UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
CAMPUS DE NATAL
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

HUGO LEONARDO DE LIMA OLIVEIRA

ANÁLISES E MINERAÇÃO DE DADOS RELACIONADOS AOS RESULTADOS DO ENADE DO CURSO DE CIÊNCIA DA COMPUTAÇÃO DO CAMPUS AVANÇADO DE NATAL.

NATAL

2018

HUGO LEONARDO DE LIMA OLIVEIRA

ANÁLISES E MINERAÇÃO DE DADOS RELACIONADOS AOS RESULTADOS DO ENADE DO CURSO DE CIÊNCIA DA COMPUTAÇÃO DO CAMPUS AVANÇADO DE NATAL.

Monografia apresentada à Universidade do Estado do Rio Grande do Norte – UERN – como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação.

.

ORIENTADOR: Dr. Isaac de Lima Oliveira Filho.
COORIENTADOR: Dr. Brismark Goes da Rocha.

NATAL

2018

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

O48a Oliveira, Hugo Leonardo de Lima
ANÁLISES E MINERAÇÃO DE DADOS RELACIONADOS AOS RESULTADOS DO ENADE DO CURSO DE CIÊNCIA DA COMPUTAÇÃO DO CAMPUS AVANÇADO DE NATAL. / Hugo Leonardo de Lima Oliveira. - Natal, 2018.
137p.

Orientador(a): Prof. Dr. Isaac de Lima Oliveira Filho.
Coorientador(a): Prof. Dr. Brismark Goes da Rocha.
Monografia (Graduação em Ciência da Computação). Universidade do Estado do Rio Grande do Norte.

1. Ciência da Computação. 2. Mineração de Dados. 3. ENADE. 4. Classificação. 5. UERN. I. Oliveira Filho, Isaac de Lima. II. Universidade do Estado do Rio Grande do Norte. III. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pela Diretoria de Informatização (DINF), sob orientação dos bibliotecários do SIB-UERN, para ser adaptado às necessidades da comunidade acadêmica UERN.

HUGO LEONARDO DE LIMA OLIVEIRA

ANÁLISES E MINERAÇÃO DE DADOS RELACIONADOS AOS RESULTADOS DO ENADE DO CURSO DE CIÊNCIA DA COMPUTAÇÃO DO CAMPUS AVANÇADO DE NATAL.

Monografia apresentada à Universidade do Estado do Rio Grande do Norte – UERN – como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação.

Aprovado em ___/___/____.

Banca Examinadora

_____
Dr. Isaac de Lima Oliveira Filho.
UERN

_____
Dr. Brismark Goes da Rocha.
UERN

_____
Dr. Carlos André Guerra Fonseca.
UERN

Ao Caos.

AGRADECIMENTOS

Agradeço primeiramente aos processos aleatórios (que podem não ser tão aleatórios assim, tudo pode estar conectado, e fazer algum sentido) que fizeram com que eu chegasse até ao final do curso. Em seguida aos meus pais por terem me dado a possibilidade de estudar em um nível superior da educação, e acreditaram em mim ao longo dos anos. Meu pai já dizia: “educação é tudo”. Se não fosse pelo o incentivo e a confiança dos meus pais, talvez teria desistido ao longo do curso, devido à tantas dificuldades que apareceram.

Agradeço a minha avó materna por sempre torcer por mim, e aos demais familiares. Aos meus amigos por darem apoio emocional, e me colocarem “para cima” em momentos não tão bons. Aos livros que li, as músicas que escutei, filmes que vi, e aos demais conteúdos que fizeram com que eu crescesse e expandisse culturalmente e intelectualmente ao longo dos anos.

Agradeço a todos os professores que tive, desde os da educação fundamental até os da educação superior, que viram em mim, algum potencial, e me incentivaram de alguma maneira.

Agradeço ao professor Dr. Carlos André, por ter me orientado nos primeiros pensamentos do trabalho. Em especial gostaria de agradecer aos meus orientadores, Dr. Isaac Filho, e Dr. Brismark da Rocha, por terem me direcionado com tanta paciência, e respeito, acreditando que seria possível finalizar esse trabalho.

RESUMO

O Exame Nacional de Avaliação do Estudante (ENADE), é uma prova que serve para avaliar o desempenho do estudante de nível superior. Tendo em vista, explorar se fatores socioeconômicos determinam o desempenho na nota final da prova, este trabalho aplica o processo *Knowledge Discovery in Databases* (KDD), com o intuito de buscar padrões por meio de técnicas que utilizam aprendizado de máquina. Os tipos de aprendizagem empregados neste trabalho, são: a aprendizagem não supervisionada, mediante regras de associação, obtidas pelo algoritmo *Apriori*, e a aprendizagem supervisionada, por meio da classificação, pelos algoritmos, J48, *Naive Bayes*, IBk, Sequential Minimal Optimization (SMO), e Multilayer Perceptron (MPL). As bases de dados do ENADE foram coletadas do site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), sendo selecionados os dados referentes aos estudantes de ciência da computação da Universidade do Estado do Rio Grande do Norte (UERN), do Campus Avançado de Natal, dos anos de 2008, 2011, e 2014. Este trabalho traça o perfil do estudante que fez a prova, obtém a correlação das variáveis selecionadas dos questionários citados, compara o desempenho dos estudantes de Natal com o cenário nacional, e identifica quais áreas do conhecimento do curso os estudantes têm um melhor rendimento. Duas ferramentas principais foram utilizadas para execução deste trabalho. Para a classificação, foi utilizada a *Waikato Environment for Knowledge Analysis* (WEKA), e a *R Studio* para a obtenção de regras de associação, cálculos estatísticos, e geração de gráficos. Após a finalização da mineração de dados, foi identificado que a situação de trabalho é um dos fatores que determinam a nota final. Estudantes que trabalhavam 40 horas semanais ou mais, tinham um rendimento inferior aos que apenas estudavam. Apesar disso o desempenho dos estudantes, está de acordo com a realidade do curso a nível nacional, com pequenas variações.

Palavras-chave: Mineração de Dados. ENADE. Classificação. Ciência da Computação. UERN.

ABSTRACT

The National Student Assessment Exam in Brazil (ENADE), is a test used to evaluate the student's performance in higher education. In order to explore whether socioeconomic factors determine the performance in the final grade of the test or not, this work will apply the Knowledge Discovery in Databases process, searching for patterns through techniques that use machine learning. The types of learning that used in this work are: unsupervised learning, through association rules, obtained by the Apriori algorithm, and supervised learning, through classification, by the algorithms J48, *Naive Bayes*, IBk, Sequential Minimal Optimization (SMO), e Multilayer Perceptron (MPL). The ENADE databases were collected from the site of Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), being selected the data referring to the students of computer science from Natal Advanced Campus of Universidade do Estado do Rio Grande do Norte (UERN). From the years 2008, 2011 and 2014. The work traces the profiles of the students who did the test, obtains the correlation of the variables selected from the questionnaires cited, compares the performance of Natal city students with the national scenario, and identify which knowledge areas from computer science course, students have a better performance.Two main softwares were used to perform this work. For the classification, it was used the Waikato Environment for Knowledge Analysis (WEKA), and the R Studio for association rules, statistical calculations, and graphics. After the data mining conclusion, was identified that work situation is one of the factors that determine the final grade. Students who worked 40 hours per week or more had a lower performance than the ones who only studied. Nevertheles, UERN students performances, is in line with the reality of the course at the national level, with small variations.

Key Words: Data Mining. ENADE. Classification. Computer Science. UERN.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O ensino superior é uma grande oportunidade para uma formação direcionada à uma das áreas do conhecimento, e para o crescimento intelectual e financeiro pessoal. Ao longo dos anos, o acesso a esse nível de escolaridade, aumentou. Segundo o relatório técnico divulgado do censo da educação superior divulgado pelo Inep (2016), o número de matriculas, em 2014, eram de 7,8 milhões, significando um aumento de 96,5%, comparado com 2003, quando as matrículas eram de 3.936.933. O número de vagas totalizava um pouco mais de 8 milhões, sendo que 3.042.977 das vagas eram do ensino à distância, e 5.038.392 presencial.

No Brasil, uma das formas de avaliar as intuições de ensino superior (IES), é por meio do Exame Nacional de Desempenho dos Estudantes (ENADE) (BARBOSA; FREIRE; CRISÓSTOMO, 2011).

Contudo, classificar a qualidade de uma instituição, ou de um curso, requer um acompanhamento mais profundo do *background* dos discentes, peça importante para que haja um processo de aprendizagem educacional.

Fatores socioeconômicos como, a renda familiar, a situação de trabalho, a aplicação de políticas de cotas, o estado civil, quantidade de horas de estudos, tipo de escola em que o estudante cursou o ensino médio, dentre outros fatores, que podem repercutir no desempenho final da nota da prova.

As intuições superiores vêm se adequando a chamada democratização do ensino superior, mas ainda há muitos problemas que precisam ser encarados, com seriedade, e devida importância (ALMEIDA et al., 2012).

Com o intuito de analisar de uma maneira mais aprofundada o desempenho acadêmico dos estudantes, alguns estudos, como os presentes na seção 1.3, estão utilizando o questionário socioeconômico, aplicado pelo ENADE, para analisar quais fatores podem influenciar no desempenho da prova, por meio da mineração de dados (MD).

Diante dos resultados desses estudos, as instituições, podem entender melhor o cenário das variáveis envolvidas, e propor algum tipo de política educacional dentro da instituição, que possa trazer melhorias futuras, para o desempenho em novas provas, visto que o ENADE, está se consolidando como um dos parâmetros de qualidade do ensino superior.

De acordo com Fayyad, Piatetsky-shapiro e Smyth (1996), há duas abordagens iniciais para se atingir os objetivos da MD: verificação, e descoberta. A verificação trata-se da constatação ou não das hipóteses colocadas inicialmente por quem conduziu a pesquisa ou estudo. Já a descoberta, parte do princípio que o sistema automaticamente busca por padrões.

Para este trabalho a abordagem escolhida, foi a descoberta, sendo quem conduz o estudo, responsável pelas análises e escolhas dos padrões encontrados, mediante os objetivos traçados.

Nesse contexto, tendo em vista a carência de estudos, o presente trabalho, analisa a base de dados de resultados do ENADE, do curso de ciência da computação, da Universidade do Rio Grande do Norte (UERN), do Campus Avançado de Natal, dos anos de 2008, 2011, e 2014.

1.1 MOTIVAÇÃO E JUSTIFICATIVA

A mineração de dados educacionais (MDE), é uma importante área da MD, que possibilita identificar quais fatores podem estar determinando o aprendizado, e o rendimento dos estudantes, desde o nível fundamental ao ensino superior, por meio da busca de padrões (ROMERO; VENTURA, 2012).

No Brasil, ainda existem poucos trabalhos relacionados a MDE, e para que se tenha cada vez mais uma compreensão de quais variáveis estão relacionadas ao desempenho do aluno, pesquisas e estudos precisam ser feitos, para que possam dimensionar o cenário de determinada instituição, e com isso, entender o contexto por trás dos seus estudantes, para que governos, ou equipes pedagógicas da própria instituição apliquem políticas que possam incentivar um melhor desempenho do estudante, uma vez que, o ambiente educacional é onde acontece o desenvolvimento dos cidadãos, e futuros profissionais (BAKER; ISOTANI; CARVALHO, 2011).

