



ANÁLISE DO DESEMPENHO DE PESSOAS COM DEFICIÊNCIA QUE PRESTARAM O EXAME NACIONAL DO ENSINO MÉDIO - ENEM

Performance Analysis of People with Disabilities at the National High School Examination - ENEM

Caue Gomes de Oliveira*

Regina Barwaldt**

Giancarlo Lucca***

Resumo: Atualmente o ingresso no ensino superior acontece de forma majoritária pela realização do Exame Nacional do Ensino Médio (ENEM), que chega a contar com cinco milhões de participantes por ano. Essa prova trouxe facilidade para cursar uma universidade, seja ela pública ou privada, porém, para pessoas com deficiência, ser um universitário ainda é um grande obstáculo. É apresentado nesse artigo uma análise dos participantes que possuem deficiências físicas e psicológicas que prestaram o ENEM, por meio de um modelo preditivo, com o intuito de auxiliar na compreensão do desempenho final do candidato. Utilizando técnicas de mineração de dados aplicado na base de dados referente à prova realizada em 2018, fez-se o uso de um algoritmo de aprendizagem de máquina, o C4.5, que gera um padrão interpretável através de uma árvore de decisão, para predizer qual foi o desempenho final do candidato, de acordo com algumas características, tais como: socioeconômicas, demográficas, étnicas, faixa-etária e formação acadêmica, verificando se tais atributos influenciam no rendimento final.

Palavras-chave: Mineração de dados educacionais. ENEM. C4.5. Pessoas com deficiência.

Abstract: Currently, admission to higher education takes place mainly by conducting the National High School Examination (ENEM), which reaches five million participants per year. This test made it easy to attend a university, be it public or private, however, for disabled people, being a university student is still a major obstacle. This article presents an analysis of the participants who have physical and psychological disabilities who provided the ENEM, through a predictive model, in order to assist in understanding the candidate's final performance. Using data mining techniques applied in the database related to the test held in 2018, a machine learning algorithm, C4.5, was used, which generates an interpretable pattern through a decision tree, to predict what was the candidate's final performance, according to some characteristics, such as: socioeconomic, demographic, ethnic, age group and academic background, checking whether such attributes influence the final performance.

Keywords: Mining of educational data. ENEM. C4.5. Disabled people.

* Bacharel em Engenharia de Computação pela Universidade Federal do Rio Grande (FURG) – Campus Carreiros. Orcid: <https://orcid.org/0000-0001-7980-5482>. E-mail: caue@furg.br

** Doutora em Informática na Educação pela Universidade Federal do Rio Grande do Sul (UFRGS). Orcid: <https://orcid.org/0000-0002-0382-3892>. E-mail: reginabarwaldt@furg.br

*** Doutor em Ciência da Computação pela Universidad Pública de Navarra (UPNA). Orcid: <https://orcid.org/0000-0002-3776-0260>. E-mail: giancarlo.lucca@furg.br



1 Introdução

O sistema global de redes de computadores, conhecido como internet, torna-se cada vez mais presente na vida das pessoas. A pesquisa TIC Domicílios (LAVADO, 2019), que mediu o uso das Tecnologias de Informação e Comunicação em relação ao acesso a internet no ano de 2018, pela população brasileira, mostra que 126.9 milhões a acessam regularmente. Logo, seu uso diário em redes sociais, jogos e aplicativos em geral, causam um crescimento significativo na quantidade de registros produzidos. Um relatório desenvolvido por *Digital Universe* (IDC, 2014), no ano de 2014, retrata a realidade da quantidade de dados produzida, constatando que em 2014 haviam 4,4ZB (Zettabytes, equivalente a 1×10^{12} Gigabytes) de informação, com estimativa de que em 2020 existissem 44ZB, um aumento considerável.

Pensando em extrair algum conhecimento destes conjuntos de dados, surgiu uma metodologia que é chamada de Descoberta de Conhecimento em Banco de Dados (DCBD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), ou em inglês *Knowledge-Discovery in Databases* (KDD), que se trata de um processo de extração de informações com algumas etapas.

Um dos passos do KDD é a Mineração de Dados (MD) (TAN; STEINBACH; KUMAR, 2005), que tem como objetivo analisar registros, extraindo conteúdo útil por meio de algoritmos que criam um modelo com base nos dados. A MD é utilizada em diversos ramos, sejam eles comerciais, empresariais, automobilísticos, tecnológicos, enfim, empresas que contêm informação excessiva, seja ela estruturada (uma estrutura planejada, com os dados em formato padrão e organizados em tabelas, por exemplo) ou não estruturada (onde estão armazenados sem estrutura, estando em vários padrões, como vídeo, som, texto ou números, e em arquivos que não estejam organizados em colunas ou linhas).

Utilizar MD e tornar importantes as informações consideradas obsoletas fez com que esse processo se tornasse relevante, gerando um investimento alto por parte de empresas. A compreensão dos dados adquirido por aplicar MD fornece um ganho para quem a está aplicando, como, por exemplo, descobrir o público alvo para aplicar técnicas de marketing específicas e aumentar o lucro, inovar produtos que já não estão mais vendendo como antes ou fazer correlações entre produtos para influenciar o consumidor a realizar mais compras, com um objetivo final de poupar custos e alavancar os negócios. Na área da educação essa técnica pode ser aplicada também, e o termo é definido por Mineração de Dados Educacionais (MDE) (SCHUNK, 1996), que visa ao desenvolvimento de métodos para explorar registros oriundos de contextos educacionais.

Neste trabalho, a mineração de dados é aplicada numa base de dados (INEP, 2017) do Exame Nacional do Ensino médio (ENEM) (ENEM, 2019) do ano de 2018, fornecida pelo Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (INEP) (INEP, 2019). Lembrando que os algoritmos responsáveis por predizer resultados baseando-se em padrões aprendidos fazem parte da técnica de Aprendizagem de Máquina (*machine learning*) (MICHIE et al., 1994).

O ENEM é um teste amplamente utilizado e no ano de 2018 contou com 6.774.891 milhões de inscritos (INEP, 2018). Dividido em cinco áreas, a média final dessa prova oportuniza ao estudante a chance de se inscrever em universidades públicas e privadas, em todo o território nacional, contando com prestígio internacional, sendo até mesmo utilizado em algumas universidades no exterior, como, por exemplo, em Portugal, Inglaterra, França, Irlanda e Canadá (MARCOS, 2019). Cada país, entretanto, conta com suas próprias exigências, podendo levar em conta a fluência em inglês, histórico escolar, atividades extracurriculares, a necessidade de uma prova local, e etc.

