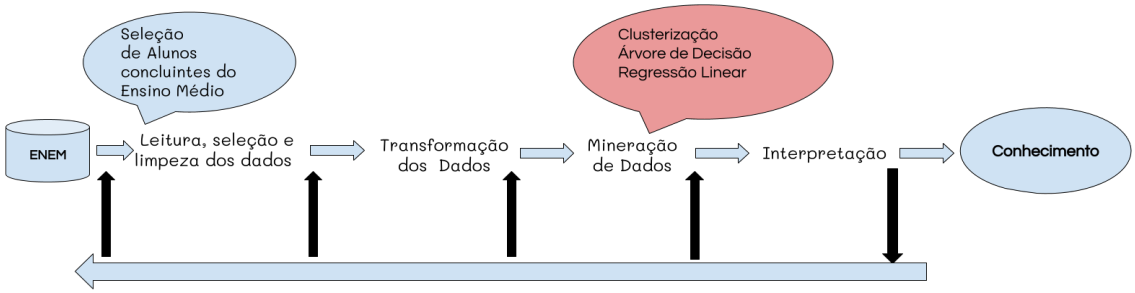


Figura 3.1: Fluxo da Descoberta de Conhecimento em bases de Dados

3.1.1 Leitura, seleção e limpeza dos dados

A primeira etapa do processo consiste selecionar uma base de dados e um escopo. A leitura dos *datasets* pode ser uma tarefa complexa devido a exigência de grandes recursos computacionais para manter a informação em memória, algumas colunas da base já podem não ser lida para economizar recursos. Por exemplo, neste trabalho foi utilizada a base de dados do ENEM com foco em alunos concluintes do ensino médio. Nesta etapa do processo a combinação de múltiplos *datasets* pode ser feita.

Também faz parte da fase de pré-processamento a limpeza informações não desejadas, como dados faltantes ou sem significado para o objetivo. A priorização não é feita somente para os registros da base, a etapa de seleção também irá retirar colunas (ou atributos) que não são interessantes.

3.1.2 Transformação dos dados

A etapa de transformação consiste em alterar o domínio dos dados de acordo com as necessidades dos algoritmos da etapa de mineração. Existem múltiplas técnicas que podem ser utilizadas neste ponto, *Label Encoding*, *Ordinal Encoding* e *One Hot Encoding*.

Label Encoding e Ordinal Encoding

O *Label Encoding* é simplesmente uma conversão direta entre o domínio atual de uma variável categórica para um domínio numérico, pode ser encarado como simplesmente trocar os valores de uma variável. Já *Ordinal Encoding* é o mesmo processo, entretanto mantendo uma relação de ordem entre as variáveis. Exemplo: Em uma base de dados de avaliações de aplicativos, os rótulos péssimo, ruim, bom, muito bom e ótimo, podem ser transformados em valores inteiros de 0 a 5. É importante destacar que, no geral, converter variáveis categóricas para numéricas utilizando esse método pode inserir viés no modelo, o que pode interferir na qualidade dos resultados. Quando o domínio categórico não tiver uma conversão adequada é indicado aplicar a técnica de *One Hot Encoding* descrita na seção abaixo.

One Hot Encoding

Como descrito em (RODRÍGUEZ et al., 2018) , este processo é usado pra transformar variáveis categóricas em uma lista de variáveis binárias. Os recursos são codificados usando um esquema de codificação *one-hot*. Em que cada rótulo da categoria se torna uma nova coluna binária criando uma matriz esparsa, onde 1 representa que o registro contém o atributo e 0 que o registro não contém, como exemplificado na tabela 3.1.

Dependência Administrativa da Escola	->	Federal	Estadual	Particular
Federal		1	0	0
Estadual		0	1	0
Estadual		0	1	0
Particular		0	0	1

Tabela 3.1: Exemplo de One Hot Encoding

3.1.3 Seleção de atributos

Seleção de atributos ou *Feature selection* é o processo de identificar a partir de um conjunto de características quais são as mais significativas para construir modelos de aprendizado de máquina (Sumaiya Thaseen; Aswani Kumar, 2017). No geral, queremos ter o menor número de atributos possíveis (ESTÉVEZ et al., 2009). Este processo tem impacto na qualidade do modelo construído, no caso de aprendizado supervisionado, pois atributos redundantes podem atrapalhar a descoberta de padrões. Também existe impacto na eficiência computacional do algoritmo e tempo de execução, pois menos recursos como memória RAM e CPU são necessários.

ChiSquare

O teste chi2 para duas variáveis é usado para verificar se duas variáveis são dependentes ou independentes, em que um maior valor da estatística chi2 apresenta que duas variáveis são dependentes. Essa definição do teste pode ser usada para seleção de atributos, para verificar quais desses atributos têm mais impacto na classificação. Em (Sumaiya Thaseen; Aswani Kumar, 2017) é apresentada uma definição formal do teste.

3.2 Aprendizado de Máquina Supervisionado

A Mineração de dados é a principal etapa do processo de descoberta de conhecimento e consiste no uso de algoritmos de aprendizado de máquina sobre as bases de dados após a etapa de transformação.

Algoritmos supervisionados são utilizados quando os registros possuem um rótulo, um valor numérico que relaciona o registro a dado evento. Quando seu valor representa algo sobre a natureza ou categoria do dado, esse rótulo é chamado de classe (SILVA;

PERES; BOSCARIOLI, 2017; REZENDE, 2003). Os rótulos podem ser inerentes à base de dados ou produzidos por um especialista. O objetivo dos algoritmos é prever esse rótulo, sendo utilizados para análise preditiva (SILVA; PERES; BOSCARIOLI, 2017).

Os métodos de aprendizado supervisionado exigem que uma parte dos registros da base de dados seja fornecida como exemplo, o conjunto de treinamento. Na fase de treinamento, é construído um modelo estatístico para associar as *features* dos respectivos registros a um rótulo.

Em termos formais, o aprendizado de máquina supervisionado consiste em construir uma função $y = F(x)$ que promova a atribuição de um rótulo y a um registro x. É possível encontrar mais detalhes em (REZENDE, 2003).

Após a fase de treinamento, o restante da base de dados é utilizado para fazer uma avaliação do modelo estatístico construído. Esta etapa é chamada de teste e é melhor detalhada na seção 3.2.1.

Se o objetivo é prever o valor numérico de um rótulo, então y é uma variável numérica (contínua ou discreta) e este cenário é chamado de Regressão. Caso y represente uma categoria do domínio, o processo é conhecido como Classificação. Neste trabalho, é usado apenas classificação.

3.2.1 Classificação

A tarefa de classificação é o processo de identificar a qual categoria do domínio dos dados um registro pertence. Quando só existem duas classes no domínio, o problema é chamado de classificação binária; casos com mais classes são denominados classificação multi-classe.

Árvores de Decisão

A árvore de decisão é um modelo de aprendizado de máquina supervisionado que apresenta regras de decisão baseadas nos atributos dos registros. Uma vantagem da árvore de decisão é que como o modelo pode ser visualizado e interpretado, um gestor pode obter conhecimento para tomada de decisão (ADEODATO, 2016).

A árvore de decisão tem os seguintes componentes: nós folhas, também chamados de nós resposta, que representam uma classe; nós internos, também chamados de nós de

decisão, que contêm um teste sobre um atributo; e arestas, ou ramificações, que conectam um nó interno a uma subárvore.

Na Figura 3.2, é possível ver um exemplo elementar de árvore de decisão. Neste modelo, "Paciente se sente bem", "Paciente tem dor" e "Temperatura do paciente" são características do paciente, portanto nós de decisão, e cada aresta está associada a um valor. Já as folhas "saudável" e "doente" são nós respostas. Após a árvore estar construída, para decidir se qualquer paciente está saudável ou doente basta seguir o fluxo da árvore começando pela raiz.

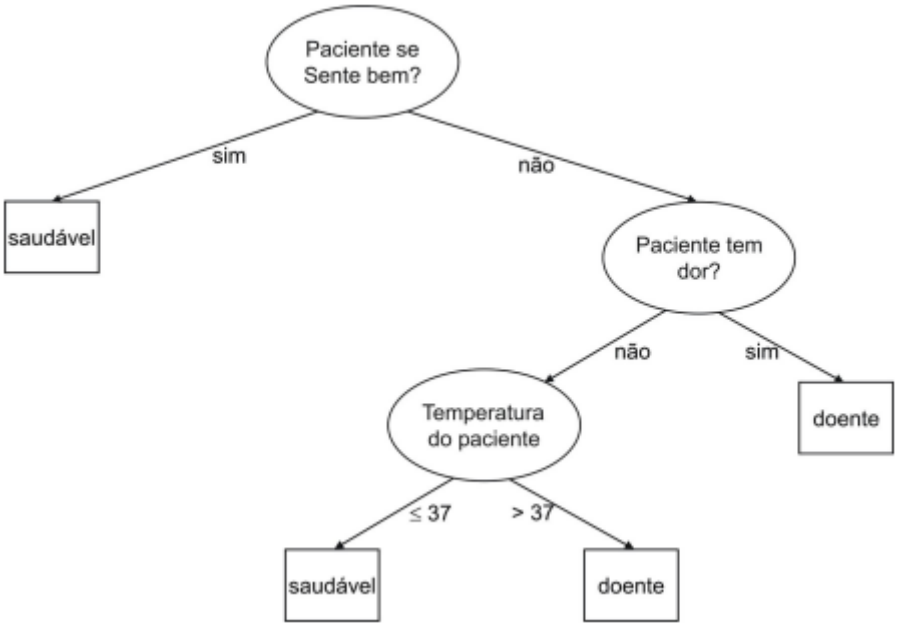


Figura 3.2: Exemplo De Árvores de Decisão
Fonte: (REZENDE, 2003)

Avaliação de Classificação

Ao utilizarmos de modelos para predição é fundamental avaliar o quão bem eles conseguem realizar a tarefa proposta. Para isso, após o treinamento do modelo tem-se a parte de teste e em seguida avaliação. A Matriz de Confusão, ou tabela de contingência, é uma das formas mais tradicionais de avaliação de modelos de classificação, ela mostra uma comparação entre as classificações reais e as classificações encontradas pelo modelo (REZENDE, 2003).

Sendo uma classificação binária, com rótulos positivo e negativo, os registros classificados corretamente são chamados de TP (True Positive) e TN (True Negative).

Quando a classificação é incorreta, os registros são denominados FP (False Positive) e FN (False Negative). A matriz de confusão para a classificação binária é uma matriz 2x2, em que o número de registros classificados corretamente está na diagonal principal.