No cenário que concerne esse trabalho, a motivação parte de fornecer um panorama para a comunidade acadêmica, dando informações sobre o desempenho dos estudantes na parte objetiva específica do curso (que compreende as questões das disciplinas ensinadas ao longo do curso), e de acordo com variáveis selecionadas dos questionários socioeconômico e de percepção da prova, apresentar possíveis determinantes do desempenho dos estudantes ao longo dos anos analisados.

Com a grande variedade de dados que o ENADE proporciona, é interessante que existam trabalhos que busquem por padrões de dados, que possam servir para o acompanhamento e avaliação dos resultados por parte da comunidade acadêmica.

Além de despertar na comunidade acadêmica a necessidade da aplicação da MDE nos próprios departamentos dos cursos das universidades, centros universitários, e faculdades, com intuito de conhecer seus estudantes para dar o suporte necessário para um bom desempenho acadêmico, e valorização do curso em questão.

1.2 OBJETIVOS

Esta seção apresenta o objetivo geral do trabalho e seus objetivos específicos.

1.2.1 Objetivo geral

O objetivo geral deste trabalho consiste em buscar padrões nas questões selecionadas dos questionários socioeconômicos e de percepção da prova, que determinem o desempenho na nota final do ENADE, para isso foram aplicadas técnicas de mineração de dados, no que diz respeito às tarefas de cunho descritiva e preditiva, usando a base de dados dos anos de 2008, 2011, e 2014, que compreende o quadro dos dados governamentais abertos oferecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

1.2.2 Objetivos específicos

Os objetivos específicos para este trabalho são:

(I). Identificar o perfil do estudante do curso de ciência da computação;

(II). Apresentar o desempenho dos estudantes ao longo dos anos analisados, fazendo comparações com o cenário nacional;

(III). Apresentar o aproveitamento dos estudantes nas áreas do conhecimento específico da prova, de acordo com a complexidade das questões;

(IV). Apresentar a correlação entre as variáveis relacionadas.

1.3 TRABALHOS RELACIONADOS

Existem poucos trabalhos no Brasil voltados para a mineração de dados educacionais, tanto com relação aos níveis de escolaridade fundamental e médio, quanto ao ensino superior. Ainda é uma área de pesquisa pouco explorada na sua totalidade, mas com grandes chances de contribuições para a sociedade, diante de um entendimento mais profundo de questões sobre a educação, visto que existem bases de dados governamentais disponíveis, por exemplo, a do ENADE, cerne deste trabalho, e do Exame Nacional do Ensino Médio (ENEM), além das próprias bases de dados das instituições com atividades educacionais presenciais e à distância. Apesar da pouca produção de trabalhos, foram selecionados, alguns que contribuíram, de certa maneira para o desenvolvimento deste.

O primeiro trabalho destacado é o de Cretton e Gomes (2016), intitulado “Aplicação de Técnicas de Mineração de Dados na Base de Dados do ENADE com Enfoque nos Cursos de Medicina”, abordando o desempenho dos alunos e seu perfil, utilizando especificamente a base de dados do ano de 2013 do ENADE. Usando a ferramenta WEKA, diante da técnica de classificação árvore de decisão, através do algoritmo J48. Uma das conclusões desse trabalho foi que os estudantes de São Paulo e Rio de Janeiro das instituições privadas sem fins lucrativos, pertencentes a organização acadêmica universidade, obtiveram uma nota menor que sessenta, na sua maioria, sendo considerado um resultado ruim.

Em Nogueira e Tsunoda (2015), com o título do trabalho de “Mineração de Dados para Análise da Relação entre as Características Socioeconômicas de Concluintes do Ensino Superior e o Desempenho desses Estudantes no ENADE 2012”, fazendo uso do questionário socioeconômico da prova, para investigar se algumas questões selecionadas poderiam influenciar no desempenho do aluno. Este trabalho selecionou as variáveis do questionário que tinham uma maior relação com o atributo “nota bruta da prova”, utilizando o filtro CfsSubsetEval. Para a mineração, a técnica de classificação foi escolhida, utilizando a árvore de decisão gerada pelo algoritmo J48. A WEKA foi escolhida como ferramenta de mineração de dados. Uma das conclusões desse trabalho foi que existe uma maior probabilidade de desempenho satisfatório para estudantes com maior renda familiar.

Outra contribuição está presente no trabalho “Prática de Mineração de Dados no Exame Nacional do Ensino Médio”, do Congresso Brasileiro de Informática na

Educação (2014). Nesse artigo, foi utilizado a técnica de associação, empregando o algoritmo *Apriori*, usado para encontrar padrões de regras nos resultados da prova e dos questionários socioeconômicos. A base de dados foi a de 2010, selecionando algumas capitais da região sudeste. Segundo os resultados, a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas que moram com o estudante são atributos que diminuem o desempenho do aluno.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está organizado da seguinte forma:

O capítulo 2 contém toda fundamentação teórica considerada importante para esse trabalho. Os materiais e metodologias utilizadas para cada parte do trabalho são apresentados no capítulo 3. O capítulo 4 trata dos resultados obtidos diante da aplicação das técnicas de MD e estatísticas nas bases de dados, além das comparações feitas por meio do desempenho dos estudantes em relação ao desempenho à nível nacional do referido curso. Na seção 5 são tiradas as conclusões diante dos resultados do trabalho e sugeridos trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo será apresentado ao leitor, conteúdos que estão relacionados a mineração de dados, tema central desta monografia, que são essenciais para compreensão do trabalho, e outros que contribuem dentro do contexto dos dados que estão sendo utilizados. Na seção 2.1 é apresentada algumas definições do que seria dado, informação e conhecimento. A subseção 2.1.1 consiste em descrever os tipos de dados. Na seção 2.2 são descritos conceitos gerais da mineração de dados. As subseções 2.2.1, 2.2.2, 2.2.3, 2.2.4, e 2.2.5, referem-se respectivamente ao aprendizado de máquina, as tarefas da mineração de dados, a técnica de classificação, a de associação, e a definição de mineração de dados educacionais. Em seguida, respectivamente nas seções 2.3 e 2.4, a ideia de dados abertos governamentais e as características do ENADE são apresentadas.

2.1 DADO, INFORMAÇÃO E CONHECIMENTO

Existem várias explicações sobre o que viria a ser dado, informação e conhecimento, que se complementam e dão ao leitor uma visão ampla sobre esses conceitos abstratos.

Contudo, para esse estudo, é utilizado uma visão “simplista”, de que os dados são uma coleção de símbolos, textos e números que não possuem significado intrínseco. Os dados processados serão considerados informações, que podem significar algo e podem ser interpretadas. O processo de entendimento das informações, será considerado como conhecimento, que pode ser útil para determinado cenário, com o objetivo de ser utilizado como uma ferramenta de ajuda, na tentativa de amenizar ou solucionar problemas (LIEW, 2007).

Para complementar essa discussão, a subseção 2.1.1, terá o enfoque em detalhar características dos dados e conceito de variável. Essa compreensão será útil quando os dados que este trabalho explora, forem devidamente apresentados, possibilitando a identificação dos tipos de dados envolvidos.

2.1.1 Características dos dados

Para começar a entender as características dos dados, primeiramente devemos entender o conceito de variável. Martins, Loura e Mendes (2007), define variável como sendo "*qualquer característica de um indivíduo ou objecto à qual se possa atribuir um número ou uma categoria.* ".

Ainda, de acordo com os autores, as variáveis podem ser quantitativas (ou numéricas) ou qualitativas (ou categóricas). As quantitativas estão relacionadas as características das quais se pode contar ou medir, enquanto as qualitativas atendem unicamente a uma classificação, podendo assumir modalidades ou categorias. As variáveis qualitativas de apenas duas categorias são chamadas de binárias. As variáveis quantitativas que se pode contar, são chamadas de variáveis quantitativas discretas, já as de medição, são variáveis quantitativas contínuas.

Martins, Loura e Mendes (2007), coloca que "*O resultado da observação da variável, sobre o indivíduo, é o dado estatístico ou simplesmente dado. ".* Com essa colocação, é possível concluir que o dado assume a característica da variável, ou seja, se uma variável é categórica, provavelmente o dado terá essa característica.

Dessa forma os dados têm características qualitativas e quantitativas. Segundo Martins, Loura e Mendes (2007), os dados qualitativos ainda podem ter características nominais e ordinais.

Quadro 1: Características dos dados.

Dados quantitativos (Valores que podem ser comparados e ordenados).	Dados Discretos	Ex: Valores quantitativos no formato: 1,2,3,4,5... (valores inteiros).
	Dados Contínuos	Ex: Temperaturas (33ºC, 70F), altura (1,78m).
Dados Qualitativos (valores ordenados e não ordenados, podendo ser comparados ou não).	Dados Nominais	Ex: formatos (triangulo, quadrado, retângulo), gênero (masculino, feminino).
	Dados Ordinais	Ex: Tamanho da roupa (P, M, G, GG), classificação de notas (ruim, média, alta).

Fonte: Própria.

O Quadro 1, apresenta um resumo com exemplos para uma melhor compreensão do que foi descrito até agora, nesta seção.

Depois de uma explicação sucinta das características dos dados, outro conceito que completa essa discussão, e encerra essa subseção, é a noção das variáveis independentes e dependentes.

Para Jung (2009), variáveis independentes são aquelas selecionadas intencionalmente pelo pesquisador para verificar a relação entre suas variações e o comportamento das outras variáveis evolvidas, utilizada em experimentos que são conduzidos com o intuito de conseguir realizar previsões e/ou obter resultados.

Já as variáveis dependentes são praticamente o resultado do experimento, correspondendo aos comportamentos apresentados diante das oscilações das variáveis independes (JUNG, 2009).

2.2 MINERAÇÃO DE DADOS

Nesta seção será dada ao leitor uma visão geral da temática principal desta monografia: a mineração de dados.

Mineração de dados, do inglês *data mining* (DM), é uma tecnologia usada para resolver problemas, utilizando análises de dados, sendo que os dados, previamente, devem estar presentes em uma base de dados. A mineração de dados é definida como sendo um processo de descobertas de padrões em dados. Os padrões descobertos devem possuir significados que possam beneficiar ou contribuir para um determinado grupo, segundo interesses estabelecidos, geralmente utilizada para vantagens econômicas (WRITTEN; FRANK, 2005).

Existem outros termos que são considerados sinônimos para a mineração de dados, os mais conhecidos são: mineração de conhecimentos a partir de dados, extração de conhecimento, análises de dados/padrões, arqueologia de dados, dragagem de dados e descoberta do conhecimento em base de dados (termo mais utilizado). Alguns autores consideraram a mineração de dados como sinônimo do processo de descoberta do conhecimento em base de dados, do termo *knowledge discovery from data* (KDD), outros têm a visão de que a mineração de dados, é apenas uma das etapas do processo KDD. (HAN; KAMBER; PEI, 2012).

O presente trabalho segue a linha de pensamento de que a mineração de dados pertence a uma das etapas do processo KDD. Dessa maneira a organização

do trabalho ficará mais detalhada, dando ao leitor uma compreensão mais ampla de toda a metodologia realizada.