De acordo com o Recenseamento Demográfico (CENSO) de 2010 (CENSO, 2010), o Brasil apresentava cerca de 45 milhões de deficientes, sendo 23,9% da população total. Do total, apenas 6,66% consegue ter um diploma de ensino superior. De acordo com Viegas (2016), em 2014 apenas 0,42% de um total de 33.377 participantes deficientes conseguiram ingressar no ensino superior através do ENEM e de vestibulares próprios. Para comparação, em 2010 cerca de 8,31% da população total de brasileiros contava com um diploma de ensino superior (CENSO, 2010), e em 2016 pode-se notar um aumento de quase o dobro, com o novo índice sendo de 15,3% (PARRODE, 2017).

Tendo em consideração o que foi exposto, neste artigo o objetivo apresentado foi aplicar uma técnica de Mineração de dados nos registros do ENEM do ano de 2018, disponibilizados pelo INEP, visando a analisar exclusivamente o público deficiente. Desses registros uma técnica de aprendizado de máquina foi utilizada para aprender padrões. E por fim, prever qual foi o desempenho final de um candidato, de acordo com suas características, tais como: socioeconômicas, demográficas, étnicas, faixa-etária e formação acadêmica, verificando se tais atributos influenciam no rendimento final.

Outros estudos realizados neste tema podem também acrescer o conhecimento sobre o assunto, como o artigo aprovado publicado na Escola Regional de Alto Desempenho (ERAD), escrito por (SANTOS et al., 2019b), no qual se realizou e uma análise do perfil dos candidatos em geral no ENEM de 2016, e o artigo publicado na *Frontiers in Education* (FIE), escrito por (SANTOS et al., 2019a), que se trata de um estudo do perfil estudantil geral ao longo de todas as edições do ENEM, de 1998 até 2017.

2 Objetivo

O objetivo geral do trabalho foi analisar o desempenho de candidatos com deficiência que realizaram a prova do ENEM em 2018, por meio de algoritmos de descoberta e conhecimento em base de dados, fundamentado em características socioeconômicas, demográficas, étnicas, faixa-etária e formação acadêmica, verificando se tais atributos influenciam no rendimento final.

3 Justificativa

Pensou-se neste estudo quando a lei Nº 13.409 foi lançada no final do ano de 2016, alterando a lei nº 12.711 que tratava de cotas. Esta alteração favoreceu o ingresso de pessoas deficientes ao nível superior, entretanto, a parcela de deficientes em cursos superiores ainda é pequena. Buscando entender quais são as suas dificuldades, objetivou-se a criação de um trabalho a fim de auxiliar a compreensão de seu desempenho final no ENEM. Conhecendo o problema, poderá ser estudada uma solução futuramente, para então realizar uma inclusão significativa dessas pessoas ao ensino superior.

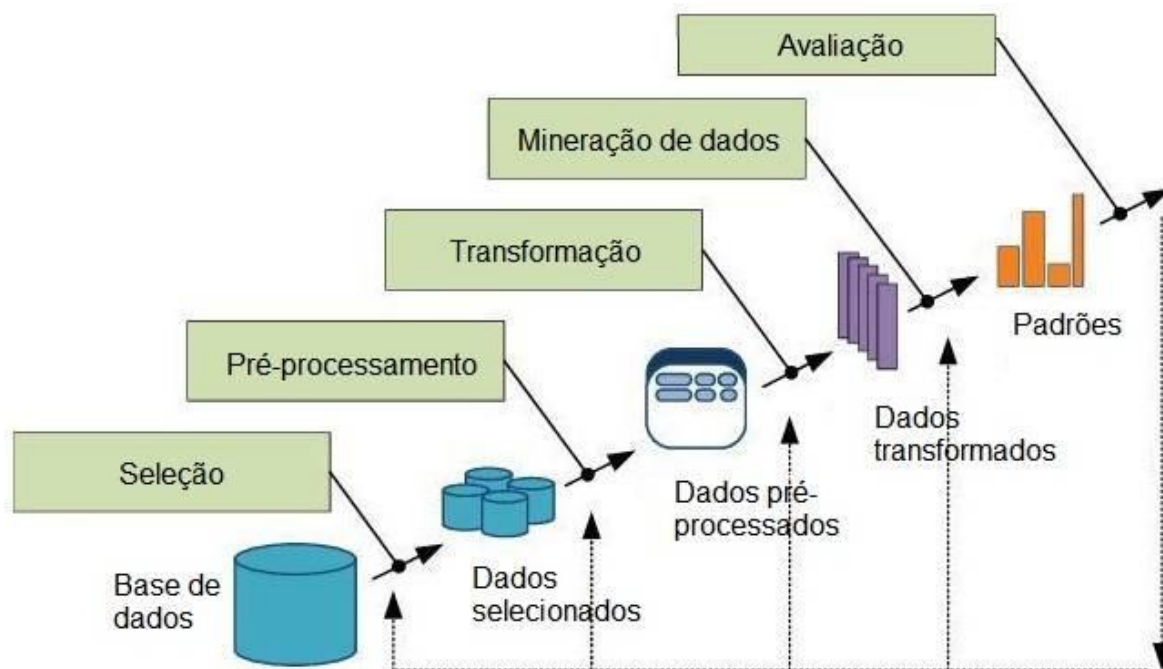
4 Metodologia

Para analisar e pré-processar essa base de dados, foram utilizados códigos desenvolvidos em linguagem Python (ROSSUM, 1995), histogramas produzidos pelo Microsoft Excel (2007) (MICROSOFT, 2007), e o pacote de software WEKA, onde é utilizado Feature Selection por meio do método Ranker, verificando o ganho de informação de todos os atributos, e também é desenvolvido um modelo de Classificador, utilizando o J48 versão 14508 (uma implementação em Java do C4.5) que gera uma árvore de decisão.



As etapas do trabalho podem ser vistas na Figura 1 a seguir, que apresenta as etapas do processo de Descoberta e Conhecimento em Bases de Dados, que engloba a técnica de Mineração de Dados.

Figura 1 – Etapas do processo de DCBD



Fonte: Adaptada de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

5 Análise dos Atributos

A base de dados utilizada é referente ao ano de 2018. Vale destacar que esses são os mais recentes disponibilizados no site do INEP (INEP, 2019). O tamanho do arquivo é de 3,3GB, e ele contém 5.513.748 instâncias, sendo a primeira um cabeçalho explicativo. Existem 137 colunas, os chamados atributos de candidatos.

A base de dados do ENEM separa os atributos em 9 módulos, são eles:

- Dados do Participante
- Dados do questionário Socioeconômico
- Dados da Redação
- Dados da Prova Objetiva
- Dados da Escola
- Dados do Local de Aplicação da Prova
- Dados dos Pedidos de Atendimento Específico
- Dados dos Pedidos de Atendimento Especializado
- Dados dos Pedidos de Recursos Especializados e Específicos para Realização das Provas.

Cada um desses módulos trata de assuntos diferentes relacionados à prova, candidato, escola, questionário, entre outros dados. Para melhor compreendê-los, foram realizadas alterações, começando então a primeira etapa da mineração de dados que é o pré-processamento.