Para entender melhor a matriz de confusão, são estabelecidas as métricas Acurácia, Precisão, *Recall* e *F-Score* que em conjunto são usadas para avaliar o teste de um modelo. Segue uma explicação sucinta de cada métrica:

Acurácia De todas as classes, o quanto é previsto corretamente. O valor deve ser o mais alto possível.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.1}$$

Precisão De todas as classes positivas que foram previstas, quantas são realmente positivas.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall Para as classes positivas, o quanto é previsto corretamente. Deve ser o mais alto possível.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

F-score É difícil comparar dois modelos com baixa precisão e alto recall ou vice-versa. Portanto, para torná-los comparáveis, é utilizado o F-Score. O F-Score ajuda a medir o recall e a precisão ao mesmo tempo.

$$FScore = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.4}$$

Curva ROC

Curvas ROC, do inglês *Receiver Operating Characteristic*, é um método visual

Classe	**Classe Prevista como True**	**Classe Prevista como False**
True	TP (True Positve)	TN (True Negative)
False	FP (False Positive)	FN (False Negative)

Tabela 3.2: Matrix de Confusão

para avaliação de modelos classificadores (PRATI et al., 2008). Para classificação binária, o gráfico ROC é construído com a True Positive Rate (TPR), calculada por $tpr = P(X|Y)$ e com a False Positive Rate (TPR), calculada por $fpr = P(Y|X)$, que, de forma simplificada, representam o quanto a taxa de acerto cresce junto à taxa de erros. Para gerar o gráfico ROC, fpr é plotada no eixo X e tpr no eixo y.

Gráficos ROC e valor AUC (apresentado a seguir) são discutidos de forma excelente em (PRATI et al., 2008), base para escrita desta seção.

Neste gráfico, um ponto $(0, 100)$ significa um modelo que acerta todas as predições, já um ponto $(100, 0)$ é um modelo que erra todas. Assim sendo, gráficos com curva próximas do canto superior esquerdo tem melhor performance. Um reta entre o ponto $(0, 0)$ e $(100, 0)$ é um modelo que não consegue diferenciar entre as duas classes.

Valor AUC

O Valor AUC - do inglês, *Area Under the Curve*, é uma sumarização do gráfico ROC. Consiste basicamente da área abaixo da curva do Gráfico ROC, sendo que esta área é uma fração de um quadrado de lado um. Desta forma, o valor AUX está sempre entre 0 e 1, em que valores próximos de 1 significam modelos que conseguem diferenciar melhor entre as duas classes.

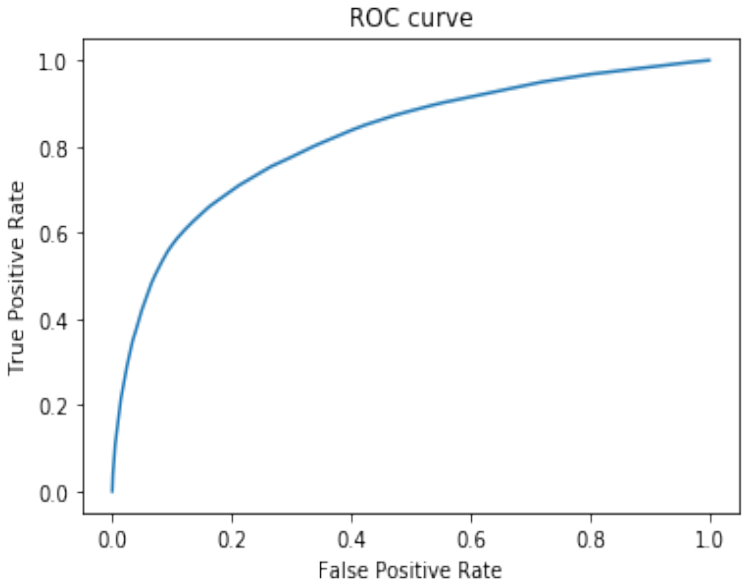


Figura 3.3: Gráfico ROC

3.3 Aprendizado de Máquina Não Supervisionado

No Aprendizado Não Supervisionado, não é necessário que os registros da base tenham algum rótulo os algoritmos aqui aprendem procurando padrões nos dados sem um objetivo especifico.

3.3.1 Clusterização

Dado um conjunto C com n objetos, o problema da k-clusterização consiste em dividi-lo em k subconjuntos disjuntos, chamados *clusters*, baseando-se na similaridade entre seus objetos.

Para medir a similaridade entre dois elementos da base, diferentes métricas podem ser utilizadas, a principal delas é as distância euclideana. Cada registro na base de dados é representado por um ponto na dimensão R^n , dados dois pontos p e q, a dissimilaridade entre eles é calculada a partir da equação 3.5, onde i é cada uma das N dimensões que o objeto possui.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{3.5}$$

Uma das técnicas de clusterização mais famosas e utilizadas é o *k-means*. Detalhes do algoritmo podem ser encontrados em (LEONI; SAMPAIO, 2017). O algoritmo forma clusters iterativamente, até que não haja mudança significativa na posição dos centróides. Um centróide corresponde ao ponto médio em R^n relativo à localização de todos os elementos de um mesmo cluster. Inicialmente, o algoritmo cria k clusters escolhendo de forma aleatória k elementos da base de dados. A cada iteração, são recalculados os centróides e todos os elementos são realocados no cluster do centróide mais próximo. Assim, o *k-means* define um agrupamento que minimiza a similaridade (isto é, a distancia média) do centróide aos objetos do mesmo cluster equação.

Outros algoritmos da biblioteca como *DBSCAN*, *Aglomerative Clustering* e *Birch*, foram testados, infelizmente a aplicação deles não se mostrou viável devido ao tamanho da base e por consequência um elevado consumo de memória RAM.

Função de Avaliação

O Índice Silhueta, é a métrica de avaliação de qualidade em clusterização mais utilizada na literatura. Para cada objeto ela indica o quanto aquele objeto é similar ao seu grupo e dissimilar ao grupo vizinho mais próximo. Para avaliar a solução, utilizamos a média de todos os índices de silhueta encontrados. Uma descrição mais detalhada compõe o resto desta seção.

Seja i um objeto pertencente ao cluster C_w, então $a(i)$ é a dissimilaridade do objeto i para cada outro objeto j de C_w. Se C_w possui um único elemento, então $a(i)$ é fixado como zero, caso contrário, $a(i)$ é calculado na equação (3.6) da seguinte forma:

$$a(i) = \frac{1}{|C_w|} \sum_{j=1} C_w | d_{i,j} \forall j \neq i \tag{3.6}$$

A distância do objeto $i \in C_w$ para cada cluster C_v, com $v \neq w$ é dada por:

$$d(i) = \frac{1}{|C_v|} \sum_{i=1} C_v | d_{i,j} \tag{3.7}$$

A dissimilaridade $b(i)$ do objeto i para cada outro objeto j do seu cluster é:

$$b(i) = min(d(i, C_v)) \tag{3.8}$$

Finalmente, o valor da silhueta $s(i)$ para o objeto i é dado por:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3.9}$$

Com isso, é possível definir como avaliar uma solução calculando a média dos índices de silhueta encontrado pelo algoritmo e a função objetivo é definida da seguinte forma:

$$\max \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{3.10}$$

3.3.2 Regras de Associação

A técnica de mineração de regras de associação é usada para encontrar afinidade entre itens de uma base de dados. Um item corresponde a cada variável que pode descrever um objeto, como Raça="Branco" ou EscolaridadeMãe="NãoSabe".

Existem múltiplas implementações de algoritmos para minerar regras de associação na literatura. O mais famoso é o *Apriori* (AGRAWAL; SRIKANT et al., 1994), existem outras implementações mais eficientes como a de (HAN et al., 2004). Estes algoritmos são usados na literatura para extrair regras de associação a partir da popularidade dos itens. Regras de Associação são relações de implicação ($A \rightarrow B$), onde lê-se A (antecedente) implica em B (consequente). Numa regra, A e B são *itemsets*, subconjuntos de itens da base de dados, obrigatoriamente disjuntos e com ao menos um elemento.

Componentes do Algoritmo

Diferentes métricas norteiam regras de associação. O suporte indica o quanto um *itemset* B é frequente na base de dados, sendo dado pela razão entre o número de registros que contém o *itemset* e o total de registros da base, representado na equação 3.11.

Caso o suporte refira-se a uma regra de associação ($A \rightarrow B$), o numerador deve indicar o número de registros que contém tanto A quanto B.

$$Support(B) = \frac{Registros\ Contendo\ B}{Total\ Registros} \tag{3.11}$$

A confiança, representa a chance de B ocorrer em um registro, sabendo-se que A ocorre a priori. É calculada como descrito na equação 3.12:

$$Confidence(A \rightarrow B) = \frac{Registros\ contendo\ A\ e\ B}{Registros\ contendo\ A} \tag{3.12}$$

Já a métrica $Lift$, refere-se ao aumento na proporção em que B ocorre quando A ocorre a priori, levando em conta a popularidade do *itemset* B. Nos casos em que os valores de *Lift* são inferiores a 1, as respectivas regras não têm significância.

$$Lift(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A) * Support(B)} \tag{3.13}$$

Para cada regra de associação, o algoritmo calcula os valores destas três métricas. Observando-as, é possível estimar a relevância de cada regra de associação encontrada no processo de mineração. Mais detalhes em (AGRAWAL; SRIKANT et al., 1994) e (HAN et al., 2004).

4 Descoberta de conhecimento na base do ENEM

Neste trabalho, cada etapa da metodologia é relacionada com um etapa do processo de Descoberta de Conhecimento descrito na seção anterior. Foi realizada uma pesquisa quantitativa utilizando de métricas e estatística para apresentar os resultados. Dito que o objetivo do trabalho é caracterizar relações entre variáveis socioeconômicas e rendimento dos alunos no ENEM a pesquisa pode ser entendida como descritiva. Por último, a pesquisa pode ser considerada documental já que o objetivo é analisar um conjunto de dados produzida por uma entidade governamental (INEP).

Todo o desenvolvimento deste trabalho foi realizado na linguagem *Python*, as vantagens dessa linguagem é sua comunidade engajadas e um grande número de bibliotecas voltadas para ciência de dados. Algumas bibliotecas são fundamentais nesta pesquisa e merecem destaque.

O pacote Pandas (MCKINNEY, 2010) é usado principalmente leitura e manipulação de *datasets*, ele já contém métodos para importação de diversos formatos de arquivo, limpeza e tratamento de dados. É uma ferramenta indispensável para trabalhar com ciência de dados no *Python*. A *Scikit-learn* ou *sklearn* (PEDREGOSA et al., 2011) é provavelmente a biblioteca mais utilizada para aprendizado de máquina em *Python*, ela contém implementações eficientes de diversos algoritmos de aprendizado de máquina. A biblioteca sklearn contém muitas ferramentas eficientes para aprendizado de máquina e modelagem estatística, incluindo classificação, regressão, agrupamento e redução de dimensionalidade. A biblioteca *Seaborn* (WASKOM; TEAM, 2020) estende a *matplotlib* (HUNTER, 2007) para permitir a criação de gráficos e soluções para visualização de dados de forma atraente e com pouquíssimas linhas de código.