Seguindo a visão de Fayyad, Piatetsky-shapiro e Smyth (1996), mineração de dados é uma das etapas de um processo conhecido como descoberta de conhecimento em bases de dados (KDD). O processo KDD envolve várias etapas, podendo possuir sub etapas. Trata-se de um processo interativo e iterativo, onde o usuário toma decisões em todo o processo deixando-o mais consistente.

As etapas básicas do processo KDD, segundo Fayyad, Piatetsky-shapiro e Smyth (1996), consistem em:

Primeiramente, na escolha de um tema a ser explorado, logo em seguida, um estudo para a compreensão do seu domínio deve ser realizado. Com o estudo realizado, pode-se definir as pretensões desejadas, para posteriormente serem analisadas.

Na sequência, selecionar um conjunto de dados que estejam diretamente relacionados com o propósito da pesquisa (dados relevantes), desse conjunto podem ser selecionadas as variáveis mais adequadas para o cenário escolhido.

Depois, os dados precisam sofrer um pré-processamento, ou seja, eles precisam ser organizados, nesta etapa, se necessário, ruídos devem ser eliminados, quando há a falta valores, uma estratégia precisa ser adotada para contornar o problema, entre outras correções relacionadas à limpeza de dados.

Seguindo com o processo, os dados geralmente passam por transformações, por exemplo, dados numéricos podem ser convertidos em nominais, de acordo com a projeção de que essas mudanças afetam de forma positiva a sequência do processo KDD.

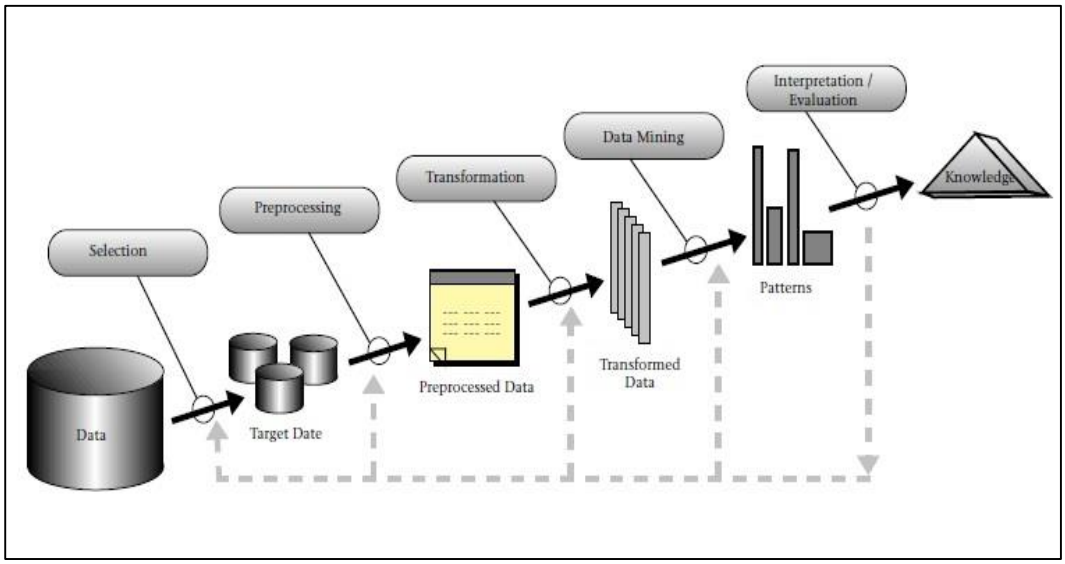
Prontamente, a mineração de dados pode ser aplicada, definindo previamente os métodos e os algoritmos de mineração de dados que serão utilizados para a busca de padrão de dados. Vale ressaltar que o desempenho da mineração de dados, depende das etapas anteriores.

Após a mineração de dados, ocorrem as interpretações e análises dos resultados obtidos, ou seja, dos padrões minerados.

A sétima e última etapa, corresponde na descoberta do conhecimento. O conhecimento pode ser incorporado a um sistema, ou simplesmente documentado e apresentado para as partes interessadas.

O processo KDD está apresentado na Figura 1, para uma melhor visualização do que foi descrito.

Figura 1: Processo KDD.

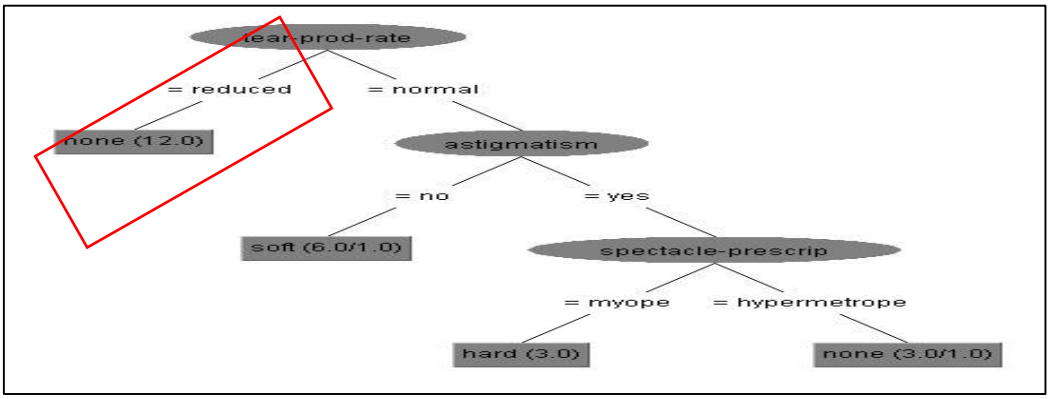


Fonte: Fayyad, Piatetsky-shapiro e Smyth (1996).

Os padrões gerados na etapa da mineração de dados podem ser representados em termos de uma estrutura que pode ser examinada e usada para informar futuras decisões. A estrutura pode ajudar a explicar algo sobre os dados. (WRITTEN; FRANK, 2005).

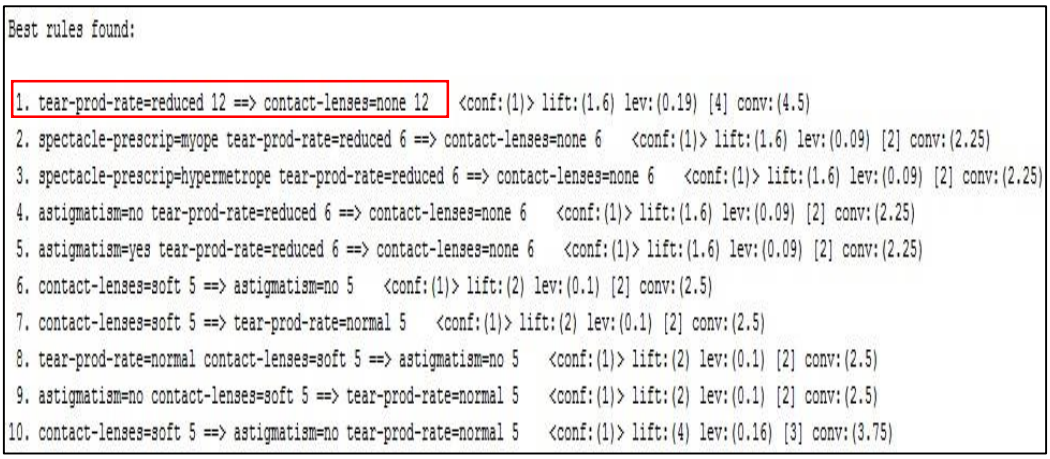
As Figuras 2 e 3, são exemplos de estruturas de árvore de decisão e regras de associação, respectivamente (nas seções posteriores serão explicadas essas e outras técnicas que a mineração de dados proporciona, mas neste ponto o leitor poderá começar a entender a praticidade da DM).

Figura 2: Estrutura de árvore de decisão.



Fonte: Própria.

Figura 3: Estrutura das regras de associação.



Fonte: Própria.

Essas estruturas são uma das possibilidades da mineração de dados, e se enquadram na abordagem da mineração de dados voltada para a visualização dos dados, seja por meio de uma árvore ou regras de associação.

O exemplo tratado nas figuras, é sobre recomendação de lentes de contatos. Se baseia, em prescrição de óculos (miopia ou hipermetropia), verificando se a pessoa possui astigmatismo e a taxa de produção de lágrimas (normal ou reduzida). As lentes podem ser classificadas em: lentes de contato suaves, dura, ou sem lentes de contato.

Mesmo para quem não tem familiaridade com a mineração de dados, visivelmente pode-se ter uma compreensão com as estruturas apresentadas, de como os dados se comportam, ou seja, os padrões.

Obviamente, a pessoa que busca entender a estrutura, tem que ter conhecimento sobre a base de dados que está sendo utilizada.

Neste caso, apesar das estruturas serem diferentes, elas possuem certas familiriades, por exemplo, a cor vermelha destacada nas Figuras 2 e 3, evedencia que ambas as estruturas possuem a seguinte regra: se a taxa de produção de lágrimas for igual a reduzida, a pessoa não poderá usar lentes de contato.

Então diante de uma base dados, essas estruturas, e outras, podem ser utilizadas para ajudar o pesquisador a encontrar quais variáveis de um conjunto de dados, aspectos, impactos, relações e outros entendimentos sobre os dados se consegue extrair, e ter importância para determinada pesquisa.

A mineração de dados pode ser aplicada em diferentes campos, da ciência aos negócios, e até em esportes (SUMATHI; SIVANANDAM, 2006).

A área da mineração de dados educacionais, foi escolhida para ser empregada neste trabalho. A subseção 2.2.5, e as seções 2.3 e 2.4, assim como os capítulos posteriores darão o contexto necessário para deixar o leitor situado em como o tema está sendo abordado.

A mineração de dados envolve uma interdisciplinaridade de vários domínios. Incluindo estatísticas, aprendizado de máquina, e bases de dados. E isso contribui significativamente para o sucesso da técnica em várias aplicações. (HAN; KAMBER; PEI, 2012).

O escopo deste trabalho não consegue abranger com detalhes todos os domínios e as mais variadas disciplinas que estão contidas e entrelaçadas com a DM, no entanto, será dedicada a subseção 2.2.1 para uma descrição condensada da área de estudo chamada aprendizado de máquina, que está intrinsicamente ligada à *data mining,* sendo possível definições compactas de técnicas da mineração de dados e de outros domínios, conteúdos e conceitos durante os capítulos seguintes, diante da necessidade, para deixar o leitor mais familiarizado, assimilando todos os pontos colocados neste trabalho.

2.2.1 Aprendizado de máquina

Para Alpaydm (2010), aprendizado de máquina consiste em que computadores (máquinas) sejam capazes de extrair um aprendizado de forma automática, utilizando algoritmos específicos para a realização dessa tarefa. O aprendizado é realizado tentando identificar certos padrões nos dados. Ressalta que talvez não seja possível identificar uma aprendizagem completa, mas sim, uma aproximação boa e útil, podendo detectar padrões e regularidades.

A mineração de dados é a aplicação do aprendizado de máquina que utiliza uma base de dados. Porém, aprendizado de máquina não é apenas voltada para problemas que envolvam um conjunto de dados, também é parte da inteligência artificial, auxiliando em tarefas como, reconhecimento de voz, e de rosto (ALPAYDM, 2010).

As técnicas de aprendizado de dados, resultam em predições para novas situações, ou no entendimento e descrição de como a predição é derivada (FRANK, 2000).