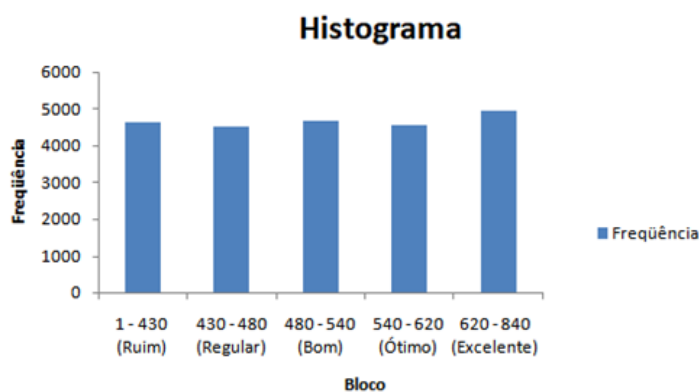
O pré-processamento de dados teve diferentes técnicas. Neste trabalho serão realizadas a filtragem, discretização, remoção e seleção dos dados. Foi iniciado com a filtragem, como o estudo tem o objetivo de analisar somente os candidatos deficientes, foi desenvolvido um código em Python para tal. O código analisa todas as colunas da base de dados que continha informação sobre qual deficiência o participante tem, e aqueles que não têm nenhuma deficiência são removidos do novo arquivo gerado. Após esse procedimento, o arquivo gerado contém 31.988 instâncias, ou seja, 31.987 candidatos que realizaram a prova são deficientes. Em porcentagem temos este número equivale a 0,58% do total.

Desse total de candidatos, existem aqueles que não realizaram a prova, e aqueles que compareceram em apenas um dia, zerando duas áreas de conhecimento. Esses candidatos também foram removidos da nossa base de dados, pois suas notas seriam baixas, o que acabaria afetando o resultado final. No fim, constatou-se que 7.372 participantes deficientes zeraram e que 1.344 compareceram em apenas um dia. Restaram então 23.270 que realizaram as duas provas.

Após a filtragem dos dados de interesse (candidatos com algum tipo de deficiência) estar correto, o próximo passo foi discretizar as classes que serão analisadas, ou seja, dividindo seu domínio em um número pré-determinado de intervalos iguais, e esse método foi realizado no Excel. Essa discretização foi realizada em alguns atributos. Precisamente os atributos: Média Final, Idade, Cor da Pele e Classe Econômica. Foram escolhidos esses atributos, pois continham um domínio de dados expressivo, considerando-se previamente que seriam aspectos interessantes para a análise dos dados.

Então, as primeiras colunas analisadas foram referentes às notas, que, no caso, foram os atributos mais importantes neste trabalho, pois a média final das notas é o atributo que queremos prever neste trabalho. Então, as notas individuais servem apenas para o cálculo da média, ou seja, a soma das cinco áreas de conhecimento divididas por cinco. Assim, foi criada uma nova coluna manualmente na base de dados. Com a média calculada, desenvolveu-se um histograma para verificar a divisão de suas faixas de intervalo, e pode-se notar que apenas 21% do total tirou acima de 620, considerado a classe Excelente nesse trabalho. O histograma se encontra na Figura 2 a seguir.

Figura 2 – Dados do atributo “média final” depois da discretização, em cinco grupos

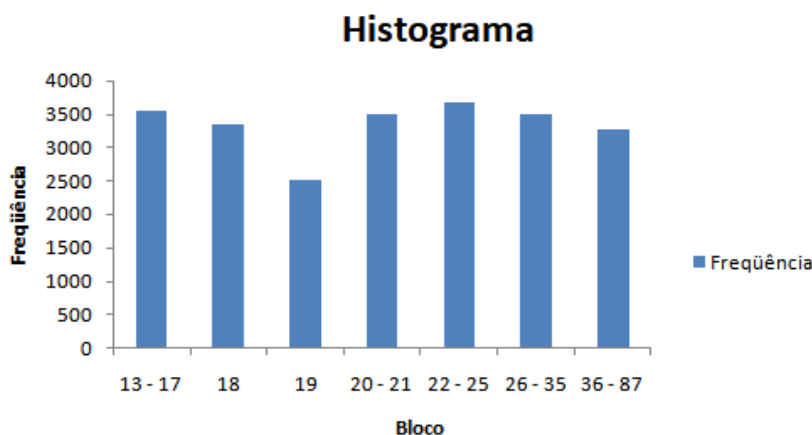


Fonte: Autor.

Então, partiu-se para a discretização das idades, que variam de 13 a 86 anos, e pode-se notar que a grande parcela dos candidatos é jovem, sendo 55,2% do total as pessoas que têm de 13 a 21 anos. O histograma relacionado à idade discretizada é apresentado na Figura 3 abaixo.



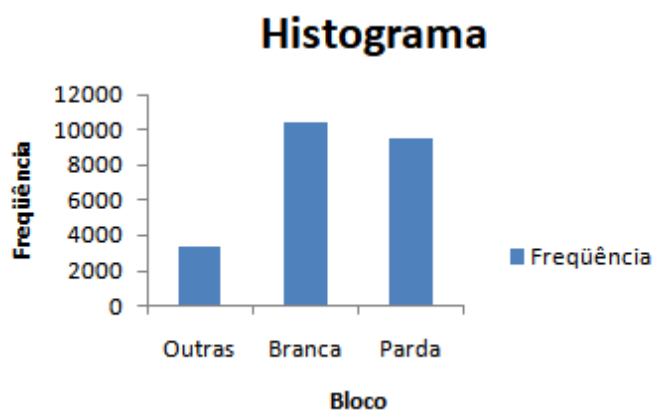
Figura 3 – Dados do atributo “idade” depois da discretização



Fonte: Autor.

Outra coluna analisada se refere à Cor da Pele, na qual os dados estão armazenados de acordo com o histograma presente na Figura 4 a seguir, sendo possível notar que brancos e pardos são a maioria, com 85,5% do total. Entretanto, pode-se notar que os grupos não ficaram bem divididos em suas faixas, pois Branca e Parda não podem ser arranjados em grupos em questão da sua maioria, restando apenas a solução de agrupar as outras etnias apresentadas na base (preta, amarela, indígena e não declarado) em um intervalo único.

Figura 4 – Dados do atributo “cor da pele” depois da discretização

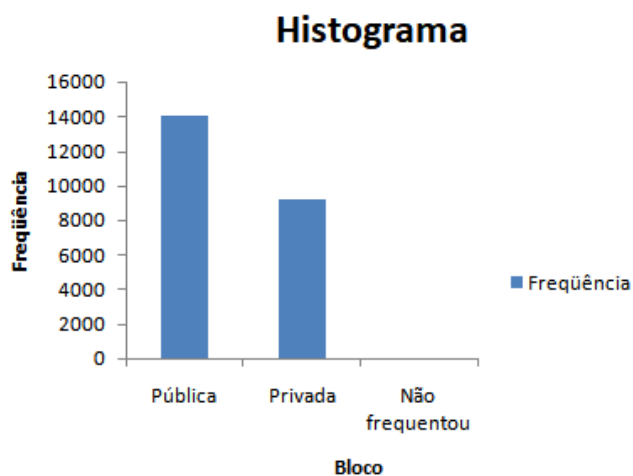


Fonte: Autor.