4.1 Leitura e Limpeza dos Dados

Sem tratamento algum, a base do ENEM 2019, no formato csv, tem aproximadamente 3,12 GBs de tamanho. Como dito anteriormente utilizou-se a biblioteca *Pandas* para leitura e tratamento da base de dados, uma de suas funcionalidades é ler a possibilidade de ler apenas algumas colunas do arquivo. Mesmo assim, é custoso manter todas as linhas em memória.

Dito que apenas os registros dos concluintes tem informação sobre a escola, ou seja, dados como código da escola ou código da unidade da federação da escola (CO_UF_ESC), só estão disponíveis para estes registros. Assim, algumas colunas da base foram lidas inicialmente em pedaços, salvas em arquivo e concatenada depois apenas com as linhas com valor de (CO_UF_ESC) válido, ou seja, de 11 a 53.

Na limpeza dos dados, foram considerados apenas linhas totalmente preenchidas e registros com notas diferentes de zero, o que resultou em alunos de todo o Brasil a 920.588 que compareceram a todas as provas e não foram desclassificados.

4.2 Tratamento dos Dados

Juntamente com o arquivo csv que compõe a base, em (INEP, 2019) também são fornecidos dicionários que especificam o valor de cada atributo, contudo, de início houve a troca de alguns rótulos para facilitar o processo de mineração de dados e trazer uma compreensão melhor do domínio de dados.

A coluna Renda Mensal Familiar, que originalmente contém 17 faixas de renda, foi transformada em Classes Sociais de acordo com critérios do IBGE. Tendo como base de que o salário mínimo de 2019 é de R\$998 reais. As famílias de classe E são aquelas que possuem os rendimentos de no máximo dois salários mínimos. Classe D de dois a quatro salários mínimos. Famílias de Classe C são aquelas de quatro a 10 salários mínimos. Por fim, as classes mais altas: a B representa mais de 10 salários mínimos e a Classe A mais de 20 salários mínimos. As colunas "Quantidade de PCs no domicílio" e "Quantidade de celulares no domicílio" foram transformadas para binária {"Não Possui" ou "Possui" }. A variável "Na sua residência tem internet?" já é binária.

Foi criada uma nova coluna para representar a nota final do estudante, “Nota Média”, que consiste da média simples das notas das provas do ENEM nas cinco áreas - Linguagens e Códigos, Ciências Humanas, Ciências da natureza, Matemática e Redação.

Para Estudo Pai e Estudo Mãe consideram-se os valores: “Fundamental Incompleto”,“Fundamental Completo”, “Médio”, “Superior” e “Não Sabe” (registros de “Pós-Graduação” foram unificados como “Superior”).

Concluindo, os atributos transformados e seus possíveis valores são os listados a seguir. Itens com os mesmos valores aparecem em conjunto.

- Unidade Federação da Escola
  - O campo contém a sigla de um das 26 unidades da federação e o distrito federal.
- Tipo Dependência Administrativa Escola
  - Municipal
  - Estadual
  - Federal
  - Particular
- Localização Escola
  - Urbano
  - Rural
- Sexo
  - Masculino
  - Feminino
- Raça
  - Parda
  - Branco
  - Amarelo

  - Indígena
  - Não Declarado

- Língua prova estrangeira
  - Inglês
  - Espanhol

- Estudo Pai e Estudo Mãe
  - NaoSabe
  - FundamentalIncompleto
  - FundamentalCompleto
  - Medio
  - Superior

- Trabalho Pai e Trabalho Mãe.
  - Grupo(A,B,C)
  - Grupo(D,E)
  - NãoSabe

- Tem PC, Tem Celular e Tem Internet
  - Sim
  - Não

- Classe Social
  - A
  - B
  - C
  - D
  - E

Devido a utilização de múltiplos algoritmos com objetivos distintos, o tratamento de dados deve ser diferente para cada técnica.

Tratamento de dados para Regras de Associação

Nas regras de associação, alguns dos rótulos foram adaptados para favorecer as características do algoritmos. A variável "Escolaridade da mãe" foi tratada para indicar se a mãe do estudante contém o ensino "Médio Incompleto", "Médio Completo" ou a opção "Não Sabe". Para raça autodeclarada, aqueles que se afirmaram pretos, pardos, indígenas e amarelos foram tomados como "Não-Branco" (critério justificado em (SOUZA; RIBEIRO; CARVALHAES, 2010)), reduzindo o domínio a {"Branco", "Não-Branco" ou "Não Declarado"}.

Tratamento de dados para Classificação

Devido às implementações da *sklearn* exigirem que apenas sejam fornecidos valores numéricos para os modelos de classificação, é necessário fazer transformações nas *features*, já que a maioria é categórica. Para utilização dos métodos de seleção de atributos da *sklearn* foi necessário a transformação dos atributos utilizando *label encoding* e *ordinal encoding*. Ademais, a transformação realizada nos dados para a árvore de decisão foi *One-Hot-Encoding*, utilizando o método da biblioteca *Pandas* chamado *get_dummies*.

Tratamento de dados para Clusterização

Os atributos escolhidos para o processo de clusterização, foram as notas das provas de cada inscrito. Para a clusterização, os valores foram normalizados usando o método *quantile_transform* da *Sklearn*. Neste método os atributos são transformados em uma distribuição uniforme, espalhando os valores dos atributos mais frequentes, reduzindo o impacto de *outliers*. O número de quantis igual a 5. Este parâmetro foi ajustado via abordagem empírica.

4.3 Classificação

O objetivo da classificação é a partir dos dados fornecidos durante a inscrição do participante prever o quanto este irá performar no ENEM. Decidir se um participante irá performar bem ou mal em algum exame é algo debatido na literatura. Em (ADEODATO, 2016; ADEODATO; FILHO, 2020; Silva Filho; Adeodato, 2019) é justificado que alunos no quartil superior da nota média possuem um bom desempenho. Neste trabalho, foi seguida esta metodologia, em que o desempenho do aluno é a média simples das cinco provas que compõe o exame. De acordo com este critério e com nossa base de dados, alunos com nota superior a $580,64$ tem um alto desempenho.

Desta forma, foi criada uma nova coluna na base de dados, em que registros no quartil superior receberam o valor 1, classe positiva, e o restantes dos dados o valor 0, classe negativa. Para classificação, foi utilizado o método *DecisionTreeClassifier* do módulo *tree* da biblioteca *sklearn* (PEDREGOSA et al., 2011). Os valores *default* dos parâmetros foram mantidos, entretanto, através de análise empírica, foi decidido que o melhor valor de max_depth era igual a 8.

Outros algoritmos como *Categorical Naive Bayes*, *KNN* e *SVC* da *sklearn* foram experimentados e obtiveram resultados nas métricas da seção 3.2.1 semelhantes. Contudo, como o objetivo é identificar desigualdades para facilitar o trabalho de gestores é essencial ter um modelo com resultados interpretáveis e não apenas obter um sistema de predição. Neste contexto , apenas a Árvore de Decisão é interessante.

4.3.1 Seleção de Atributos

Utilizando de conhecimento do domínio e se baseando na pesquisa de (FRANCO et al., 2020), foi feita uma seleção prévia dos atributos que naturalmente não seriam interessantes para a pesquisa. Atributos relacionados a necessidades e especificidades do participante, dados do local de prova foram descartados. Dados como idade e estado civil também, já que o foco é em alunos do Ensino Médio. Essa filtragem inicial na base completa é estritamente necessária, pois rodar qualquer algoritmo com todas as colunas exigiria muito consumo de memória RAM e não agregaria valor. Desta forma, os atributos escolhidos como entrada do método de *feature selection* são os descritos na seção 4.2.

Neste trabalho, a utilização de um procedimento formal de seleção de atributos é necessária apenas para a classificação, já que o objetivo proposto para a clusterização demanda os atributos relacionados às notas.

Utilizando do modulo de *feature selection* da *Sklearn* foi utilizada uma métrica para seleção dos atributos através de uma pontuação, de forma que *features* com maior pontuação tem maior impacto na classificação. A métrica escolhida foi a *ChiSquare*, apresentada na seção 3.1.3, pois ela é muito aplicada quando as *features* são categóricas.

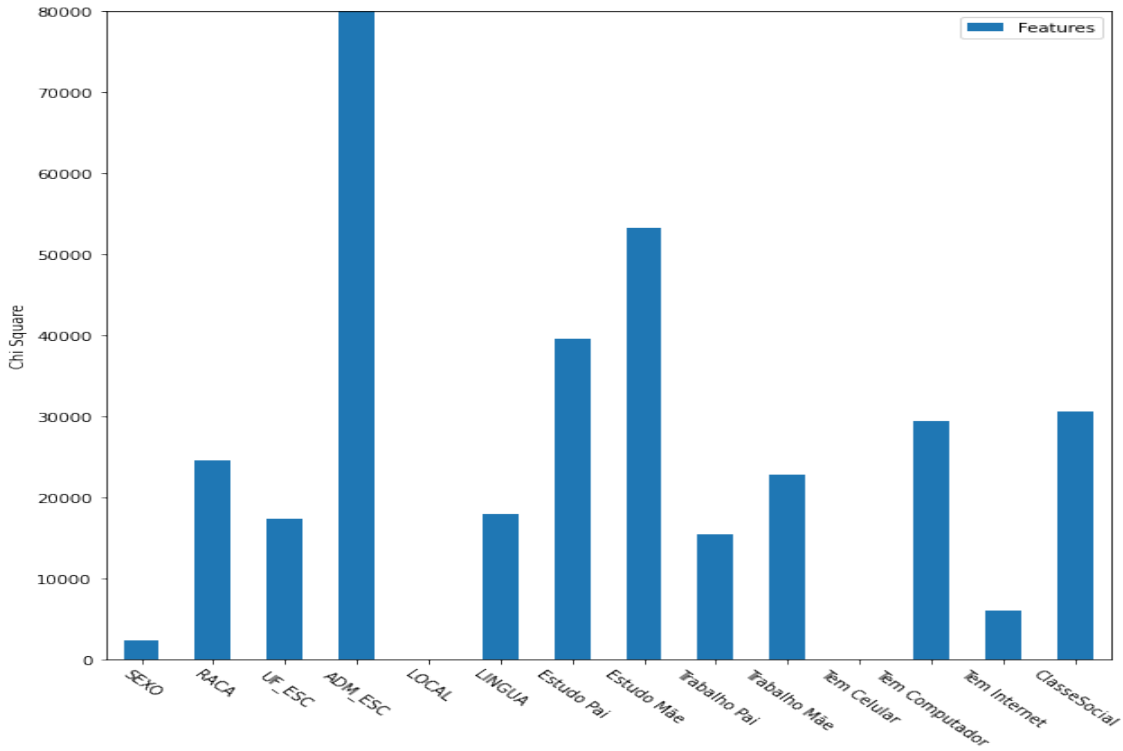
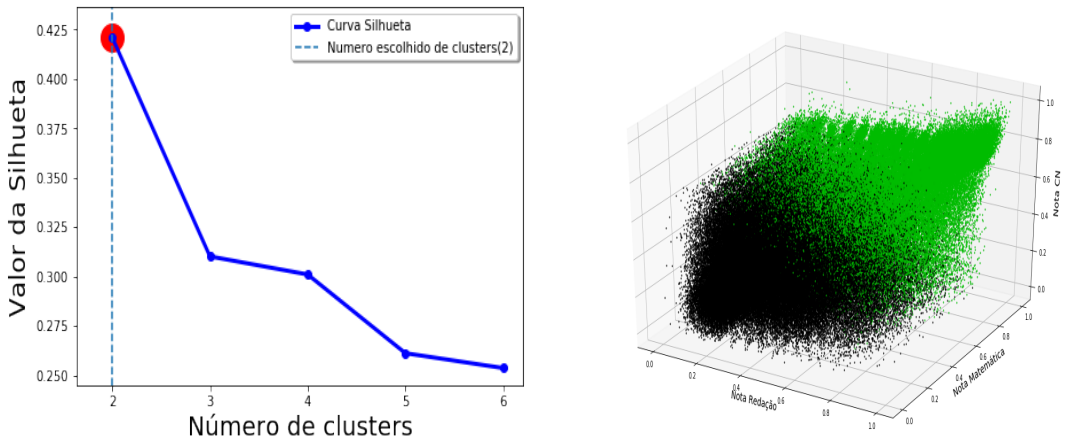