O conceito de aprendizado envolve adquirir conceitos gerais, de exemplos de treinamento específico em um conjunto de dados. Os dados são classificados em membros ou não membros do conceito envolvido. Um tipo de classificação utilizada é a booleana, podendo conter valores TRUE ou FALSE. O treinamento ocorre utilizando inferências em um conjunto de dados (MITCHELL, 1997).

Em outras palavras, toda base de dados escolhida deve ser utilizada no procedimento de treinamento, com intuito de gerar modelos genéricos diante dos dados apresentados, que podem ser utilizados em testes de futuras amostras, ou simplesmente no estudo das características do modelo, ou seja, o estudo das variáveis que se destacaram. Em capítulos posteriores, a fase de treinamento será retomada.

Para Domingos (2012), os algoritmos de aprendizado, consistem em três componentes: Representação, avaliação e otimização. A escolha da representação do conhecimento é um critério chave, e isso depende do domínio que se tem em extrair o conhecimento expresso por determinada técnica.

Os parâmetros de avaliação dependem dos algoritmos utilizados. É necessária para distinguir o desempenho dos classificadores. A escolha da técnica que possui uma otimização mais eficiente é a chave para o aprendizado. Quando na avaliação é verificado que duas ou mais técnicas que foram escolhidas para o aprendizado possuem funções de avaliação com valores semelhantes, deve-se escolher o que possui a mais adequada otimização.

Domingos (2012), aponta que o objetivo fundamental do aprendizado de máquina é a generalização para além dos exemplos do conjunto de treinamento.

De acordo com Ayodele (2010), os algoritmos de aprendizagem de máquina são organizados em uma taxonomia. Os tipos de aprendizado de máquina são:

- Aprendizado supervisionado – É gerada uma função por meio do algoritmo escolhido, que mapeia as entradas para as saídas desejadas. A classificação é geralmente a metodologia utilizada. Uma função mapeia um vetor em uma de várias classes, observando vários exemplos de entrada e saída da função.
- Aprendizado não supervisionado – Onde o aprendizado ocorre sem a necessidade da disponibilidade de exemplos de classes.

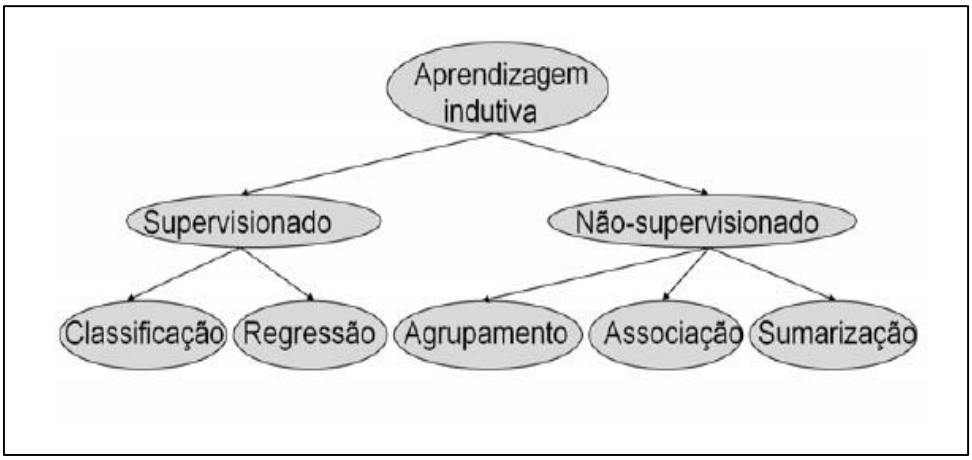
- Aprendizado semi-supervisionado – junção dos aprendizados supervisionados e não supervisionados.
- Aprendizagem de reforço – Onde a aprendizagem ocorre por meio da interação com o ambiente. Cada ação tem algum impacto no ambiente. O ambiente por sua vez, fornece um *feedback* para o algoritmo de aprendizagem.
- Transdução – A aprendizagem consiste em tentar prever novas saídas com base nos treinamentos da entrada e saída, e novas entradas.
- Aprendizagem por meio do aprendizado – onde o algoritmo aprende seu próprio viés indutivo com base em experiência passada.

Apesar de existir esses diferentes tipos de aprendizagem, as técnicas tradicionais da DM, incorporam na sua grande maioria, o aprendizado supervisionado e o não supervisionado, visto que o objetivo final dessas abordagens é tentar compreender a base de dados em questão, por meio da busca de padrões.

Na Figura 4, podemos observar uma hierarquia de aprendizagem voltada para a mineração de dados.

O aprendizado supervisionado e o não supervisionado são provenientes da aprendizagem indutiva, processo que gera as generalizações diante de uma base de dados (GAMA et al., 2015).

Figura 4: Hierarquia do aprendizado para DM.



Fonte: Gama et al. (2015).

As técnicas de classificação, regressão, agrupamentos, associação, e sumarização que estão da Figura 4 serão apresentadas na subseção 2.2.2, em seguida.

2.2.2 Tarefas da mineração de dados

De acordo com Larose (2005), as tarefas de mineração de dados mais comuns são: Descrição, Classificação, Estimativa ou Regressão, Predição, *Clustering* ou Segmentação, e Associação. Seguindo com as definições em Larose (2005):

Descrição – Em alguns casos, pesquisadores e analistas, estão apenas tentando descobrir caminhos para descrever padrões e tendências, observando o comportamento dos dados. Os modelos de mineração de dados devem ser transparentes com relação aos resultados obtidos, ou seja, o modelo deve descrever padrões claros que sejam passíveis de interpretação e explicação intuitiva. Alguns métodos de mineração são mais adequados que outros. Por exemplo, as árvores de decisão, são autoexplicativas e intuitivas para os humanos, por outro lado, as redes neurais, criam um modelo complexo, não sendo fácil de interpretar e explicar.

Classificação – O objetivo mais comum utilizando essa técnica, é a classificação de dados em uma determinada classe alvo. Por exemplo, classificar um conjunto de dados, segundo a faixa de renda, que pode ser: renda alta, renda média, e renda baixa. A classe alvo em uma classificação, deve ser uma variável categórica.

Estimativa ou Regressão – Similar à classificação, exceto com relação à classe, que deve ser numérica, em vez de categórica. O campo da análise estatística fornece vários métodos amplamente utilizados na regressão. Incluindo, estimativas pontuais e estimativas de intervalos de confiança, regressão e correlação linear simples, e regressão múltipla.

Predição – A predição é similar às tarefas de classificação e regressão, exceto que para predição, os resultados serão encontrados no futuro, ou seja, a predição utiliza dados do presente para prever situações futuras. A tarefa de predição geralmente utiliza as técnicas de classificação e a regressão.

Clustering, agrupamentos ou segmentação – Difere-se da classificação, regressão e predição, por não possuir classe alvo. O objetivo consiste separar registros que possuem características semelhantes em subgrupos, ou clusters relativamente homogêneos. *Clustering* refere-se ao agrupamento de registros,

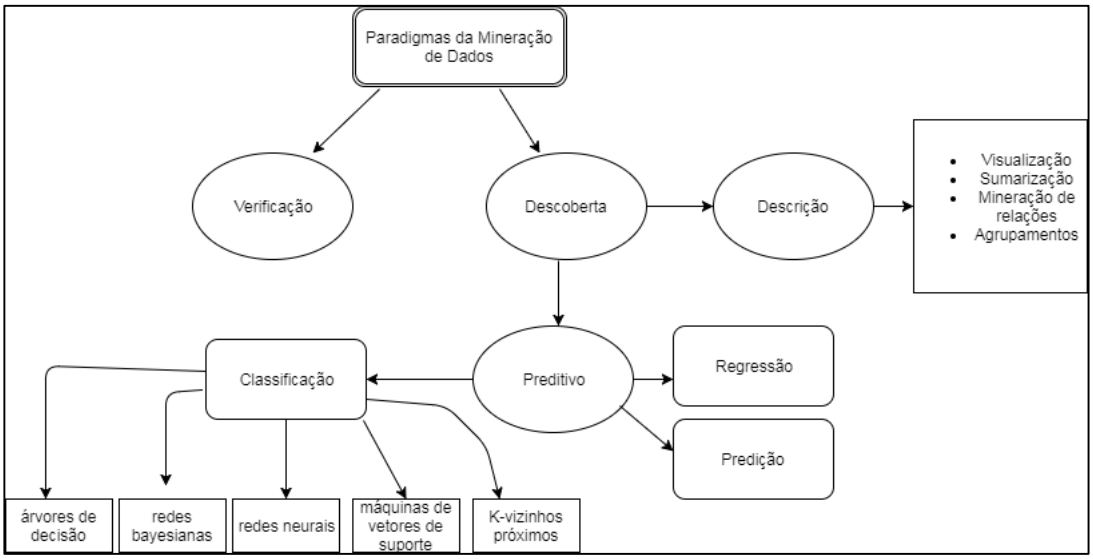
observações, ou classes de objetos semelhantes. Um cluster é um conjunto de registros que são semelhantes entre si e diferente dos registros em outros clusters.

Associação – Tarefas de associação, utilizam regras de associação para quantificar a relação entre dois ou mais atributos. Regras de associação possuem expressões na forma $X \rightarrow Y$, onde um antecedente que implica em um consequente.

Segundo Fayyad, Piatetsky-shapiro e Smyth (1996), o processo KDD, possui dois objetivos distintos, e a escolha depende das pretensões do pesquisador. Os objetivos podem ser de verificação e descoberta. Se for escolhida a verificação, o pesquisador se limita a aplicar técnicas, como análises estatísticas para verificar hipóteses definidas previamente. Caso contrário, opta-se pela verificação de padrões, sem fazer nenhuma suposição inicial.

A Figura 5 sintetiza tudo o que foi descrito, acrescentando alguns detalhes, dando ao leitor um resumo do que se pode chamar de paradigmas da mineração de dados. Ressaltando que as técnicas de classificação presentes na Figura 5, são as técnicas escolhidas para este trabalho, e que existe uma variedade de técnicas para essa finalidade.

Figura 5: Resumo dos paradigmas da mineração de dados.



Fonte: Própria.

O processo KDD foi encarado como sinônimo da mineração de dados na Figura 5, simplesmente para agregar contexto em uma demonstração mais generalista, deixando o entendimento mais claro e amplo.

A Figura 5, possui alguns termos da técnica descritiva que ainda não foram comentados. Como por exemplo a visualização e a sumarização.

A visualização e sumarização estão bastante relacionadas, se completam. A visualização consiste na exibição dos dados, por exemplo em gráficos de duas dimensões, ou mais. Já a sumarização consiste na caracterização dos dados, permitindo destacar de forma mais acentuada a relevância dos dados. Essas técnicas ainda serão apresentadas quando forem empregadas em capítulos posteriores.

A tarefa de associação é uma das técnicas que está inclusa nesse segmento de DM. A tarefa de associação será devidamente apresentada na subseção 2.2.4, assim como as técnicas de classificação, presentes na Figura 5, na subseção 2.2.3.

Para este trabalho a abordagem da descoberta foi adotada, inicialmente com as técnicas de descrição. As técnicas de descrição selecionados para esse trabalho foram as de visualização, sumarização e associação. Em seguida, já com certo conhecimento sobre os dados, devido a aplicação das técnicas de descrição, entra em cena a abordagem preditiva, com o emprego da classificação.