A seguir, observou-se a coluna referente à escola de ensino médio na qual o candidato cursou. Temos que os participantes que não frequentaram ensino médio representam 0,16% do total, entretanto não é possível agregar em escola pública ou privada. Em relação à escola privada foi considerado que se o participante alguma vez chegou a ter aula numa instituição particular, então ele entra para o grupo de Privada, que acaba contando com 39,65% dos estudantes. Os outros 60,19% representam os alunos que cursaram em escola pública. O histograma da divisão é apresentado na Figura 5.



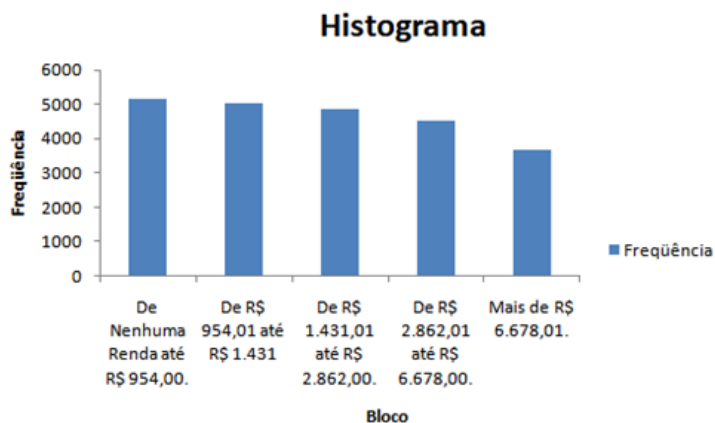
Figura 5 – Dados do atributo “tipo de escola” depois da discretização



Fonte: Autor.

O último atributo que se mostrou necessário discretizar foi a da Classe Econômica do candidato, que mostra em qual faixa de renda ele se enquadra. E o seu histograma pode ser visto a seguir na Figura 6.

Figura 6 – Dados do atributo “classe econômica” depois da discretização



Fonte: Autor.

Após as operações de discretização apresentadas na seção anterior, algumas colunas se mostram desnecessárias para nosso estudo, pois não contêm informações pertinentes. Como, por exemplo, colunas com muitos dados vazios, códigos, vetor com respostas da prova e outras informações. Tais dados foram removidos da nossa base.

Finalmente, na última etapa do nosso pré-processamento, foi realizada a seleção dos dados. Como a quantidade de registros continua expressiva, utilizou-se a técnica de Feature Selection para visualizar quais fornecem um ganho de informação maior, ou seja, quais que

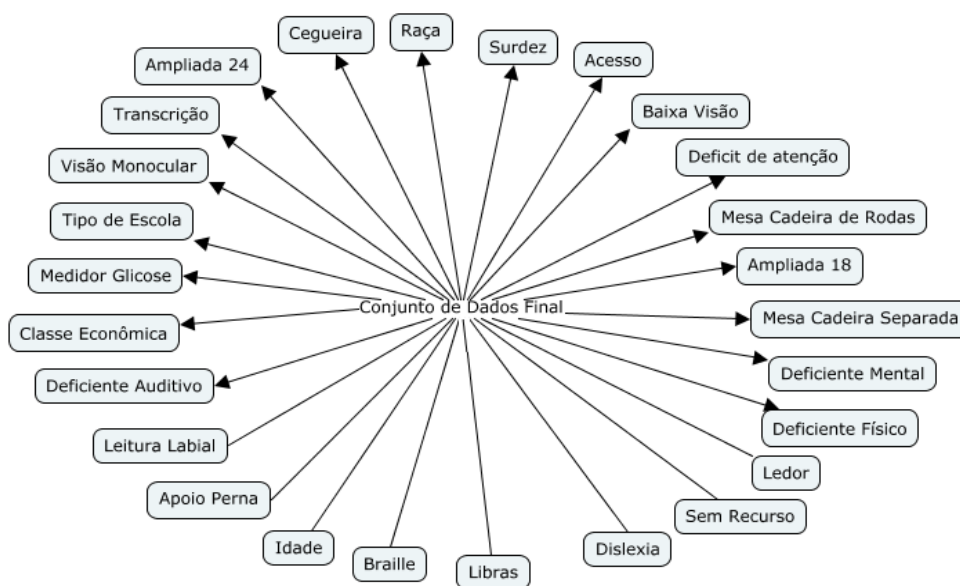


mais se relacionam com as 5 classes geradas do atributo alvo Média Final, que são: Ruim, Regular, Bom, Ótimo e Excelente.

Alguns atributos apresentaram um grande ganho de informação, porém são dados que interferem diretamente na nota, como as notas individuais de cada área de conhecimento e as notas de cada competência avaliada na redação. Esses dados foram desconsiderados no conjunto principal para análise, mas foram inseridos separadamente em outros dois conjuntos para critério de análise.

Com o auxílio da ferramenta CmapTools (IHMC, 2005), um esquema gráfico foi criado para uma melhor visualização dos atributos finais escolhidos, e está presente na Figura 7 a seguir.

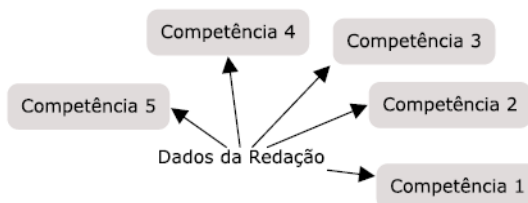
Figura 7 – Conjunto de dados final



Fonte: Autor.

Os dados que interferem diretamente na média final podem ser vistos abaixo, na Figura 8 e na Figura 9, e foram inseridos no conjunto de dados final, como o segundo e terceiro grupo de análise respectivamente.

Figura 8 – Segundo conjunto para análise, com as competências da redação



Fonte: Autor.