Figura 4.1: *Features* com maior significado de acordo com o processo de seleção Chi-quare

Vemos que os cinco atributos com mais relevância segundo esse critério são Tipo Administrativo da Escola, Estudo da Mãe, Estudo do Pai, Classe Social e Tem Computador. Vale destacar que características como Sexo, Localização da Escola e se o candidato possui celular são as que menos tem relação com o desempenho. O parâmetro de quantos atributos seriam utilizados foi ajustado via abordagem empírica.

4.3.2 Treinamento e Avaliação

Com os atributos selecionados, a base foi dividida entre treino e teste, sendo que 25% dos registros ficaram para teste. O método de seleção foi amostragem estratificada, para que



(a) Silhueta em função de K

(b) Visualização 3D dos dados agrupados

Figura 4.2: Valor de Silhueta dado número de clusters e Visualização 3D dos dados agrupados.

as classes aparecessem de maneira proporcional às suas quantidades originais no treino e no teste.

4.4 Clusterização

Para decidir qual o melhor número de clusters, o algoritmo *k-means* foi testado com valores de k variando de 2 a 6, conforme indica o gráfico na figura 4.2(a), sendo o melhor valor de Silhueta encontrado para $k = 2$. A Silhueta foi calculada usando o método *silhouette_score* da biblioteca, que é um implementação da equação apresentada em 3.10.

Nestes testes, o número máximo de iterações é de 300, o valor padrão da implementação do algoritmo na biblioteca. A figura 4.2(b) tem-se uma visualização 3D da base de dados já agrupada. O eixo X é a nota de Redação, eixo Z de Matemática e eixo Y de Ciências da Natureza.

O resultado da clusterização é um novo rótulo para cada registro em que identifica o cluster a qual o mesmo foi agrupado. Esse rótulo é adicionado como um novo atributo na base de dados. Por fim, os agrupamentos resultantes serão usados para identificar relações entre as características socioeconômicas com o desempenho dos aluno com as regras de associação.

4.5 Regras de Associação

O *toolkit Orange* (DEMVsAR et al., 2013) é uma ferramenta de Mineração de Dados que usa programação visual, focada para exploração de dados. Para este trabalho, foi utilizado o widget *Orange3-Associate*, em que é implementado o algoritmo de mineração de regras de associação apresentado em (HAN et al., 2004). O uso desta ferramenta é extremamente intuitivo visto que não é exigido um formato específico para entrada dos dados e execução rápida, mesmo para o grande volume de dados. Como este processo é realizado depois da Clusterização, o algoritmo foi executado considerando-se toda a base de dados com o rótulo de cada cluster. Os valores das métricas, que permitem validar e entender melhor o comportamento das regras de associação, foram limitados da seguinte forma: suporte mínimo de 20%, confiança de 70% e *lift* maior que 1. Esses valores foram escolhidos baseados nas métricas apresentadas em (SILVA; MORINO; SATO, 2014).

5 Resultados e Discussões Aprendizado de Máquina Supervisionado

5.1 Classificação

A tarefa de classificar um aluno como Alto Desempenho ou Baixo Desempenho a partir dos dados de inscrição foi realizada pela Árvore de Decisão com até 82% de acurácia. Entretanto, como visto na seção 3.2.1, a acurácia não é suficiente para medir a qualidade de um modelo. Especificamente nesta tarefa, como as classes são muito desbalanceadas, a acurácia é uma armadilha. Imaginando para o nosso problema, em que a classe predominantes é de 75%, um modelo que "chutasse" apenas esta classe teria 75% de acurácia, mas não seria um modelo que têm um resultado significativo.

Dito isso, os modelos serão avaliados usando as métricas apresentadas na Seção 3.2.1 e na Seção seguinte é feita uma leitura sobre as regras escolhidas pela árvore de decisão. O modelo, que foi construído utilizando os atributos selecionados pela técnica *ChiSquare*, alcançou resultados satisfatórios. Porém, como é possível observar na tabela 5.2 o modelo tem dificuldades em acertar a classe positiva, falhando em oferecer robustez para classificar registros como alto desempenho. A tabela 5.1 resume os resultados.

Tabela 5.1: Métricas de avaliação dos modelos de classificação

	Resultados
Acurácia	0.825
Precisão	0.67
Recall	0.53
F-Score	0.59
AUC	0.72

Classe	**Classe Prevista como True**	**Classe Prevista como False**
True	53% (30569)	8% (1445)
False	47% (26958)	92% (158175)

Tabela 5.2: Matriz de Confusão Resultante

5.1.1 Árvore de Decisão Resultante

Nas Árvores de Decisão é possível seguir o fluxo das regras desde a raiz até as folhas para classificar os registros.

Na raiz da Figura 5.1 tem-se a decisão mais importante, se o registro pertencer a escola estadual, então seguiremos o ramo da direita, caso contrário o ramo da esquerda. Novamente, para o segundo nível, tem-se duas possibilidades para cada nó. Para facilitar a compreensão, serão discutidas 4 subárvores. Duas estão do lado esquerdo da raiz, Figuras 5.5 e 5.4, e duas do lado direito, Figuras 5.3 e 5.2. Nas imagens, nós com cor laranja indicam classe de baixo desempenho (classe negativa) e nós com cor azul, alto desempenho (classe positiva). Quanto mais esbranquiçado, maior dúvida em classificar aquele nó.

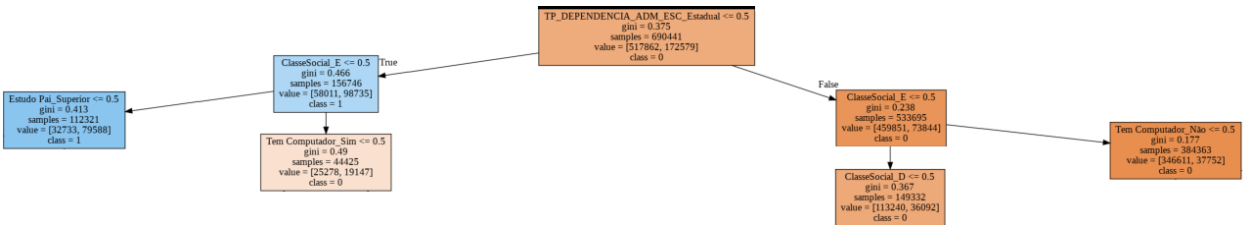


Figura 5.1: Raiz da árvore

Nos ramos à direita da raiz, ou seja, que representam estudantes de escola estadual, todos os registros são classificados como baixo desempenho. Os atributos não permitem ao modelo distinguir, entre os alunos de escola estadual, quais têm bom desempenho. Podemos ver na Figura 5.2 que os nós têm forte cor laranja, indicando certeza do modelo. Já no ramo direito central da figura vemos alguns nós esbranquiçados, principalmente aqueles com pai e mãe com ensino superior.

No ramo esquerdo central, aqueles que não possuem computador já são classificados por baixo desempenho. Naqueles que possuem computador, é verificado se o aluno estuda em escola municipal; caso afirmativo, é classificado como baixo desempenho. No outro ramo, ou seja, referente a alunos que estudam em escolas particulares ou federais, a tendência é classificar como um, entretanto o modelo tem bastante dúvida. Só existe certeza quando a escolaridade do pai é superior. Mostrando a importância desses atributos.