Todo o processo de mineração de dados realizado neste trabalho será detalhado devidamente dentro da sequência do processo KDD nos próximos capítulos.

2.2.3 Técnica de classificação

Como já foi exposto na subseção anterior, a classificação, ferramenta da DM, foi escolhida como um dos mecanismos que estão à frente na busca de padrões de dados, tendo como objetivo, classificar um conjunto de dados de acordo com determinada classe.

As técnicas de classificação analisam a saída, ou seja, a classe, de um conjunto de dados, com o intuito de aprender mais sobre os dados. A Classificação pode ser chamada de método de aprendizado de máquina supervisionado, porque usa a classe para a detecção de padrões (WRITTEN; FRANK, 2005).

Para ficar mais claro o conceito de classe, veja o exemplo dado na Figura 6. Na figura podemos perceber uma base de dados que possui cinco variáveis: *Outlook* (percepção, nesse caso, do tempo climático), *Temperature* (temperatura), *Humidity* (humidade), *Windy* (ventania) e *Play* (jogo). Sendo a variável *Play* a classe escolhida. Os valores da classe são binários categóricos: *yes* ou *no*. A classe pode ser chamada de variável independente. As demais são chamadas de variáveis dependentes. Sendo que *Outlook* e *Windy* possuem valores categóricos, e *Temperature* e *Humidity* valores

numéricos. Os valores de todas as variáveis, inclusive a classe, podem ser chamados de atributos. Neste exemplo a variável *Play* foi escolhida como classe para a verificação de algum jogo ou brincadeira (sim ou não), que será determinada de acordo com as características dos atributos das demais variáveis. Dessa maneira, os padrões encontrados poderiam influenciar na tomada de decisão, de ir ou não para o jogo ou brincadeira, por exemplo.

Figura 6: Seleção de classe em uma base de dados.

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Fonte: WITTEN e FRANK (2005).

Existe uma variedade de técnicas de classificação, no entanto, para este trabalho serão aplicadas as seguintes técnicas de classificação: J48, *instance-bases learning with parameter k* (IBK), *Sequential Minimal Optimization* (SMO), *Naive Bayes* e *Multilayer Perceptron*.

Essas técnicas foram escolhidas devido a suas características parecidas no quesito dos tipos de classes e atributos que são suportados, com pequenas diferenças, portanto possibilitando uma futura comparação nos resultados.

Os tipos de classes e atributos suportados pelas técnicas que são usadas, estão sinalizadas nos quadros 2 e 3. Vale ressaltar que os tipos de dados podem depender de como os algoritmos foram implementados (adequações necessárias), podendo ter pequenas variações, dependendo da ferramenta de mineração de dados utilizada.

A ferramenta de DM utilizada neste trabalho foi o WEKA. O próprio programa fornece as informações de dados suportados para cada classificador. Sendo descrita no capítulo de metodologia.

Quadro 2: Tipos de classes suportadas.

	Classificador	Classe Nominal	Classe binária	Classe faltando valores	Classe numérica	Classe de data.
J48	Tree	✓	✓	✓		
Naive bayes	Bayes	✓	✓	✓		
MLP	Function	✓	✓	✓	✓	✓
IBK	Lazy	✓	✓	✓	✓	✓
SMO	Function	✓	✓	✓		

Fonte: Própria.

Quadro 3: Tipos de atributos suportados.

	Atributos de datas	Atributos Unários	Atributos nominais vazios	Atributos faltando valores	Atributos Binários	Atributos Nominais	Atributos Numéricos
J48	✓	✓	✓	✓	✓	✓	✓
Naive bayes		✓	✓	✓	✓	✓	✓
MLP	✓	✓	✓	✓	✓	✓	✓
IBK	✓	✓	✓	✓	✓	✓	✓
SMO		✓	✓	✓	✓	✓	✓

Fonte: Própria.

A seguir uma descrição breve dos métodos de classificação presentes no trabalho. Todas as técnicas que serão descritas pertencem a aprendizagem supervisionada.

O primeiro classificador a ser descrito, utiliza regras no formato IF/THEN para a generalização do modelo criado, ou seja, se determinado caso, então pertence a determinado classe. Por exemplo: se amanhã estiver fazendo sol, então posso ir à praia.

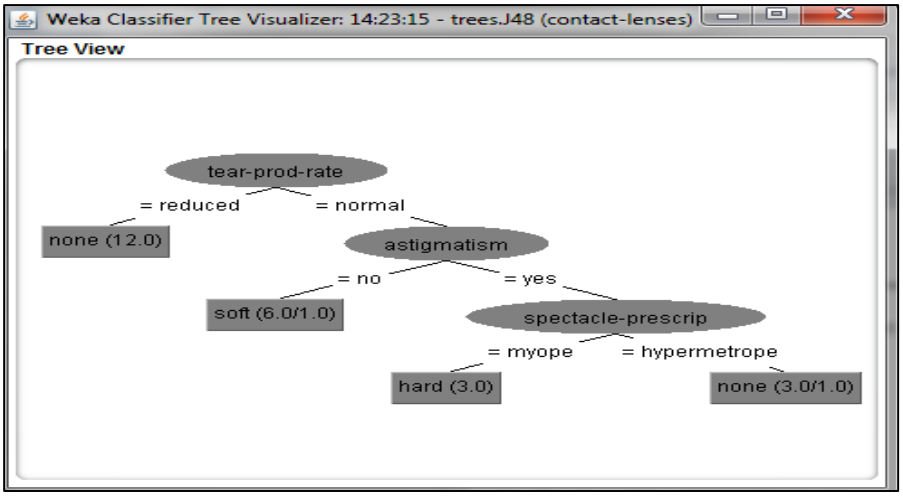
O j48 é uma extensão do ID3. É um algoritmo que se baseia na criação de uma árvore de decisão. Uma implementação de código aberto, feita em Java, do algoritmo C4.5, estando presente na ferramenta de mineração de dados, WEKA (KAUR; CHHABRA, 2014).

A árvore de decisão possui um nó raiz, nós internos e folhas. O nó raiz não possui borda de entrada, podendo possuir zero, ou mais bordas de saída. Os nós

internos, possuem exatamente uma borda de entrada, e na saída, uma ou mais bordas, são nesses nós que são feitos os testes nos atributos. Os nós chamados de folhas, possuem exatamente uma borda de entrada e não tem bordas de saída. Os nós folhas, denotam as classes. A árvore de decisão possui duas fases, fase de construção da árvore, ou do modelo, e a validação da árvore (SAKHARE; JOSHI, 2014).

As bordas citadas tanto as de entrada e saída podem ser consideradas como aberturas nos nós para as arestas. Arestas na representação da árvore podem ser entendidas como os galhos. Observe a Figura 7 para uma melhor compreensão.

Figura 7: Árvore gerada pelo J48 na ferramenta WEKA.



Fonte: Própria.

A abordagem adotada para a criação da árvore é a de “dividir para conquistar”. Primeiramente, é selecionado um atributo para ser o nó raiz. Esse nó raiz, cria um ramo para cada valor possível, dessa maneira, dividindo em subconjuntos. Esse processo é repetido em cada ramo. O processo para, quando todas as instâncias restantes tiverem a mesma classificação em um nó, caso contrário, enquanto restar instâncias para serem classificadas, novos ramos e nós são criados (WITTEN; FRANK, 2005).

Na fase da construção da árvore é feita a análise dos dados de treino (*training set*), e na validação da árvore, a criação do modelo, que será usado para classificação futura, ou objetos que ainda não foram classificados (GOYAL; MEHTA, 2012).

A base do algoritmo consiste nos conhecimentos das propriedades da *shannon entropy*, e *the gain information*. Para mais detalhes, consulte o trabalho de Hssina et al. (2014).

O próximo classificador a ser descrito está presente na categoria dos que possuem um aprendizado baseado em instâncias. Instancias podem ser encaradas como um conjunto de características presentes em cada linha de um conjunto de dados, ou seja, cada linha é uma instância. Esses algoritmos são conhecidos por armazenarem todas as instâncias da fase de treinamento para ser utilizado na classificação de uma nova instância.

IBK é um classificador da categoria *lazy* (aprendizado preguiçoso) que usa a técnica *k-nearest-neighbour*, ou k- vizinhos próximos, onde utiliza-se a distância como métrica. A quantidade K de vizinhos pode ser determinada automaticamente, ou escolhida. A distância é utilizada como parâmetro do método de busca, podendo ser utilizada a distância euclidiana, Chebyshev, Manhattan e Minkowski. (VIJAYARANI; MUTHULAKSHMI, 2013)

Trata-se de modificação do algoritmo original *nearest-neighbour*, que considera apenas um vizinho mais próximo. O parâmetro K permite uma classificação mais refinada, mas isso depende de cada base de dados. Um novo padrão é estimado, quando uma classe aparece com maior frequência dentre seus k-vizinhos. (BEZERRA, 2006).

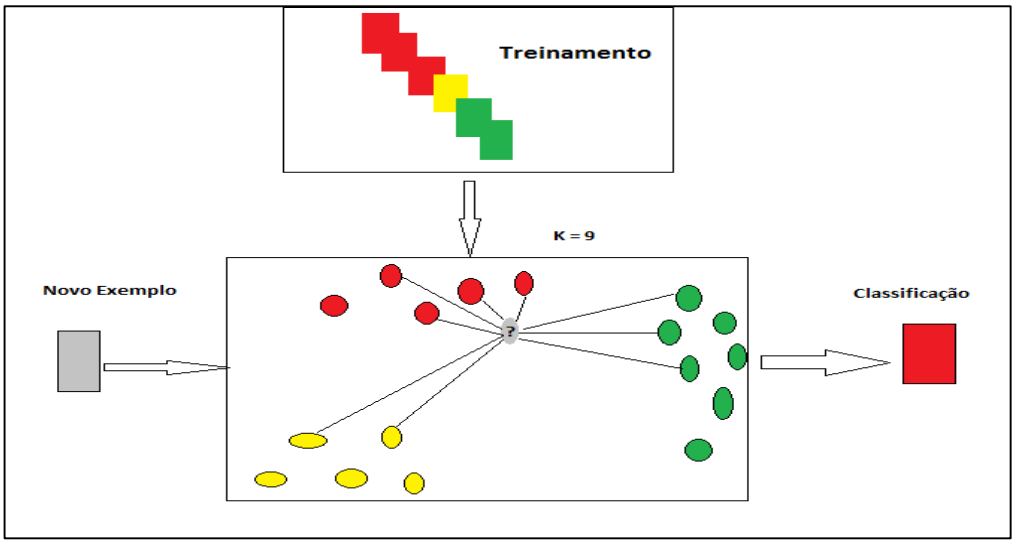
De acordo com Sutton (2012), pode-se resumir o algoritmo nos seguintes passos: Um inteiro positivo K é especificado, juntamente com uma nova amostra, seleciona-se as k entradas da base de dados que estão mais próximas da nova amostra, encontra-se a classificação mais comum dessas entradas, e esta é a classificação que damos à nova amostra.

A aprendizagem entre os K vizinhos pode ser representada em um plano de duas dimensões, podendo mostrar as conexões formadas com as métricas ou áreas com certo alcance, delimitadas por fronteiras, ficando visível dessa forma, as classes em que determinadas instâncias pertencem.