Figura 9 – Terceiro conjunto para análise, com as notas individuais



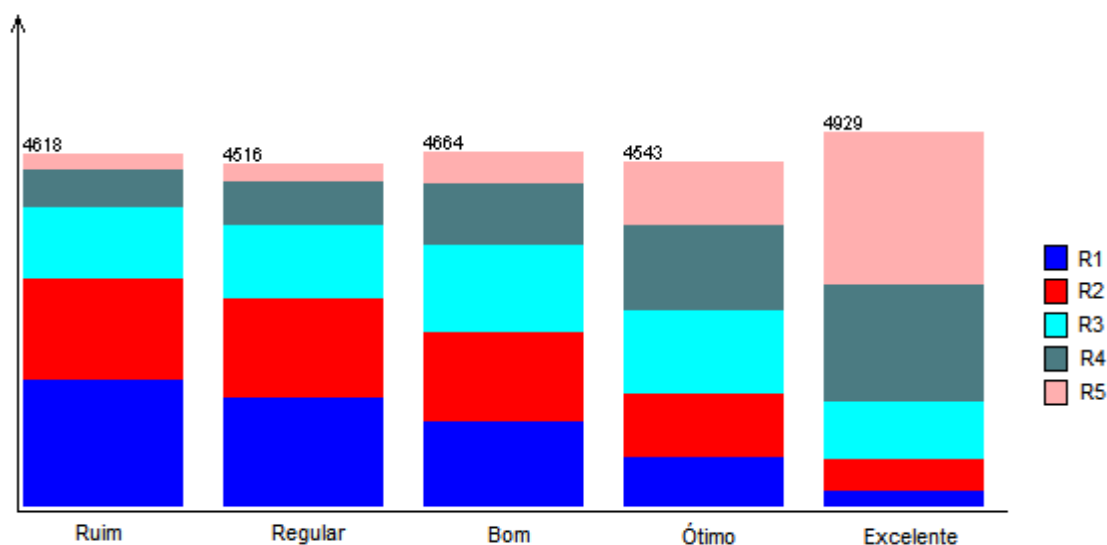
Fonte: Autor.

6 Resultados

De forma geral, ao inserir o conjunto de dados no WEKA, pode-se notar alguns resultados interessantes em relação aos atributos de Classe Econômica, Tipo de escola, Raça, Idade e Média Final.

Quando analisamos o atributo de Classe Econômica do candidato, no caso sua renda mensal, podem-se notar algumas informações pertinentes. Dos resultados obtidos, temos por exemplo que o desempenho final do candidato é diretamente proporcional à renda, ou seja, quanto mais renda o candidato tem, melhor será seu desempenho final. Resultado obtido ao visualizar a Figura 10. Lembrando que R1 equivale a menor renda enquanto R5 consiste na maior renda.

Figura 10 – Relação entre os atributos “Classe Econômica” e “Média Final”

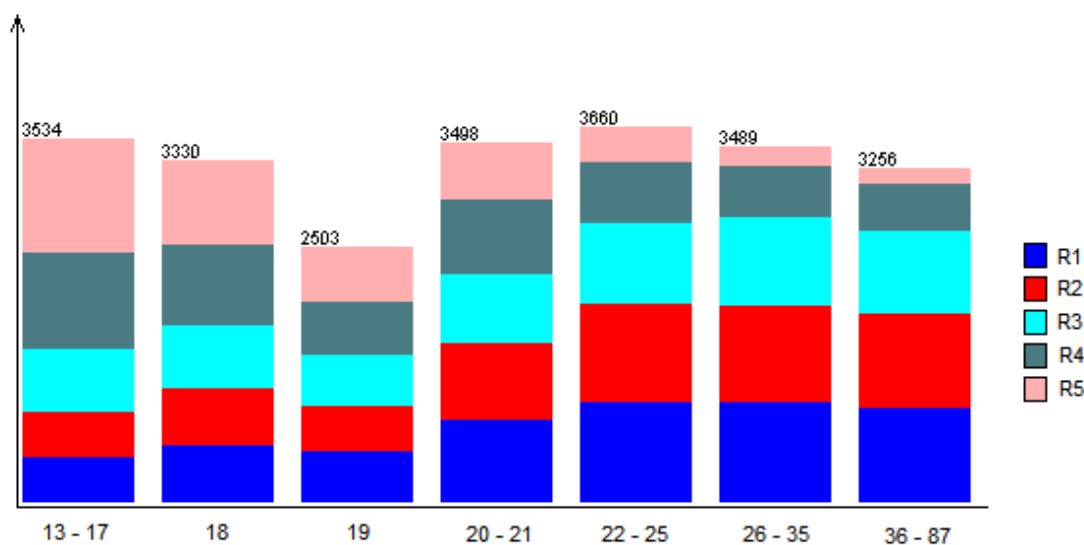


Fonte: Autor.

Já em relação à Idade, visualiza-se que candidatos que dispõem de mais renda são os mais novos, e conforme a faixa etária aumenta, o índice de renda diminui, resultado apresentado na Figura 11.



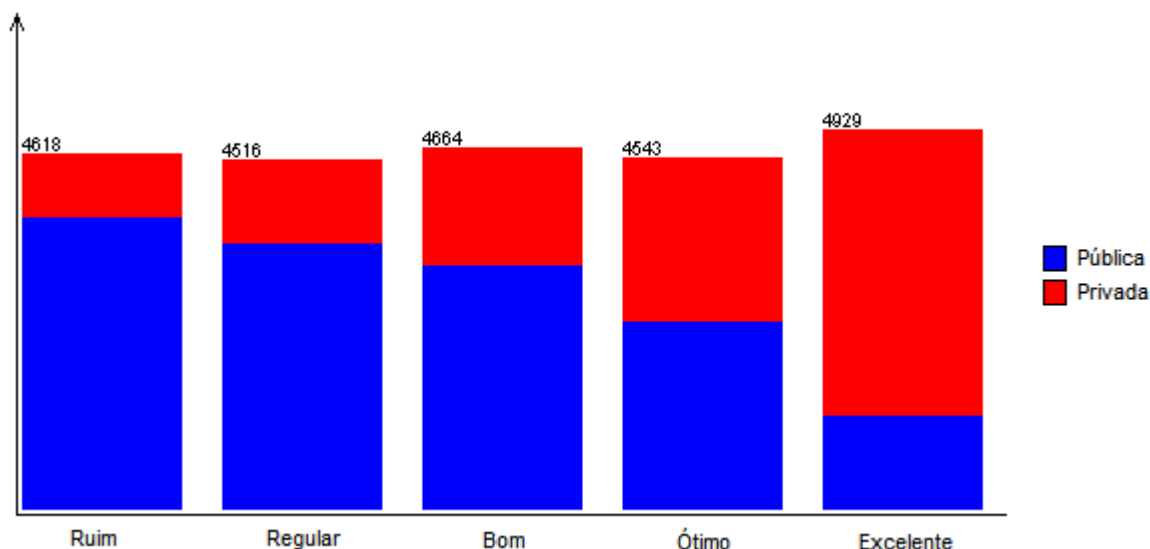
Figura 11 – Relação entre os atributos “Classe Econômica” e “Idade”



Fonte: Autor.

Na análise do atributo Tipo de Escola, ao ser relacionado com o atributo da Média Final, nota-se que o candidato que cursou uma escola pública tem um resultado pior do que aquele que cursou uma escola particular, apresentando indícios de que uma escola particular oferece um ensino mais adequado. A Figura 12 a seguir mostra a informação.