Observando a subárvore extrema esquerda (Figura 5.5) vemos que todos os regis-

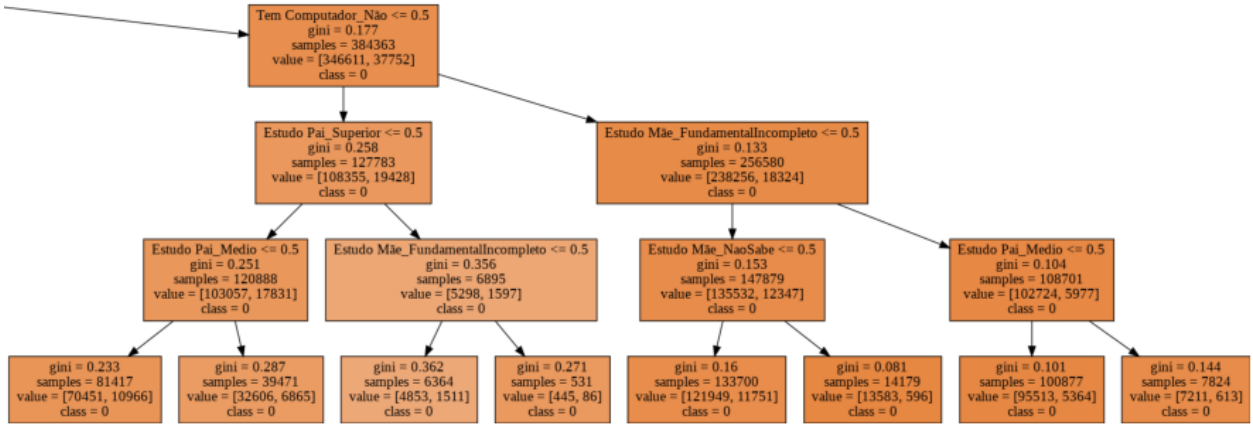


Figura 5.2: Ramo extremo direito da árvore

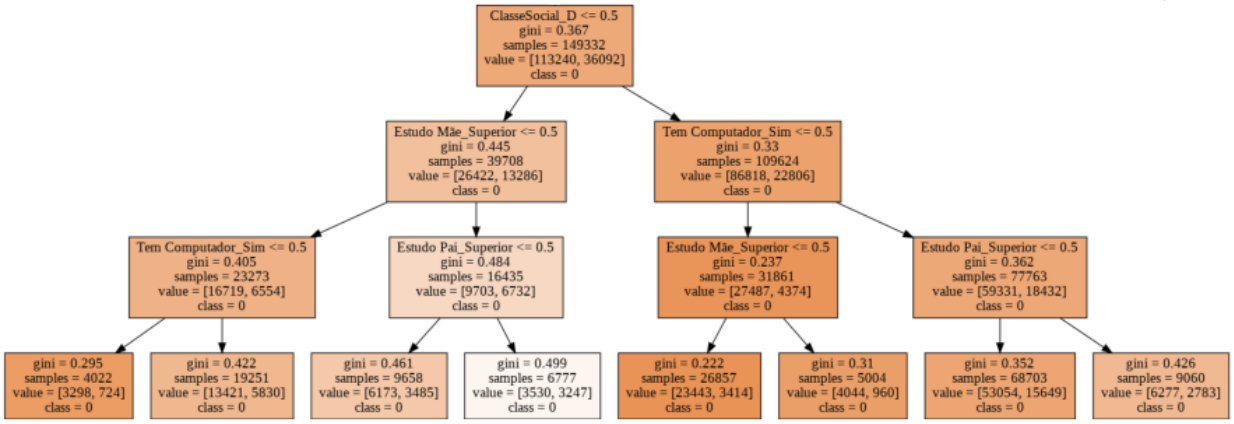


Figura 5.3: Ramo direito central da árvore

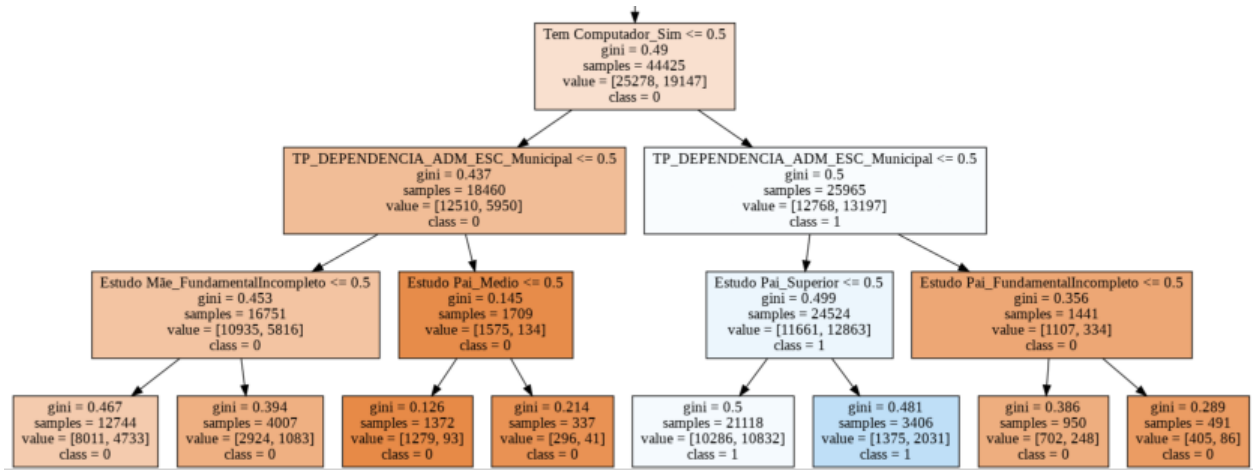


Figura 5.4: Ramo esquerdo central da árvore

tros são azuis, exceto os que pertencem a escola municipais. Reforçando o viés do modelo em classificar alunos de escolas estaduais ou municipais como de baixo desempenho.

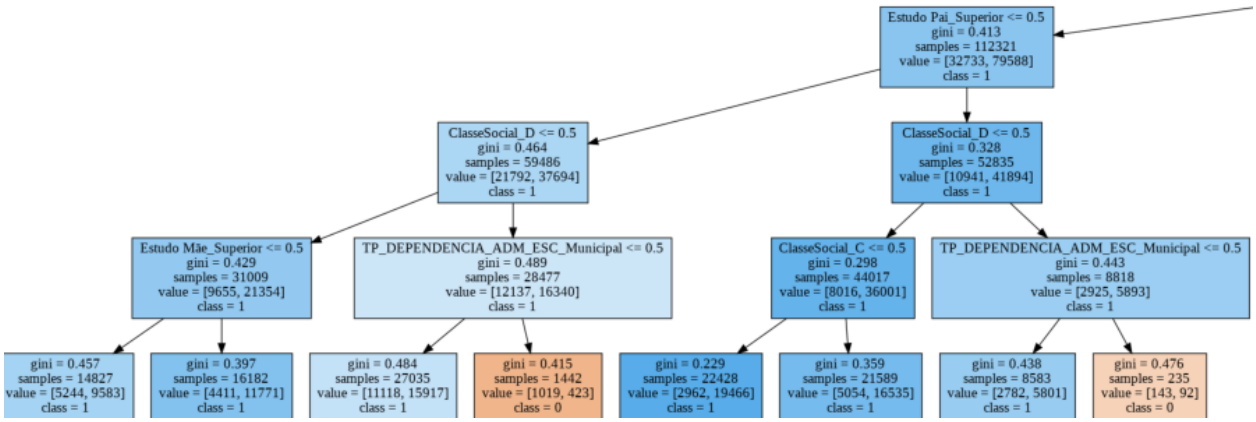


Figura 5.5: Ramo extremo esquerdo da árvore

6 Resultados e Discussões Aprendizado de Máquina Não-Supervisionado

Nesta seção, serão apresentados os resultados da clusterização e da mineração de regras de associação. No final, é feita uma análise discutindo a relação de vários atributos com a nota média. É importante ressaltar que a classificação apresentada no capítulo anterior não tem nenhuma implicação com os resultados dessa seção.

6.1 Grupos identificados na base

Do total de registros, 497.288 foram rotulados como "Cluster A" e 423.300 como "Cluster B". A Figura 6.1 apresenta características da distribuição das notas dos registros de cada grupo nas provas do ENEM.

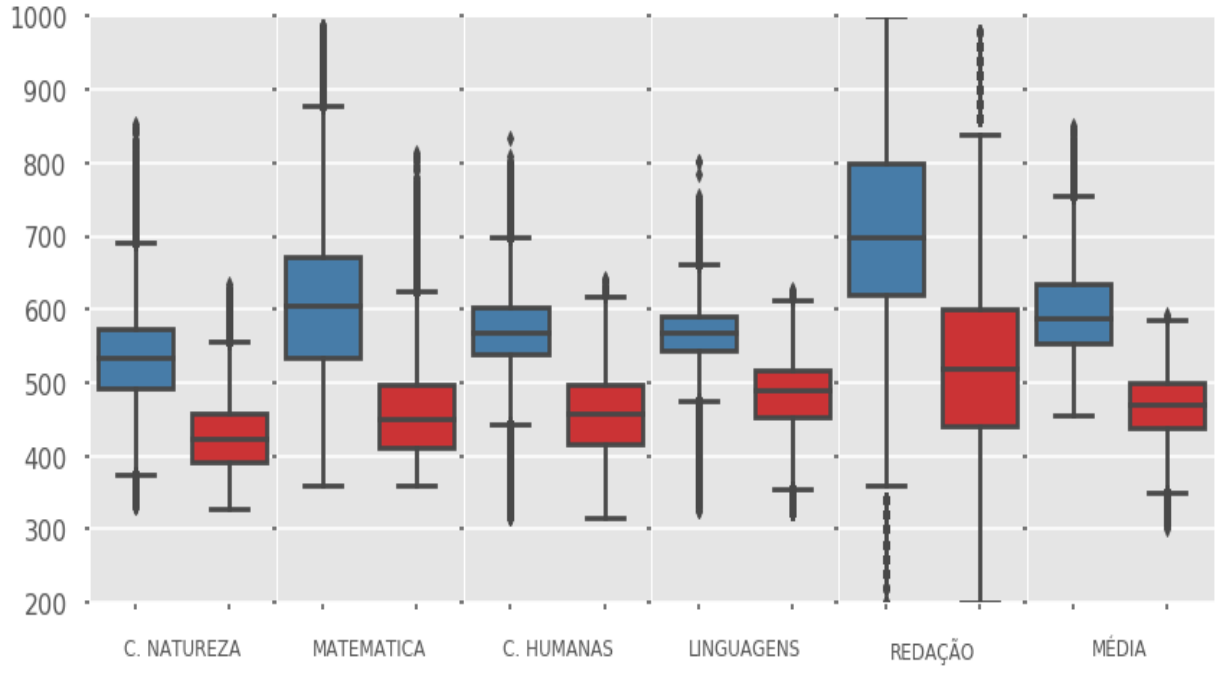


Figura 6.1: Distribuição das notas para os clusters resultantes

A Figura 6.1 apresenta as notas de cada uma das 5 provas e uma coluna que é a Nota Média do aluno. Nesta figura e nas próximas figuras deste texto, a cor vermelha está associada ao Cluster A e a cor Azul ao Cluster B.

No gráfico, é possível constatar que, para todos os eixos, o primeiro quartil do Cluster B é maior que o terceiro quartil do Cluster A, ou seja, a nota de 75% dos alunos do Cluster B é estritamente maior que a de 75% dos estudantes do Cluster A, Na comparação da média, o valor mínimo do Cluster B é muito próximo à mediana do Cluster A. Concluindo a análise, é factível afirmar que os dois grupos tem desempenhos diferentes no exame, sendo que o Cluster B pode ser considerado de bom desempenho e o Cluster A de desempenho insatisfatório.

Neste gráfico, também é possível identificar que Ciências da Natureza, Ciências Humanas e Linguagens têm uma distribuição similar, com os dois clusters próximos. Já em Matemática e Redação, há uma distância maior entre os grupos, indicando que esses atributos têm uma importância maior para definir em qual grupo um registro estará.