Dentro das pesquisas que foram feitas, não foi encontrado incorporado na ferramenta WEKA, a estrutura de visualização do classificador IBk, nem em pacotes ou *plugins* para esse fim. Contudo a falta desse recurso não trará prejuízo para o desenvolvimento e conclusões do trabalho, já que a ferramenta possui outros parâmetros que auxiliarão na busca do conhecimento diante dos resultados.

A Figura 8 traz uma representação da fase de treinamento do IBk, deixando mais claro de perceber o que foi descrito. As instâncias são colocadas em um plano, e de acordo com o K e o cálculo métrico estabelecidos, novas instâncias são classificadas.

Figura 8: Fase de treinamento do KNN.



Fonte: Própria.

O próximo classificador é do tipo *function*, no qual, utiliza funções para criar as generalizações. Assim como as árvores de decisão, o *Sequential media optimization* (SMO), próximo classificador a ser descrito, usa a abordagem dividir para conquistar.

O SMO é uma implementação chamada de otimização mínima sequencial, criada para resolver um problema de otimização da programação quadrática, do inglês, *quadratic programming* (QP). Com o SMO é possível repartir o problema da QP em uma série de “QPs” menores, aumentando a eficiência na resolução do problema. O problema da QP, está presente na construção das máquinas de vetores de suporte, ou, *Support Vector Machines* (SVMs), (PLATT, 1998).

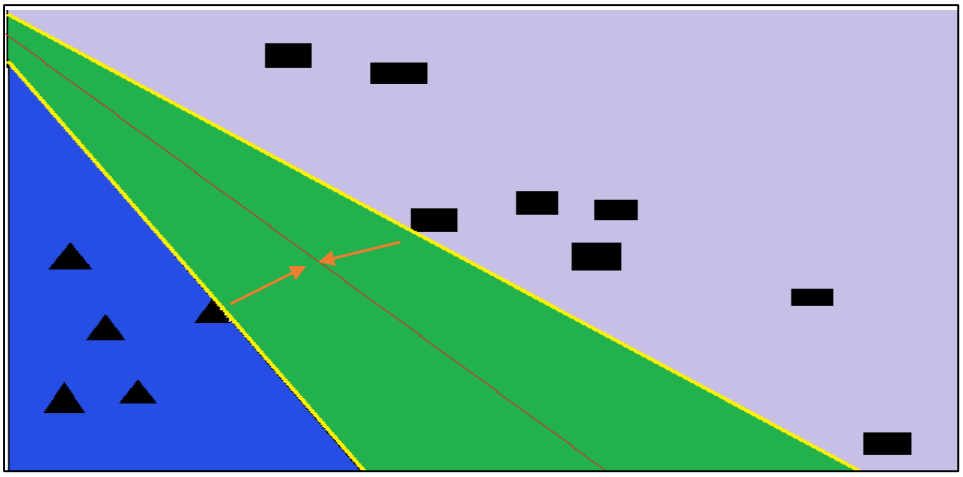
Uma SVM é uma ferramenta para resolver problemas de reconhecimento de padrões e problemas de regressão (ALI, 2003).

As SVMs são baseadas na teoria do aprendizado estatístico, consiste em construir um hiperplano como superfície de decisão, onde as classes sejam separadas com uma margem máxima, aumentando a distância entre as classes (GIRARDELLO, 2010).

O algoritmo SMO, utiliza a estratégia dividir para conquistar, e resolve o problema de forma analítica para cada parte dividida. A primeira parte do treinamento de uma SMO, envolve a determinação de dois multiplicadores de Lagrange, e segue com atualizações das entradas e de funções, com testes de verificação, calcula-se o desvio entre a saída da função e a classificação alvo, então é verificado se a classificação alvo é menor que o *threshold*, se sim, o algoritmo termina, se não, continua com os testes (HUANG; YAN, 2014).

A Figura 9 apresenta uma representação do SMO em um plano de duas dimensões, com duas classes, onde percebe-se que a técnica empregada é capaz de separar ao máximo as classes. Onde a linha vermelha seria a divisão do melhor hiperplano, que foi criada com base nas linhas amarelas, que são os vetores de suporte. Dessa forma os triângulos e quadrados podem ser classificados mediante certa divisão.

Figura 9: Representação de um hiperplano com duas classes.



Fonte: Própria.

Em seguida, o classificador *Naive Bayes*, presente na classe *Bayes,* onde utiliza conhecimentos probabilísticos.

Naive Bayes é um classificador do tipo Bayes, baseado no teorema de Bayes. Trabalha com a modelagem de incerteza através do uso das probabilidades. A forma de representação, é em grafos, chamados de redes bayesianas. As redes bayesianas possuem nós acíclicos orientados. Os nós representam variáveis aleatórias. Além dos nós, os grafos possuem arcos. Os arcos têm a função de unir dois nós, e representa uma dependência probabilística entre as variáveis associadas (BRAZILIAN SYMPOSIUM ON COMPUTER GAMES AND DIGITAL ENTERTAINMENT, 2008).

O aprendizado do classificador, ocorre por meio da independência condicional, no qual todos os atributos são independentes dada a classe. No documento feito por Leung (2007), é explicado de forma sucinta a classificação.

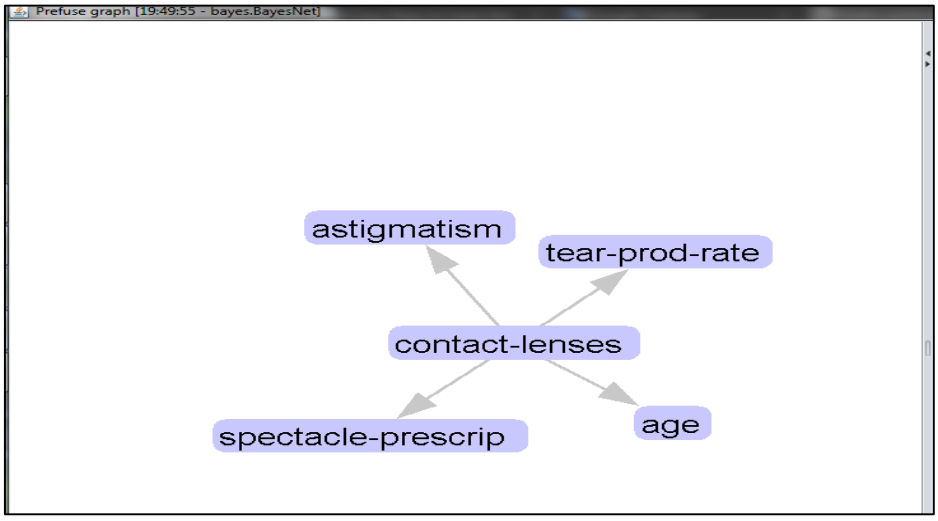
Esse o método é chamado de "*naive",* ou seja, ingênuo, por assumir que determinada característica particular de uma classe não está relacionada com quaisquer outras características, tratando os atributos de forma independente, mesmo que essas características dependam de outras.

Apesar da visão ingênua, essa técnica tem bons resultados. No entanto, pode ter um desempenho não tão satisfatório em bases de dados que possuem valores de atributos redundantes, porque um atributo com valor muito repetido pode influenciar na decisão da classificação, já que são os atributos são tradados de forma independente (WITTEN; FRANK, 2005).

Por exemplo, na classificação da fruta laranja com valores de atributos, verde, redonda, e quatro de diâmetro. Cada atributo, possui independentemente uma probabilidade que auxiliará na classificação. Um limão, que é verde e redondo, mas possui três de diâmetro, pode ser classificado como sendo uma laranja, já que possui características muito parecidas com as da laranja.

A Figura 10, traz uma representação da rede bayesianas do exemplo das lentes de contato, já citada. Perceba que essa representação não possui valores associados para uma maior interpretação, apenas a visualização dos atributos e da classe. A estrutura foi retirada do WEKA.

Figura 10: Representação das redes bayesianas.



Fonte: Própria.

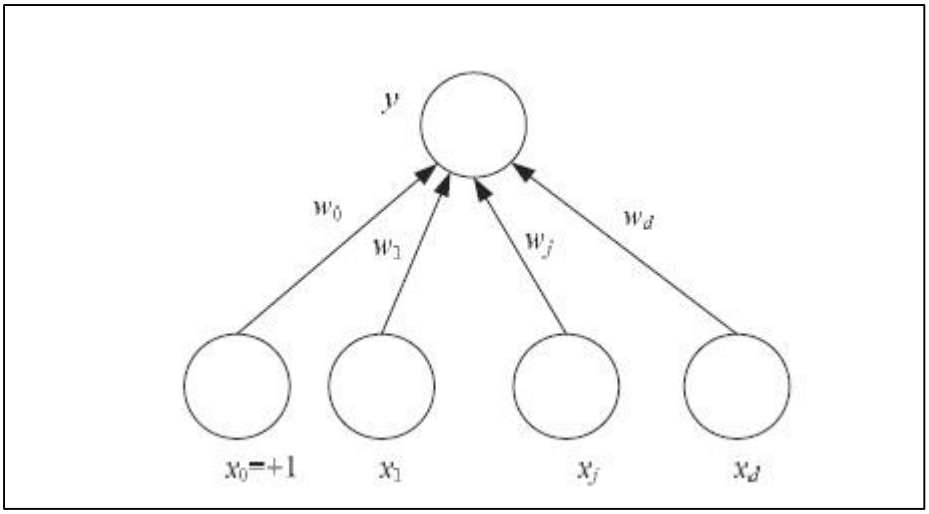
A última técnica a ser apresentada, é o *Multilayer Perceptron,* ou *perceptrons* de múltiplas camadas. Pertence aos classificadores do tipo *function*.

O *Multilayer Perceptron* (MLP), é a representação de *perceptrons* em um gráfico de nós e arestas ponderadas, que pode ser chamado de “rede de neurônios” (WITTEN; FRANK, 2005).

Perceptrons são elementos básicos de processamento de uma “rede de neurônios”. Esse elemento possui, unidades de entradas, um *bias*, configurado com valor igual um, que “manipula” a função de ativação, uma unidade de saída, e pesos que conectam diretamente a entrada e a saída (ALPAYDIN, 2010).

A representação de um simples *perceptron* é demonstrada na Figura 11. Onde se tem as unidades de entradas representadas pela letra X, onde o primeiro nó da esquerda para direita simboliza o *bias.* Em seguida se tem os pesos atribuídos para cada entrada, letra W. E saída sendo a letra Y.

Figura 11: Representação de um *perceptron*.



Fonte: Alpaydin (2010).

Dessa forma o MPL consiste em um sistema de simples neurônios, ou nós, interconectados. Os nós estão conectados por pesos e sinais de saída, que são uma função da soma das entradas para o nó modificado, seja por uma função transferência não linear simples, ou função de ativação (GARDNER; DORLING, 1998).

A modificação de um nó representa que esse nó está recendo um “sinal”, simbolizando o impulso elétrico de um neurônio real, e ocorrendo a chamada sinapse, que ocorre de um neurônio para outro. A função de soma é ponderada pelos pesos

associados a cada entrada. Então, para que haja essa comunicação, um nó precisa sofrer uma transferência não linear simples, ou o termo mais empregado, função de ativação, que pode ser a própria função de soma.

A função de ativação indica sob quais condições as sinapses irão ser ativadas. Existem muitos tipos de funções de ativação, dentre elas, Linear, *McCulloch-Pits*, *Signum*, sigmoide, e rampa, sendo escolhidas por propósitos específicos (BAN; CHANG, 2013).