Figura 12 – Relação entre os atributos “Tipo de Escola” e “Média Final”

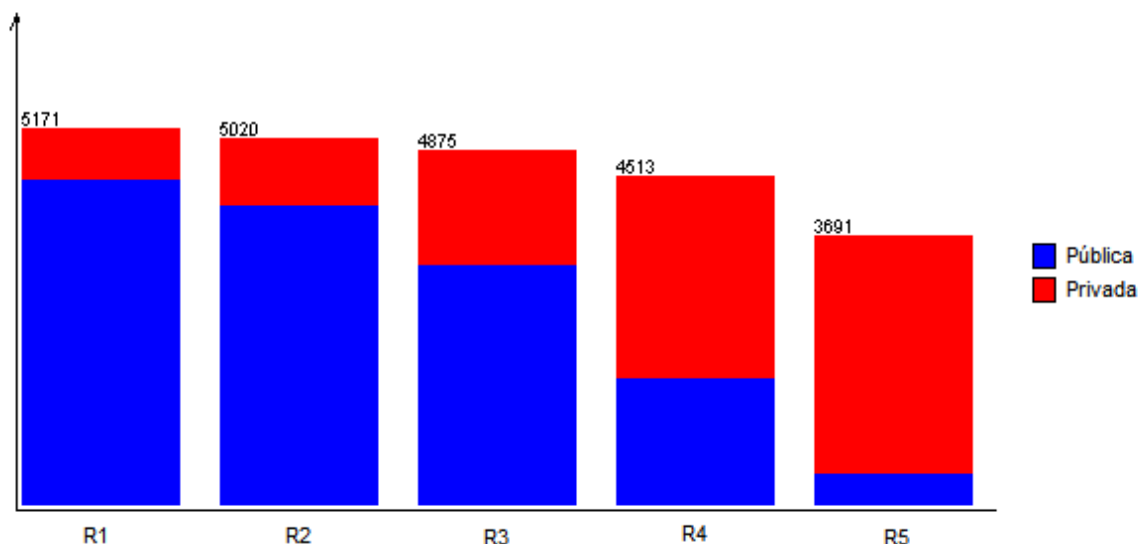


Fonte: Autor.

Ao ser analisada o atributo de Classe Econômica, tem-se evidências que candidatos que cursaram ensino particular no ensino médio são aqueles com as melhores rendas mensais. A Figura 13 contém esses dados.



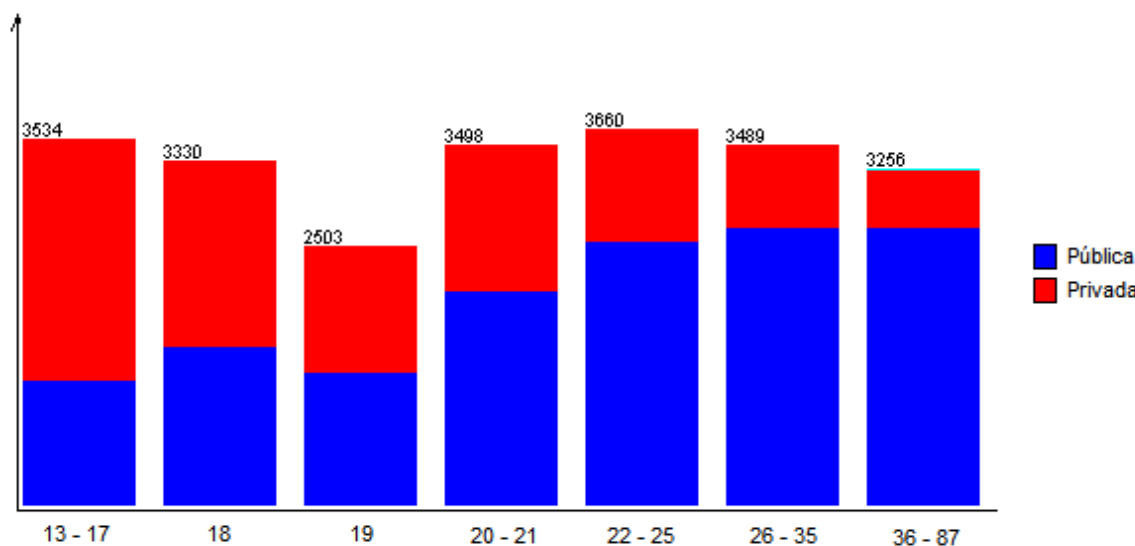
Figura 13 – Relação entre os atributos “Tipo de Escola” e “Classe Econômica”



Fonte: Autor.

Se aplicada no atributo Idade, pode-se notar que os candidatos mais jovens são os que mais têm acesso ao ensino particular, enquanto, conforme a faixa etária cresce, os candidatos cursam o ensino público em sua maioria. Resultado na Figura 14 a seguir.

Figura 14 – Relação entre os atributos “Tipo de Escola” e “Idade”

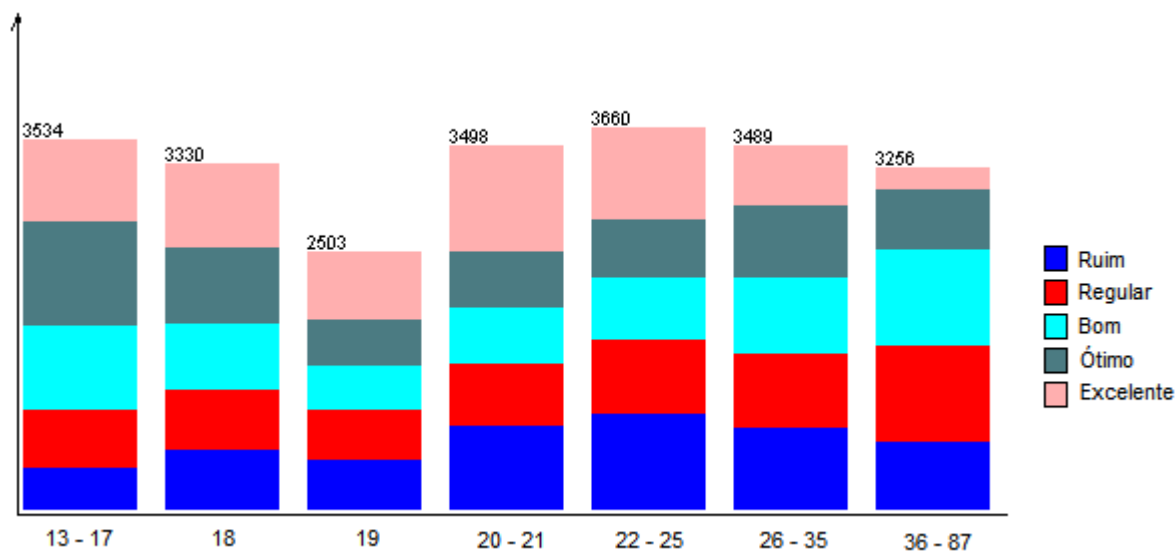


Fonte: Autor.