Nos próximos parágrafos, são analisadas as informações de inscrição do participante para descrever os clusters de forma socioeconômica. Para isso, é interessante ter algum conhecimento do domínio no geral.

A sociedade brasileira é uma das mais desiguais do mundo, no trabalho (CAMPELLO et al., 2018) é apontado que uma parcela significativa da população é excluída dos direitos básicos e que a renda é um dos principais fatores. Os candidatos do ENEM, por serem dessa sociedade refletem esse contraste. Na base de dados, 76% do total de estudantes possuem renda familiar entre R$ 0 e R$ 2994,00, ou seja, menor ou igual a 3 salários mínimos (dado o ano de 2019). Desta forma, uma grande parte dos candidatos é da Classe E (65%) e da Classe D (21%). A maioria dos candidatos pertencem a escolas estaduais (77%), enquanto menos de 1% dos alunos da base estudam em escola municipal. Aproximadamente 16% cursam em escolas particulares e 5% nas federais. A maioria dos participantes do exame é do sexo feminino. Por isso, é inerente que essas variáveis sejam mais constantes do que outras.

A tabela 6.1 informa para cada cluster o percentual de algumas características relevantes. O Cluster A é caracterizado por ser bastante homogêneo com relação aos atributos socioeconômicos, 93% dos registros cursaram ensino médio em escola estadual e 58% escolheram realizar a prova de espanhol. Além disso, no grupo A, 52%, informou que a mãe não concluiu o ensino médio e 62% afirmou que não possui PC em casa.

No Cluster B, aproximadamente 66% dos registros pertencem às classes econômicas D, C, B e A. Neste grupo, $50,2\%$ se autodeclararam brancos, 70.61% afirmaram possuir PC e 70.14% relataram que a mãe concluiu o ensino médio e 67.31% escolheram realizar a prova de Inglês.

Posto isso, ao observar a tabela 6.1 nota-se uma oposição entre os dois clusters sobre o olhar socieconômico. Para os critérios de Raça, Estudo da Mãe, Língua e Tem Computador o valor mais frequente é discordante quanto ao outro cluster. Assim, é possível afirmar que além da distância no desempenho do exame existem grandes diferenças sociais entre membros dos dois grupos. Esse resultado sugere que candidatos com perfil socioeconômico similar tendem a ter performances equivalentes na prova.

		Cluster A (%)	**Cluster B(%)**
Sexo	Feminino	62	54
	Masculino	38	46
Raça	Branco	29	50
	Não-Branco	69	48
	Não Declarado	02	02
Adm. Escola	Estadual	93	59
	Federal	02	09
	Municipal	01	01
	Particular	05	31
Língua	Espanhol	58	33
	Inglês	42	67
Estudo Mãe	Médio Completo	44	70
	Médio Incompleto	52	28
	Não Sabe	05	02
Tem PC	Não	62	29
	Sim	38	71
Classe Social	A	00	03
	B	01	07
	C	05	20
	D	17	27
	E	77	44

Tabela 6.1: Atributos socioeconômicos divididos pelos clusters

6.2 Regras de Associação

Ao usar regras de associações na base de dados, é possível conhecer relações entre as variáveis socioeconômicas e tipificar os clusters. As métricas de Suporte e Confiança foram limitadas a 20% e e 70%, respectivamente. Todas as regras apresentadas têm *Lift* maior do que 1, ou seja, são positivamente correlacionadas. Como a técnica valoriza a popularidade de aparição de um item, naturalmente atributos mais populares se destacam nas regras, como o estudo em Escola Estadual e Classe E. Devido aos limiares estabelecidos, como os itens no Cluster B são menos frequentes, este cluster tem menos regras nas tabelas.

No total, foram geradas 52 regras. As regras geradas estão divididas nas Tabelas 6.2 e 6.3 e estão ordenadas de forma decrescente pela confiança.

Para explicar como uma regra pode ser interpretada, segue um exemplo com a Regra 16, "Classe=E $\rightarrow$ ADM_ESC=Est", que tem valores 0.556 de Suporte e 0.896 de Confiança. A regra deve ser lida da seguinte maneira: alunos da Classe E e que estudam em escola estadual são aproximadamente 55% do total da base de dados (Suporte) e 89% dos alunos que são da Classe E estudam em escola estadual (Confiança). Ou seja, 11% dos alunos que pertence a Classe E estuda em escola Particular, Federal ou Municipal.

Como a clusterização foi baseada nas notas, é possível ler o rótulo Cluster A ou Cluster B, como um indicativo de desempenho na prova, sendo que os alunos do Cluster A têm notas mais baixas.

Podemos ver algumas regras que derivam imediatamente das porcentagens relatadas na seção anterior: a Regra 10 aponta que 92% dos alunos do Cluster A estudam em escola estadual, a Regra 32 que 77% destes alunos são da Classe E e a Regra 49 mostra que Não possuir computador acarreta em estar no Cluster A, com 70% de confiança.

As regras, no geral, reforçam as características do Cluster A, e é possível entender este cluster a partir da Regra 35, onde, com 75% de confiança, candidatos que são da Classe E, estudam em escola estadual e não possuem computador estão no Cluster A.

Sobre o Cluster B, tem-se as Regras 51 e 52, com aproximadamente os mesmos valores de suporte confiança. É interessante ver que 70% dos candidatos do Cluster B possuem Computador ou a mãe completou o Ensino Médio e estes registros correspondem a aproximadamente 32% da base toda. Com valores de suporte e confiança praticamente

iguais, as Regras 31 e 33 também relacionam as características do Cluster B citadas acima: a Regra 31 mostra que estar no Cluster B e ter computador implica em a mãe ter completado o ensino médio; a Regra 33 reafirma, já que indica que as mães dos alunos do Cluster B que possuem computador concluíram o ensino médio.

	Antecedent → Consequent	**Supp**	**Conf**
1	Classe=E, EstudoMãe=Medio_Inc, Cluster=A → ADM_ESC=Est.	0.231	0.964
2	Classe=E, Cluster=A, TemPC=Não → ADM_ESC=Est.	0.286	0.962
3	EstudoMãe=Medio_Inc, Cluster=A → ADM_ESC=Est.	0.267	0.958
4	Cluster=A, TemPC=Não → ADM_ESC=Est.	0.318	0.956
5	Classe=E, EstudoMãe=Medio_Inc, TemPC=Não → ADM_ESC=Est.	0.224	0.953
6	Classe=E, Cluster=A → ADM_ESC=Est.	0.397	0.950
7	EstudoMãe=Medio_Inc, TemPC=Não → ADM_ESC=Est.	0.245	0.948
8	Classe=E, EstudoMãe=Medio_Inc → ADM_ESC=Est.	0.305	0.934
9	Classe=E, TemPC=Não → ADM_ESC=Est.	0.372	0.933
10	Cluster=A → ADM_ESC=Est.	0.500	0.926
11	ADM_ESC=Est., EstudoMãe=Medio_Inc, TemPC=Não → Classe=E	0.224	0.915
12	TemPC=Não → ADM_ESC=Est.	0.426	0.910
13	EstudoMãe=Medio_Inc → ADM_ESC=Est.	0.370	0.909
14	EstudoMãe=Medio_Inc, TemPC=Não → Classe=E	0.235	0.909
15	ADM_ESC=Est., Cluster=A, TemPC=Não → Classe=E	0.286	0.898
16	Classe=E → ADM_ESC=Est.	0.556	0.896
17	Cluster=A, TemPC=Não → Classe=E	0.298	0.893
18	EstudoMãe=Medio_Comp, Cluster=A → ADM_ESC=Est.	0.209	0.883
19	ADM_ESC=Est., TemPC=Não → Classe=E	0.372	0.872
20	EstudoMãe=Medio_Inc, TemPC=Não → Classe=E, ADM_ESC=Est.	0.224	0.867
21	ADM_ESC=Est., EstudoMãe=Medio_Inc, Cluster=A → Classe=E	0.231	0.866
22	EstudoMãe=Medio_Inc, Cluster=A → Classe=E	0.240	0.861
23	Cluster=A, TemPC=Não → Classe=E, ADM_ESC=Est.	0.286	0.859
24	TemPC=Não → Classe=E	0.398	0.851
25	Classe=E, EstudoMãe=Medio_Comp → ADM_ESC=Est.	0.224	0.844

Tabela 6.2: Regras de Associação- Parte 1

	Antecedent → Consequent	**Supp**	**Conf**
26	EstudoMãe=Medio_Inc, Cluster=A → Classe=E, ADM_ESC=Est.	0.231	0.830
27	ADM_ESC=Est., EstudoMãe=Medio_Inc → Classe=E	0.305	0.825
28	EstudoMãe=Medio_Inc → Classe=E	0.327	0.804
29	ADM_ESC=Est., Cluster=A → Classe=E	0.397	0.795
30	TemPC=Não → Classe=E, ADM_ESC=Est.	0.372	0.793
31	EstudoMãe=Medio_Comp, Cluster=B →TemPC=Sim	0.251	0.779
32	Cluster=A → Classe=E	0.418	0.774
33	Cluster=B, TemPC=Sim → EstudoMãe=Medio_Comp	0.251	0.774
34	Classe=E, ADM_ESC=Est., TemPC=Não → Cluster=A	0.286	0.770
35	Classe=E, ADM_ESC=Est., EstudoMãe=Medio_Inc → Cluster=A	0.231	0.757
36	EstudoMãe=Medio_Inc → Classe=E, ADM_ESC=Est.	0.305	0.750
37	Classe=E, TemPC=Não → Cluster=A	0.298	0.747
38	ADM_ESC=Est., TemPC=Não → Cluster=A	0.318	0.747
39	Cluster=A → Classe=E, ADM_ESC=Est.	0.397	0.736
40	Classe=E, EstudoMãe=Medio_Inc → Cluster=A	0.240	0.733
41	Classe=E, ADM_ESC=Est., EstudoMãe=Medio_Inc →TemPC=Não	0.224	0.733
42	ADM_ESC=Est., EstudoMãe=Medio_Inc → Cluster=A	0.267	0.721
43	Classe=E, ADM_ESC=Est., Cluster=A →TemPC=Não	0.286	0.720
44	ADM_ESC=Est. → Classe=E	0.556	0.720
45	Classe=E, EstudoMãe=Medio_Inc → TemPC=Não	0.235	0.718
46	Classe=E, TemPC=Não → ADM_ESC=Est., Cluster=A	0.286	0.718
47	Classe=E, ADM_ESC=Est. → Cluster=A	0.397	0.714
48	Classe=E, Cluster=A →TemPC=Não	0.298	0.712
49	TemPC=Não → Cluster=A	0.333	0.711
50	Classe=E, EstudoMãe=Medio_Inc → ADM_ESC=Est., Cluster=A	0.231	0.706
51	Cluster=B →TemPC=Sim	0.325	0.706
52	Cluster=B → EstudoMãe=Medio_Comp	0.322	0.701