Os detalhes das funções não serão colocados por conta do escopo deste trabalho, contudo o leitor pode consultar as referências citadas para uma maior aproximação com relação ao tema, já que envolve conhecimentos específicos da estrutura dos neurônios cerebrais e de funções que fazem a operacionalidade de uma rede neural.

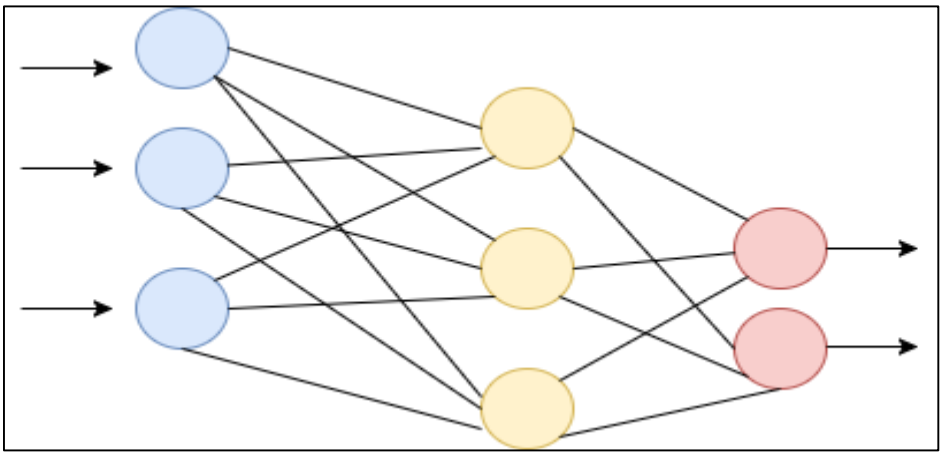
A comunicação entre os nós ocorre com o intuito de escolher a melhor opção de conexão para se chegar a uma resposta. No contexto do trabalho, a rede neural MPL é usada para chegar a uma melhor classificação.

O MLP é uma rede do tipo *feed-forward,* possuindo somente conexões entre camadas adjacentes*.* Essas redes geralmente possuem camadas ocultas entre as camadas de entrada e saída, e utilizam o algoritmo de aprendizado conhecido como *back propagation* (FUNAHASHI, 1988).

As camadas ocultas, são camadas de neurônios que podem ser ativadas diante de características significativas apresentadas nas entradas. Podendo ser chamadas de detectores de aprendizagem diante das características. A quantidade de camadas ocultas, influenciam diretamente no aprendizado da rede. Com muitas camadas ocultas a rede simplesmente consegue memorizar corretamente as respostas para cada padrão, em vez aprender uma solução generalizada (TOURETZKY; POMERLEAU, 1989).

A Figura 12 trata-se de uma representação de uma rede MLP, com conexões entre as camadas adjacentes, percebe-se que não existe ligações entre os neurônios da mesma camada. O exemplo possui uma única camada oculta.

Figura 12: Representação de um MPL com uma camada oculta.



Fonte: Própria.

O processo de treinamento, ou aprendizagem, ocorre por meio de uma função para avaliar o desvio dos valores de saída previstos, podendo essa função ser de erro, custo ou perda. Onde se tem os valores alvo, e um conjunto ideal de pesos (SILVA; SÁ; ALEXANDRE, 2008).

Durante a fase de aprendizado, os pesos são ajustados para ser capaz de predizer a classe correntemente diante das tuplas de entrada. O aprendizado da rede é chamado de conexionista devido às conexões entre as unidades (HAN; KAMBER; PEI, 2012).

Essas foram descrições compactas sobre as técnicas de classificação que serão aplicadas nesse neste trabalho, que ainda serão retomadas em capítulos posteriores, acrescentando conteúdos que completarão as suas funcionalidades, que trará ao leitor uma clareza maior sobre essas técnicas.

2.2.4 Técnica de associação

Outra técnica que será empregada nesta monografia, é a técnica de associação pertencente a abordagem descritiva da mineração de relações, pertencente ao método de aprendizagem não supervisionada.

Métodos de associação, são aqueles que utilizam regras de associação. Regras de associação possuem expressões na forma $X \rightarrow Y$, onde temos um antecedente que implica em um consequente, em que X e Y são itemsets. Itemsets,

são conjuntos com elementos, que podem ser frequentes (candidatos) ou não. ROMÃO et al. (1999).

Por exemplo, $\{pão, leite\} \rightarrow \{café\}$, além de ser uma regra de associação, são itemsets. Um itemset com K elementos é chamado de K-itemset. A ideia do método é mostrar que o primeiro conjunto pode implicar no segundo conjunto, ou seja, utiliza a lógica proposicional. O contrário não quer dizer a mesma coisa, $\{café\} \rightarrow \{pão, leite\}$, seria outra regra (AMO, 2018).

Um conceito que compõe os algoritmos de associação é o de suporte. Podemos chamar de suporte a frequência com que um determinado conjunto pode aparecer em um cenário escolhido. Outro conceito envolvido é o de confiança que é a quantidade de vezes que determinados elementos escolhidos aparecem juntos nos itemsets ($X \cup Y$), dividido pela quantidade de aparições de X nos itemsets. (VASCONCELOS; CARVALHO, 2004). A seguir as fórmulas, para uma melhor compreensão:

$$Suporte = \frac{Frequência\ \ de\ X\ e\ Y}{Total\ de\ T}$$

$$Confiança = \frac{Frequência\ \ de\ X\ e\ Y}{Frequência\ de\ X}$$

Na fórmula do suporte, o $Total\ de\ T$, refere-se ao total de transações. Entende-se por transação a quantidade de linhas de um determinado agrupamento de dados. (VASCONCELOS; CARVALHO, 2004).

Uma boa regra de associação tem sua confiança maior que o grau mínimo de confiança, e o suporte maior que o grau mínimo de suporte. O grau mínimo de suporte e de confiança são determinados pelo usuário (AMO, 2018).

2.2.4.1 Apriori

O algoritmo Apriori é um dos mais utilizados quando se trata de mineração de dados voltada para geração de regras de associação. Usando a abordagem Depth-First Search ou busca de profundidade, o algoritmo é capaz de gerar conjuntos de

itemsets candidatos (padrões) de K elementos a partir de conjuntos de itemsets de $K-1$ elementos. (VASCONCELOS; CARVALHO, 2004).

O objetivo do algoritmo é encontrar relações entre os dados enquanto eles são separados. À medida que esses possíveis candidatos a itemsets são gerados, o algoritmo também calcula o valor do suporte e da confiança, verificando se os possíveis candidatos a itemsets estão dentro dos valores mínimos permitidos. A cada iteração, a lista de elementos de um candidato a itemset aumenta em um elemento, o algoritmo termina quando o conjunto de K-1 itemsets encontrado seja vazio. (VASCONCELOS; CARVALHO, 2004).

2.2.5 Mineração de dados educacionais

Ao longo da última década, diante do crescimento e disponibilidade de dados sobre educação escolar em seus diferentes níveis, ficou visível a necessidade da exploração desses dados, que trazem consigo a busca de entender comportamentos, resultados, fenômenos e questões diversas que abrangem o ambiente educacional, surgindo assim, a mineração de dados voltada para dados educacionais.

Mineração de dados educacionais, do inglês, *educational data mining* (EDM), é uma subárea da DM, e tem o objetivo de tentar entender como acontece o aprendizado dos estudantes, identificando configurações ou modelos que revelem, quais variáveis estão envolvidas com os resultados educacionais, sejam positivos ou negativos, de modo que, possibilite um direcionamento sobre o tema, descrevendo e explicando fenômenos educacionais. (ROMERO; VENTURA, 2013).

Há duas linhas principais de trabalhos em que a EDM vem contribuindo, são: análise de dados e criação de modelos que visa compreender melhor os processos de aprendizagem e o desenvolvimento de métodos mais eficazes voltados para softwares educacionais (principalmente por causa da educação à distância), que deem suporte à aprendizagem do aluno. (BAKER; ISOTANI; CARVALHO, 2011).

A mineração de dados educacionais, utiliza as técnicas tradicionais da mineração de dados, como, classificação, clustering e associação, obtendo resultados de sucesso no domínio educacional, contudo não se limita a essas técnicas. Os métodos provenientes da EDM, consistem em destilação de dados, descobertas com

modelos, *knowlede tracing* (KT), e *nonnegative matrix factorization*. (ROMERO E VENTURA, 2013).

Na visão de Baker e Yacef (2009), o trabalho da EDM está classificado em: Predição, utilizando as técnicas de classificação, regressão e *density estimation*. *Clustering* ou agrupamento. Mineração de relações, que incluem, regras de associação, correlações, sequências e causas. Destilações de dados para julgamento humano (decisões). E por fim, as descobertas com modelos.

Baker e Yacef (2009), ainda apresentam as principais áreas de aplicações da EDM, sendo as seguintes: modelagem de estudantes, modelos de estrutura de conhecimento de um domínio, suporte pedagógico, e busca por evidências empíricas.

A modelagem de estudantes, consiste na criação de um modelo, que possibilite a verificação de aspectos importantes que representem informações das características do estudante, como, motivação, metacognição, conhecimento atual em determinada disciplina, dificuldades, que possam ajudar na análise de certos acontecimentos, como por exemplo, a retenção, e a reprovação em disciplinas. A predição de tais acontecimentos pode ajudar a comunidade escolar a tomar medidas que possam amenizar esses problemas, e assim, melhorar significativamente o desempenho do aprendizado do aluno. O modelo ainda pode ser aplicado em softwares educacionais que acompanham o desempenho dos estudantes em atividades educacionais, possibilitando feedback em tempo real.

Modelos de Estrutura de conhecimento de um domínio, são formados por conhecimentos específicos de um domínio. Tenta-se descobrir ou aperfeiçoar um determinado modelo de domínio.

A EDM pode servir para um suporte pedagógico, onde tenta verificar qual melhor suporte pedagógico se encaixa em determinada situação, qual é o mais efetivo para certo grupo.

A quarta área da EDM é a busca por evidências empíricas, refinando e estendendo teorias educacionais e fenômenos do âmbito educacional, visando obter uma compreensão mais profunda dos fatores-chave que afetam a aprendizagem, podendo apresentar soluções para sistemas de aprendizagem, como parte do objetivo.

2.3 DADOS ABERTOS GOVERNAMENTAIS

Com a devida cobrança da população e órgãos que pregam a transparência governamental, os principais governos democráticos do mundo, estão tentando ao longo dos últimos anos, disponibilizar dados para a população, com isso, surge a nomenclatura de dados abertos governamentais.

Para uma definição mais formal, de acordo com o *Worl Wide Web Consortium* (W3C), Dados Abertos Governamentais, “são a publicação e disseminação das informações do setor público na web, compartilhadas em formato bruto e aberto, compreensíveis logicamente, de modo a permitir sua reutilização em aplicações digitais desenvolvidas pela sociedade” (W3C, 2018).

Como trata-se de um conceito relativamente novo, vários autores, enxergavam que os governos começavam a demonstrar uma abertura no compartilhamento de dados sobre o governo, e procuraram estabelecer diretrizes e formulações de conceitos envolvendo os dados abertos do governo, um deles, descreveu as três leis dos dados abertos governamentais.