Finalizando, ao utilizar o atributo de Média Final no atributo de Idade, nota-se que os candidatos mais novos são os que têm um melhor desempenho na prova, de acordo com a Figura 15.

Figura 15 – Relação entre os atributos “Tipo de Escola” e “Média Final”



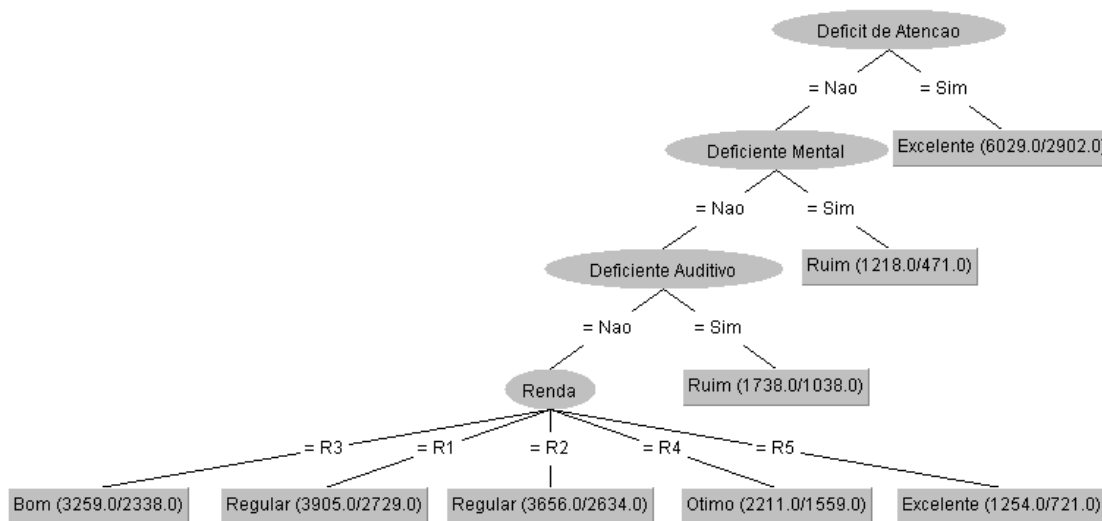
Fonte: Autor.

Foi considerado no nosso problema um *baseline* de 21,1818%, resultado alcançado ao utilizar o algoritmo zeroR no software WEKA, método que efetua uma classificação aplicada somente à classe majoritária.

A base de dados utilizada neste trabalho foi expressiva, e utilizar diretamente o algoritmo de aprendizado de máquina c4.5 não foi viável, pois, desta forma, a árvore de decisão gerada foi extensa e não interpretável. Por isso, utilizou-se o conjunto de dados pré-processado deste trabalho. Executou-se o algoritmo considerando pelo menos 1200 instâncias por folha, valor que foi escolhido por favorecer a visualização da informação e, desta forma, tornando a árvore mais legível. Esse valor significa que do total (23270 registros) foram analisados somente 1200, ou seja, 5,15% da árvore de decisões gerada. Mostramos na Figura 16 o resultado obtido ao utilizar o primeiro conjunto de dados finais.



Figura 16 – Árvore de decisões gerada após aplicar no conjunto de dados escolhido pelo Feature Selection e filtrado



Fonte: Autor.

Dessa árvore podem-se abstrair alguns indícios a partir da análise das regras de decisão geradas. Por exemplo, o modelo presume que se o candidato tem Déficit de Atenção, então o seu resultado será Excelente. Essa regra surgiu pelo fato do conjunto de dados conter em sua maioria participantes com Déficit de Atenção que tiveram um rendimento bom na prova. Dentre os outros resultados, temos que, se o candidato tiver Déficit Mental ou Deficiência Auditiva, sua média final será Ruim. Por fim, um candidato cuja deficiência não é nenhuma das outras apresentadas na árvore, de acordo com sua renda terá o seguinte rendimento:

- a) R1: Regular;
- b) R2: Regular;
- c) R3: Bom;
- d) R4: ótimo;
- e) R5: excelente.

Além disso, considerando o rendimento do modelo, temos a acurácia com resultado de 36,1238%. Valor que, apesar de baixo, é superior ao *baseline* do conjunto original.

Para critério de comparação, foram obtidos novos resultados submetendo o algoritmo no conjunto de dados formado por atributos selecionados pela técnica de Feature Selection como competências da redação. Isto ocorreu na tentativa em gerar um modelo mais preciso e genérico. A árvore gerada com esta abordagem é apresentada na Figura 17.

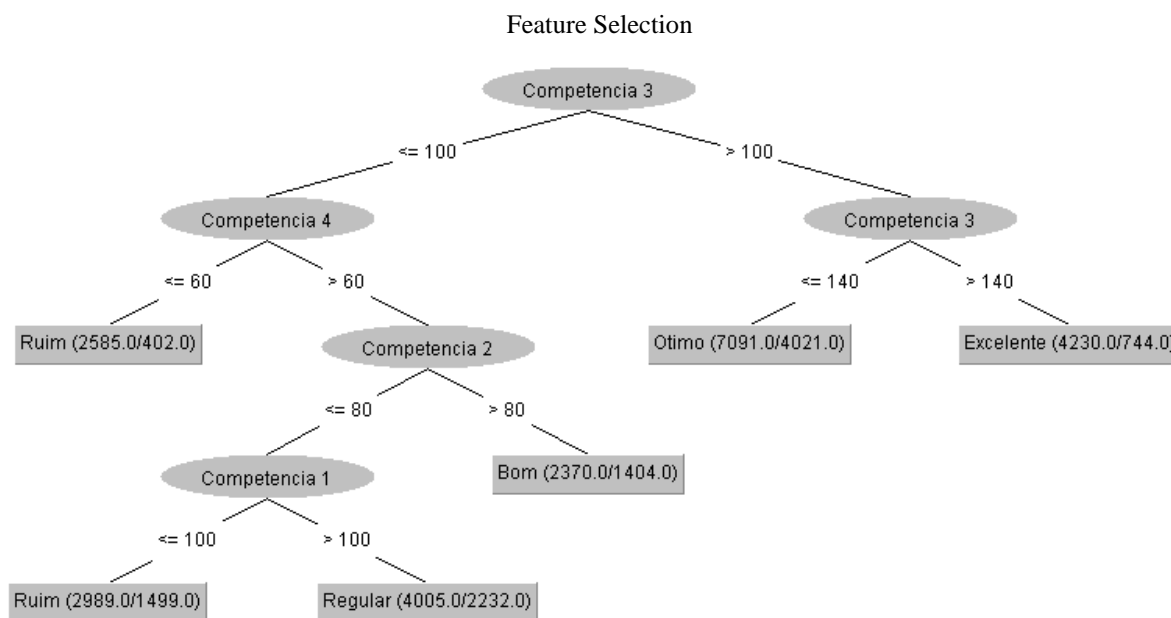
Lembrando que cada competência varia de 0 a 200 e tem o seguinte significado:

- a) Nota da competência 1 - Demonstrar domínio da modalidade escrita formal da Língua Portuguesa.
- b) 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.
- c) Nota da competência 3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
- d) Nota da competência 4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.