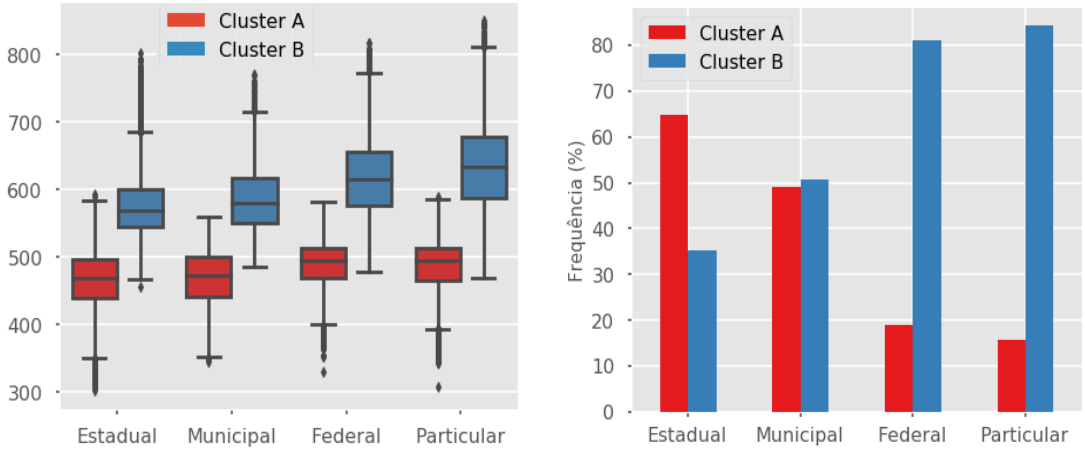
Tabela 6.3: Regras de Associação- Parte 2

7 Uma visão sobre a desigualdade no ENEM

Nesta seção serão analisadas, de forma mais aprofundada, algumas variáveis socieconômicas e a relação delas com o desempenho do estudante. No trabalho (SILVA et al., 2020), essa análise foi feita para os candidatos de Minas Gerais. Esta seção amplia o estudo e mostra como o resultado relativo a Minas Gerais foi representativo para o país.

No boxplot da Figura 7.1a, é possível observar a distribuição da nota média pelos diversos tipos administrativos de escola. Já a Figura 7.1b, apresenta a população dos estudantes de cada tipo administrativo nos clusters. Nos dois gráficos, os dados de alunos de escolas federais e escolas particulares é similar, mostrando que as escolas federais alcançam um bom desempenho no exame. É interessante ressaltar que as escolas federais são públicas, e que o financiamento é feito pelo governo federal. Esses institutos são caracterizados pela sua autonomia da gestão e alta capacitação docente e para isso possuem um alto volume de investimentos.

Ao analisarmos a renda dos candidatos, é possível observar na Figura 7.2 que, quanto maior a renda, maior a probabilidade de um estudante estar no Cluster B. Apontando que a renda familiar tem impacto expressivo no desempenho do aluno. Informações que não estão na base de dados, tal como, se o candidato realizou cursos preparatórios ou cursos de língua estrangeira e se o candidato trabalha permitiram construir uma análise



(a) Nota Média por tipo de escola

(b) População nos clusters por tipo de escola

Figura 7.1: Distribuição da nota média e entre clusters de acordo com tipo de escola

Língua	Total (%)	Média Renda Familiar (R$)	Desvio Padrão (R$)
Espanhol	46%	2065	2187
Inglês	53%	3802	4348

Tabela 7.1: Renda por Língua Estrangeira escolhida

mais aprofundada do impacto da Renda no desempenho.

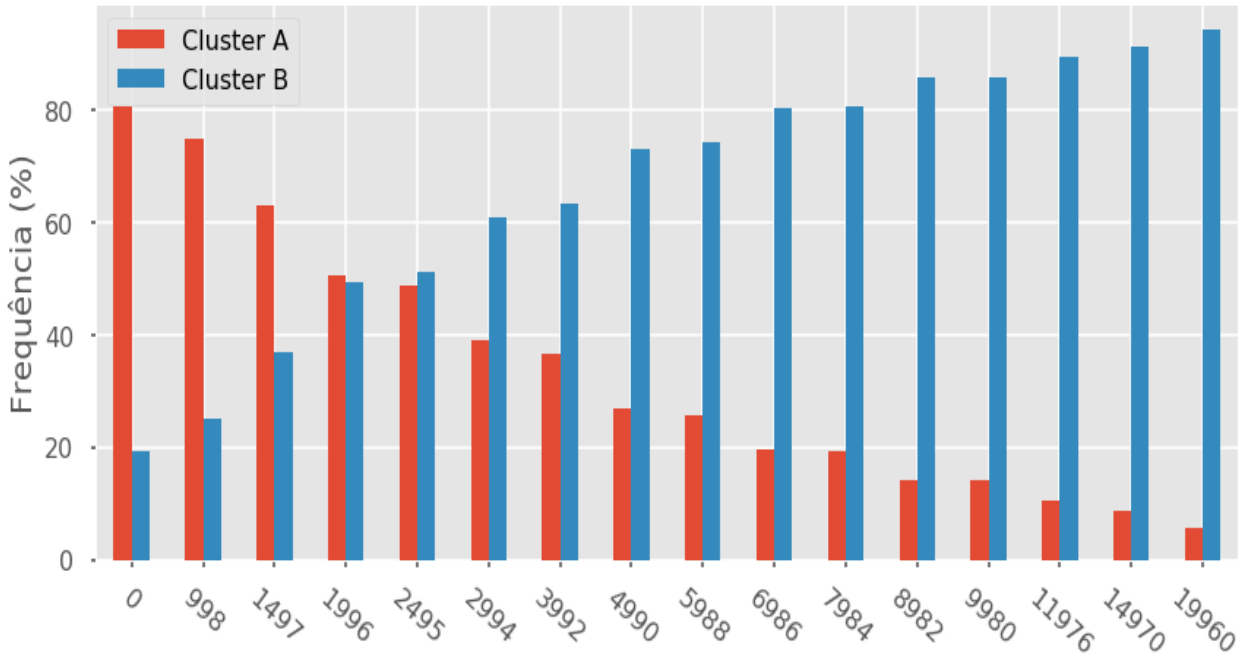


Figura 7.2: *Features* Distribuição de cada cluster por renda

Com relação à escolha da língua da prova estrangeira, o gráfico da Figura 7.3 Cluster B apresenta, em sua maioria, alunos que escolheram inglês. Uma razão para o alto número de alunos que escolhe espanhol, considerando que esta língua não faz parte do currículo, pode ser a de que alunos que escolhem inglês têm um maior costume ou proficiência nesta língua, enquanto quem escolhe a outra opção tem como motivação a similaridade com a língua portuguesa. Assim, podemos ver que há uma relação na escolha da língua no desempenho. Considerando que a prova de linguagens tem 45 perguntas e 5 são de língua estrangeira, um aluno com domínio do tópico poderia acertar 11% desta prova. Contudo, esse desempenho insatisfatório na prova de linguagens não justifica uma nota ruim no geral. Outro fator a ser considerado é a de que alunos que escolhem inglês tem a renda, na média, 1737 reais maior do que aqueles que escolhem espanhol, a tabela 7.1 resume os valores.

No gráfico da Figura 7.4b, é fácil verificar que a maioria dos participantes auto-

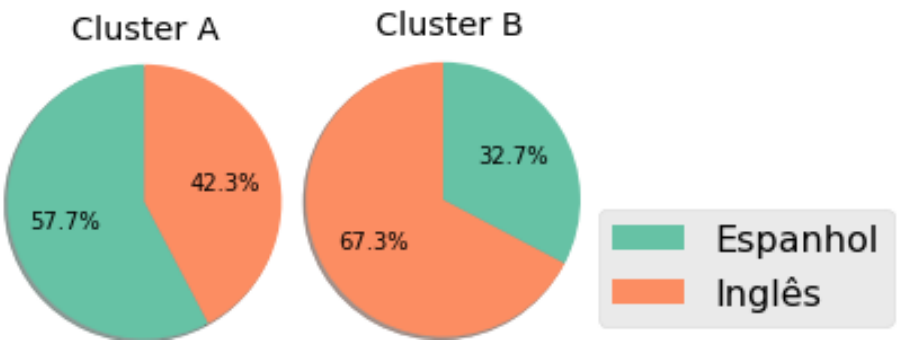
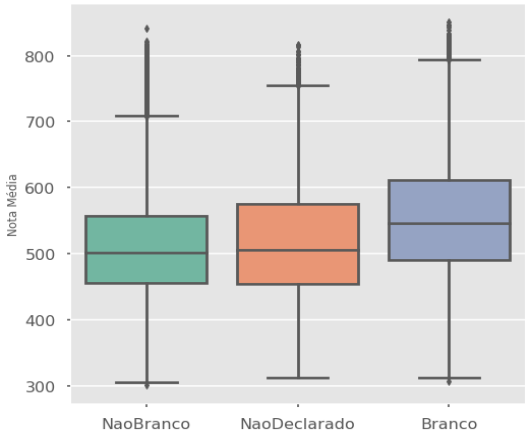
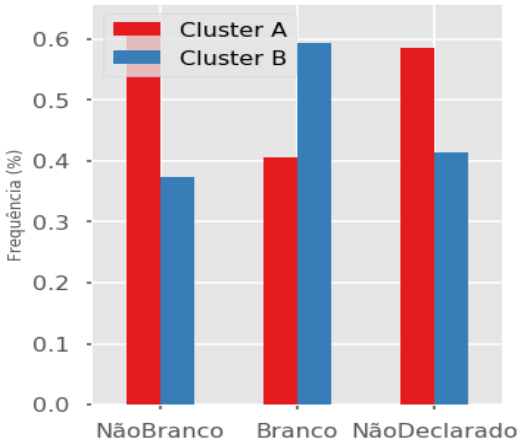


Figura 7.3: Distribuição dos estudantes nos clusters pela língua estrangeira escolhida



(a) Nota Média por Raca Autodeclarada



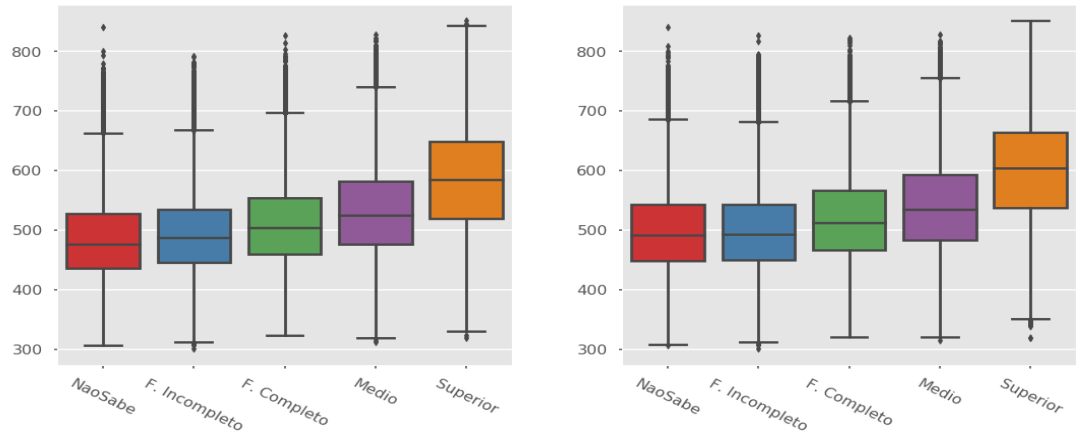
(b) População nos clusters por raça autodeclarada

Figura 7.4: Distribuição da nota média e entre clusters de acordo com a raça autodeclarada

declarados brancos estão no Cluster B e que a maioria dos participantes autodeclarados não-brancos ou que não fizeram autodeclaração estão no Cluster A. Dito isso, no boxplot da Figura 7.4a, é possível ver que a mediana dos participantes autodeclarados brancos e não-brancos têm uma distância de aproximadamente 50 pontos.