- e) Nota da competência 5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Figura 17 – Árvore de decisões gerada após aplicar no segundo conjunto de dados selecionado pelo



Fonte: Autor.

Nessa figura é possível notar que o atributo Competência é levado em consideração, isto porque seu valor interfere na nota da redação, diretamente relacionada à média final. Destaca-se que neste modelo a acurácia subiu para 55,7284%; logo, apresentou uma maior taxa de acertos.

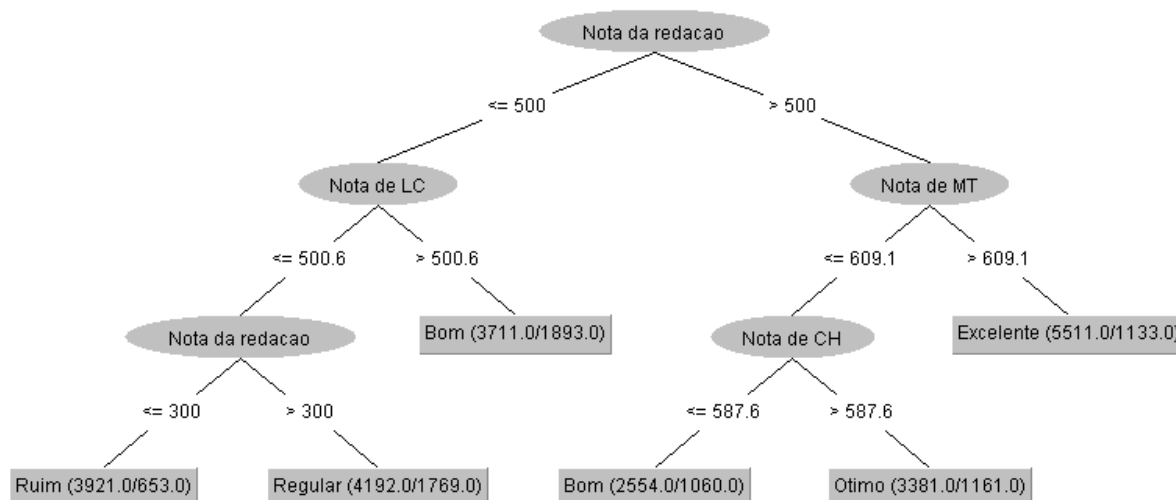
Pode-se então presumir que, se o candidato obtiver uma nota acima de 140 na competência 3, que avalia a capacidade de defender um ponto de vista, então seu desempenho será Excelente. Outra regra interessante pressupõe que, se o candidato obtém nota abaixo de 100 na competência 3 e abaixo de 60 na competência 4, que avalia o conhecimento de mecanismos linguísticos para a argumentação, então seu desempenho será Ruim. De forma geral, o participante que sabe como defender um ponto de vista e que sabe construir uma boa argumentação tem possibilidade de um melhor desempenho. Outras regras:

- a) Competência 3 entre 100 e 140: Ótimo;
- b) Competência 3 menor que 100, competência 4 maior que 60 e competência 2 maior que 80: Bom;
- c) Competência 3 menor que 100, competência 4 maior que 60, competência 2 menor que 80 e competência 1 maior que 100: Regular;
- d) Competência 3 menor que 100, competência 4 maior que 60, competência 2 menor que 80 e competência 1 menor que 100: Ruim.

Por fim, uma última análise é realizada. Desta vez, o classificador foi executado no terceiro conjunto de dados. A acurácia calculada foi a mais alta, sendo 66,6266%. Pode-se dizer que esse acréscimo existiu em virtude dos atributos de notas das áreas de conhecimento e a nota da redação estarem inclusas, dados que interferem diretamente na Média Final. A árvore gerada está na próxima imagem, Figura 18.



Figura 18 – Árvore de decisões gerada após aplicar no conjunto de dados completo



Fonte: Autor.

Como esses atributos interferem diretamente na Média Final, apresentam um ganho de informação considerável, aparecendo na árvore gerada. Das regras obtidas, supõe-se que ao obter uma nota maior que 500 na Redação e maior que 609.1 em Matemática, o candidato terá um desempenho médio Excelente. Assim, considera-se que quando o participante tem uma boa nota nessas áreas, então também sairá bem nas outras áreas. Visualiza-se também que quando a nota na área de Linguagens é menor que 500,6 e a nota de redação menor que 300, seu desempenho será Ruim sem nem considerar outras áreas. Assim, aquele candidato que não tem domínio na língua portuguesa, tais como problemas na escrita, erros de gramática, compreensão, interpretação, argumentação e construção de ideias, terá uma Média Final Ruim.

7 Conclusão

O objetivo geral da proposta foi analisar o desempenho de candidatos com deficiência que realizaram a prova do ENEM em 2018, por meio de algoritmos de descoberta e conhecimento em base de dados, fundamentado em características socioeconômicas, demográficas, étnicas, faixa-etária e formação acadêmica, verificando se tais atributos influenciam no rendimento final. Como visto, o ENEM é uma das provas mais importantes do Brasil, pois sua nota pode ser utilizada para o ingresso em diversas universidades, tanto nacionais quanto internacionais. Além disso, mesmo com o auxílio de cotas, a inclusão de deficientes no ensino superior ainda acontece de forma lenta, por isso uma análise em torno dos dados relacionados a este assunto se torna importante.

Pode-se observar que o acesso a escolas particulares e uma renda mais alta contribuem de forma elevada para um melhor desempenho. Notou-se também que um número pequeno de deficientes possui um diploma de ensino superior, e um número menor ainda tenta ingressar em uma universidade. Entre os candidatos, os mais novos apresentam uma melhor renda e têm mais acesso a um ensino privado, o que causa um rendimento melhor nas faixas etárias iniciais.

Em relação à aplicação do algoritmo C4.5 para criação de árvores de decisões, os resultados obtidos foram interessantes. É confirmado o fato da renda influenciar no desempenho, e três resultados quanto a algumas deficiências são obtidos, assumindo que um candidato com Déficit de Atenção se sairá de forma Excelente, e um Deficiente Mental e Auditivo terão um desempenho Ruim. Lembrando que uma pessoa que apresenta Déficit de