Nas Figuras 7.5a e 7.5b, verifica-se a distribuição da nota média pelos vários níveis de escolaridade dos pais. A forma que as notas se distribuem pelos diversos níveis escolares é similar com relação a mãe e ao pai. Vale destacar que a mediana do grupo referente ao ensino superior é superior ao terceiro quartil do grupo relativo ao ensino médio. De fato, mais de 50% dos inscritos com pai ou mãe que cursou o ensino superior têm nota maior que 75% dos estudantes de cada uma das outras categorias.

De forma geral, ao compararmos os resultados desta seção com (SILVA et al., 2020), podemos identificar que atributos socieconômicos se comportam para todos os registros do Brasil de forma equivalente a Minas Gerais, mostrando como os resultados



(a) Nota Média por Escolaridade da Mãe (b) Nota Média por Escolaridade do Pai

Figura 7.5: Distribuição da nota média de acordo com a escolaridade do Pai e da Mãe

daquele trabalho são significativos.

8 Conclusão e Trabalhos Futuros

O acesso ao ensino superior é o sonho de milhões de jovens, e realizar uma boa prova do ENEM é fundamental para alcançar essa meta. Entretanto, as desigualdades sociais impõem barreiras invisíveis à maior parte desses jovens. Por isso, analisar o desempenho no exame a partir de uma perspectiva socioeconômica é importante para fundamentar discussões sobre como impedir a manutenção dessa desigualdade.

A pesquisa apresentada é uma continuação da relatada em (SILVA et al., 2020), em que foi desenvolvida uma proposta de descoberta de conhecimento na base de dados do ENEM 2019 para concluintes do ensino médio das escolas de Minas Gerais. Com o objetivo de auxiliar processos de tomadas de decisão, a partir da identificação e compreensão de como as desigualdades afetam os alunos de ensino médio que prestam o exame, neste trabalho de conclusão de curso, a base de dados foi expandida para compreender todos os concluintes do Brasil. Nesta proposta de descoberta de conhecimento, foram utilizadas técnicas de classificação, clusterização e mineração de regras de associação.

Com relação ao aprendizado supervisionado, foram aplicadas Árvores de Decisão para classificar candidatos com relação a alto e baixo desempenho no exame baseado nos dados da inscrição. A seleção de atributos mostrou que características como Tipo de Administrativo da Escola, Renda e Estudo dos pais são as que mais influenciam no desempenho do aluno. Os resultados mostram a dificuldade do modelo em decidir quando os candidatos terão bom desempenho. Trabalhos futuros neste tópico envolvem a utilização de técnicas para melhorar a qualidade da predição da classe minoritária, como Florestas Aleatórias (BIAU; SCORNET, 2016) e técnicas de criação sintética de amostras como SMOTE (CHAWLA et al., 2002).

A clusterização foi feita utilizando *K-means* para agrupar a base de dados a partir dos atributos referentes às notas dos candidatos nos cinco eixos que compõe o exame. O número de de grupos definido foi 2, pois, para este valor, foi encontrado a melhor medida de silhueta. De forma geral, é possível afirmar que um grupo é de candidatos que tiveram um baixo desempenho (Cluster A) e outro de estudantes que tiveram um alto

desempenho (Cluster B). Ao analisar os atributos socieconômicos dos grupos no Cluster A, existe um predomínio de alunos com características socioeconômicas similares entre si, destacando-se a educação em rede estadual de ensino, a baixa renda familiar e o fato de que a mãe não completou o ensino médio. Em contrapartida, no Cluster B, de alunos com alto desempenho, observa-se a presença de alunos com rendas mais altas, estudo nas redes particular ou federal de ensino e cuja mãe completou ao menos o ensino médio. Trabalhos futuros neste tópico envolvem a utilização de outras métricas para avaliação da clusterização, como avaliações externas (RENDÓN et al., 2011).

Neste trabalho, foi discorrido que o desempenho de alunos de escolas federais no ENEM é similar a de alunos em escolas particulares. Levantando o debate de que é necessário analisar o modelo de gestão dessas escolas para identificar o que é possível replicar em outras redes públicas de ensino. Em trabalhos futuros, a base do censo escolar pode ser explorada para verificar as características comuns entre escolas particulares e federais.

Também podem ser construídas outras análises comparando múltiplas unidades da federação de forma a verificar se algumas características socioeconômicas se comportam diferente em cada região.

Por último, vale dizer que dados importantes para entender o desempenho dos alunos foram retirados dos questionários socioeconômicos nas últimas edições e isso é uma limitação da pesquisa. Como é relatado em (FRANCO et al., 2020) as principais informações retiradas são: sobre o ensino fundamental do candidato, sobre a realização de cursos de língua estrangeira, interesse em política, e o quão motivado ele está para o próximo ciclo de estudo.

Bibliografia

ADEODATO, P. J. Data mining solution for assessing brazilian secondary school quality based on enem and census data. In: *Proc. 13° CONTECSI*. [S.l.: s.n.], 2016. p. 2658–2679.

ADEODATO, P. J.; FILHO, R. L. S. Where to aim? factors that influence the performance of brazilian secondary schools. In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*. [S.l.: s.n.], 2020.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499.

BARROS, R. P. d. et al. Determinantes do desempenho educacional no brasil. Instituto de Pesquisa Econômica Aplicada (IPEA), 2001.

BIAU, G.; SCORNET, E. A random forest guided tour. *Test*, Springer, v. 25, n. 2, p. 197–227, 2016.

CAMPELLO, T. et al. Faces da desigualdade no brasil: um olhar sobre os que ficam para trás. *Saúde em Debate*, SciELO Public Health, v. 42, p. 54–66, 2018.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

COSTA, E. et al. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, v. 1, n. 1, p. 1–29, 2013.

DEMVsAR, J. et al. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, v. 14, p. 2349–2353, 2013. Disponível em: ⟨http://jmlr.org/papers/v14/demsar13a.html⟩.

ESTÉVEZ, P. A. et al. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, IEEE, v. 20, n. 2, p. 189–201, 2009.

FILHO, R. L. S.; ADEODATO, P. J. Data mining solution for assessing the secondary school students of brazilian federal institutes. In: IEEE. *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2019. p. 574–579.

FRANCO, J. J. et al. Usando mineração de dados para identificar fatores mais importantes do enem dos últimos 22 anos. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 1112–1121.

GOMES, T.; GOUVEIA, R.; BATISTA, M. Dados educacionais abertos: associações em dados dos inscritos do exame nacional do ensino médio. In: *Anais do Workshop de Informática na Escola*. [S.l.: s.n.], 2017. v. 23, n. 1, p. 895.

HAN, J. et al. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, Springer, v. 8, n. 1, p. 53–87, 2004.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

INEP. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Microdados do Enem por Escola*. 2019. Http://portal.inep.gov.br/web/guest/microdados. Online: acessado 03 Julho 2020.

INEP. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Microdados do ENEM 2019*. 2020. ⟨http://portal.inep.gov.br/web/guest/microdados⟩. Online: acessado 03 Julho 2020.

JR, E. B. Questões epistemológicas em mineração de dados educacionais. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2019. v. 30, n. 1, p. 1541.

LEONI, R. C.; SAMPAIO, N. Desempenho das escolas públicas e privadas da região do vale do paraíba: Uma aplicação da técnica de agrupamentos kmeans com base nas variáveis do enem 2015. *Cadernos do IME-Série Estatística*, v. 42, p. 31, 2017.

LIMA, P. d. S. N. et al. Análise de dados do enade e enem: uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior*, Universidade de Sorocaba, v. 24, n. 1, p. 89–107, 2019.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PRATI, R. et al. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008.

RENDÓN, E. et al. Internal versus external cluster validation indexes. *International Journal of computers and communications*, v. 5, n. 1, p. 27–34, 2011.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003.

RODRÍGUEZ, P. et al. Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, Elsevier, v. 75, p. 21–31, 2018.

Silva Filho, R. L. C.; Adeodato, P. J. L. Data mining solution for assessing the secondary school students of brazilian federal institutes. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2019. p. 574–579.

SILVA, L. A.; MORINO, A. H.; SATO, T. M. C. Prática de mineraçao de dados no exame nacional do ensino médio. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2014. v. 3, n. 1, p. 651.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. [S.l.]: Elsevier Brasil, 2017.

SILVA, V. A. A. da et al. Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no enem 2019 utilizando mineração de dados. In: *SBIE 2020 - Trilha 1 ()*. [S.l.: s.n.], 2020.

SIMON, A.; CAZELLA, S. Mineração de dados educacionais nos resultados do enem de 2015. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, n. 1, p. 754.

SOUZA, P. F. d.; RIBEIRO, C. A. C.; CARVALHAES, F. Desigualdade de oportunidades no brasil: considerações sobre classe, educação e raça. *Revista Brasileira de Ciências Sociais*, SciELO Brasil, v. 25, n. 73, p. 77–100, 2010.

Sumaiya Thaseen, I.; Aswani Kumar, C. Intrusion detection model using fusion of chi-square feature selection and multi class svm. *Journal of King Saud University - Computer and Information Sciences*, v. 29, n. 4, p. 462–472, 2017. ISSN 1319-1578. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S1319157816300076⟩.

WASKOM, M.; TEAM the seaborn development. *mwaskom/seaborn*. Zenodo, 2020. Disponível em: ⟨https://doi.org/10.5281/zenodo.592845⟩.