As três leis fundamentais dos dados abertos governamentais, foram propostas por Eaves (2009), e são as seguintes:

1. Se o dado não for encontrado e indexado na web, ele não existe.
2. Se ele não estiver disponível em formato aberto e compreensível por máquina, ele não poderá ser reaproveitado.
3. Se um *framework* legal não permitir que os dados sejam reutilizados, ele não é útil.

Um grupo de trinta advogados que defendiam em 2007, um governo mais aberto e transparente com relação aos dados do governo, se encontraram em Sebastopol, na California (EUA), para estabelecer uma compreensão mais profunda sobre os dados abertos do governo, e estabeleceram os oito princípios dos dados abertos governamentais (MALAMUD, 2007).

Os oito princípios dos dados abertos governamentais que foram firmados são:

1. Dados Completos. Todos os dados públicos devem estar disponíveis. Os dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou privilégios.

2. Dados Primários. Os dados devem ser apresentados de acordo a fonte, com o maior nível possível de granularidade, não em forma agregada ou modificada.

3. Dados Apropriados e atualizados. Os dados devem ser disponibilizados tão rapidamente, quanto necessário para preservar o valor dos dados.

4. Dados Acessíveis. Os dados devem estar disponíveis para o mais amplo alcance de usuários para a mais ampla gama de propósitos.

5. Dados compreensíveis por máquinas. Os dados devem estar razoavelmente estruturados para permitir o processamento automatizado.

6. Dados não discriminatório. Os dados devem estar disponíveis para qualquer pessoa, sem exigência de registro.

7. Dados não proprietários. Os dados devem estar disponíveis em um formato sobre o qual nenhuma entidade tenha controle exclusivo.

8. Dados Livres de Licença. Os dados não devem estar sujeitos a qualquer norma de direitos autorais, patentes, marcas comerciais ou segredos comerciais. Podem ser permitidas restrições razoáveis de privacidade, segurança e privilégios.

Com o acesso aos dados governamentais, muitos projetos e aplicações podem ser criadas, em diferentes áreas, com a motivação de trazer algo novo, ou de utilidade para a sociedade.

No manual dos dados abertos: governo, de W3c (2018), é citado algumas áreas e atividades em que os dados abertos estão sendo utilizados. Entre essas áreas estão:

- Transparência e controle democrático;
- Participação popular;
- Empoderamento dos cidadãos;
- Melhores ou novos produtos e serviços privados;
- Inovação;
- Melhora na eficiência de serviços governamentais;
- Melhora na efetividade de serviços governamentais;
- Medição do impacto das políticas;

- Conhecimento novo a partir da combinação de fontes de dados e padrões.

O Brasil ainda tem um caminho longo para ser percorrido quando se trata de dados abertos, tanto em projetos e aplicações que podem ser criadas com os dados que já estão disponíveis, quanto no quesito de deixar cada vez mais transparente os dados governamentais, e na facilidade de serem encontrados. Contudo, o país pode-se orgulhar de já ter dado uma longa caminhada.

Segundo a edição anual do Índice de Dados Abertos, produzida pela Open Knowledge e Diretoria de Análise de Políticas Públicas da Fundação Getúlio Vargas, o Brasil, subiu quatro posições no ranking mundial de Dados Abertos, ficando na oitava colocação, e na liderança quando se trata da América Latina (FGV, 2017).

Os dados abertos governamentais, podem ser encontrados geralmente nos órgãos governamentais de interesse específicos de uma determinada área, por exemplo, dados sobre a educação, podem ser coletados no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

O Brasil também possui o Portal Brasileiro de Dados Abertos, onde reúne uma variedade de dados abertos, que até então, conta com mais de 2.800 conjuntos de dados, que estão disponíveis para uso e reuso pela sociedade, podendo apenas ter a exigência de dar os devidos créditos aos autores.

2.4 ENADE

O Exame Nacional de Desempenho dos Estudantes (ENADE), faz parte dos três eixos de avaliação das Instituições de Educação Superior (IES), do Sistema Nacional de Avaliação da Educação Superior (SINAES). O ENADE, consiste na avaliação do desempenho de estudantes, enquanto que os outros dois eixos, consistem na avaliação institucional e avaliação de cursos. (BARBOSA; FREIRE; CRISÓSTOMO, 2011).

Em 2004 com a lei 10.661/2004, o SINAES, criado pelo Ministério da Educação (MEC), começou a ser implantado, e no quesito desempenho acadêmico do aluno, o ENADE substituiu o antigo Exame Nacional de Cursos (ENC), popularmente conhecido como provão. (FREIRE; CRISÓSTOMO; CASTRO, 2008)

O ENADE é realizado pelo INEP, sendo a Comissão Nacional de Avaliação da Educação Superior (CONAES), órgão colegiado, quem estabelece as diretrizes do exame, coordenação e supervisão do SINAES. (MICRODADOS DO ENADE, 2014)

Os conteúdos presentes no exame, são de natureza de conteúdos curriculares dos cursos e formação em saberes gerais. Possui questões de diferentes complexidades, explorando os momentos da vida acadêmica do estudante. O ENADE é realizado anualmente, de acordo com a área do curso. Os cursos são divididos em três grupos, e cada grupo realiza a prova, a cada três anos. (RISTOFF; LIMANA, 2018).

O grupo I, é formado por áreas da saúde, ciências agrárias, e áreas afins. Ciências exatas, licenciaturas e áreas afins, compõe o grupo II. O grupo III é formado pelas ciências sócias aplicadas, ciências humanas e áreas afins.

Até o ano de 2011, eram avaliados estudantes ingressantes e concluintes, contudo a partir de 2014, somente estudantes concluintes começaram a ser avaliados pelo exame. A prova possui quarenta questões, dez relacionadas aos conhecimentos gerais e trinta aos específicos do curso, com questões de múltipla escolha e discursivas.

O ENADE, ainda possui, um questionário de impressões dos estudantes sobre a prova, um questionário dos estudantes, possuindo questões de nível socioeconômico e organização didático-pedagógica do curso, e um questionário do coordenador (a) do curso. (RISTOFF, 2014).

Quanto aos resultados, o exame traz indicadores de avaliações das IES, que estão se consolidando na análise da qualidade do curso. O primeiro deles, é o conceito ENADE, que é calculado a partir da padronização do desempenho médio dos concluintes nas provas de formação geral (25%) e componente específico (75%). O conceito é representado em uma faixa de classificação de 1 a 5 (NOVAES; SALES, 2015).

Além do conceito ENADE, existe o Conceito Preliminar de Curso (CPC) e o Índice Geral de Cursos (IGC). O conceito ENADE faz parte dos parâmetros estabelecidos para o cálculo do CPC e IGC. O corpo docente, a percepção do discente sobre as condições do processo formativo (questionário do ENADE), e desempenho dos estudantes, integram a composição do CPC. O IGC, é o indicador que representa as avalições realizados sobre todos os cursos de uma IES, envolve média dos últimos CPCs, média dos conceitos de avaliação dos programas de pós-graduação stricto

sensu, e distribuição dos estudantes entre graduação e pós-graduação stricto sensu. Tanto o CPC quanto IGC, são representados em uma faixa de classificação de 1 a 5. (BOLETIM ENADE, 2015).

O ENADE, o CPC e o IGC, são adicionados à uma lista de outros indicadores, que não serão apresentados neste trabalho, que sustentam os três eixos de avaliação das IES, e por isso não devem ser avaliados isolados para determinar a qualidade de uma IES ou curso.

Para este trabalho, serão utilizadas algumas questões do questionário socioeconômico, que serão apresentadas na seção de apresentação dos dados, e o questionário de percepção da prova, que foi utilizado na sua totalidade. Como já foi apresentado o intuito da pesquisa, então vale acrescentar que questões que não foram utilizadas, podem conter uma importância, que devido ao tempo estimado para a conclusão do trabalho, não foi possível acrescentá-las, e podem servir para o aprimoramento deste trabalho, principalmente as questões da parte da organização didático-pedagógica do curso.

.

3 MATERIAIS E METODOLOGIA

Neste capítulo será apresentado de forma mais detalhada a base de dados original que está sendo utilizada para a execução deste trabalho. Também consta neste capítulo como foi a preparação dos dados no processo KDD. Além de apresentar a ferramenta WEKA que foi essencial para a aplicação dos classificadores na fase de mineração de dados. A seção 3.1 apresenta dos dados utilizados. Seguida pela seção 3.2 que descreve a metodologia para o processo KDD. E finalizando com a seção 3.3, coma descrição básica da ferramenta WEKA, e das configurações principais dos classificadores, e os parâmetros que serviram para as análises. Dentro das seções contém subseções relacionadas a sua seção principal.

3.1 DADOS UTILIZADOS

A base de dados utilizada neste trabalho foi a do ENADE, correspondendo aos anos de 2008, 2011 e 2014. Referindo-se ao curso de ciência da computação da Universidade do Estado do Rio Grande do Norte (UERN), campus avançado de Natal.

Os dados foram retirados diretamente do Portal Brasileiro de Dados Abertos. Contendo a base de 2014, cento e cinquenta e seis variáveis, a de 2011, cento e vinte e sete, e a de 2008, cento e noventa e oito.

Esse conjunto de variáveis estão distribuídas nos questionários socioeconômico, didático pedagógico e de percepção da prova. Também são referentes ao perfil do aluno, as notas e gabaritos, à instituição e do curso, e da prova, nos aspectos de presença e inscrição.

Quanto a quantidade de registros dos alunos de ciência da computação da UERN, campus de Natal, o Quadro 4 contém a quantidade de estudantes que foram selecionados para fazer a prova, os que estavam presentes, e os ausentes.

Dos setenta e três alunos presentes no exame, vinte um eram ingressantes, e cinquenta e dois concluintes. O presente trabalho, como já foi citado, visa explorar os dados referentes aos alunos concluintes. Dessa forma a base de dados em estudo, conta com cinquenta e dois estudantes.

Quadro 4: Estudantes selecionados para o ENADE, Campus Natal.

Ano	Quantidade de alunos chamados para realizar o exame.	Quantidade de alunos presentes no exame.	Ausentes.	Concluintes.	Ingressantes.
2014	26	25	1	26	-
2011	19	17	2	19	-
2008	40	31	9	10	30
Total	85	73	12	55	30

Fonte: INEP.

3.1.1 Descrição dos dados

Quanto às questões selecionadas, foram escolhidas dez questões do questionário socioeconômico, e as nove questões presentes no questionário de percepção da prova. Os gabaritos das provas objetivas específicas também serão utilizados.

As questões escolhidas do questionário socioeconômico e o padrão de resposta adotado, estão presentes na Tabela 1.

As questões do questionário socioeconômico foram selecionadas de acordo com a compatibilidade e presença das mesmas nos anos em questão (visto que, algumas questões estão presentes em determinado ano, mas não nos demais), e algumas passaram por adaptações nas alternativas para se obter padrões de respostas de múltipla escolha comum a todos os anos.

As questões da Tabela 1 que contém asteriscos, são as que foram ajustadas, acomodando as respostas para todas as edições. Para ficar claro quais adaptações foram feitas, os quadros 5, 6 e 7, apresentam tais ajustes.

Tabela 1: Questões do questionário socioeconômico escolhidas.

Questão 1.Qual o seu estado civil?	A ( ) Solteiro(a). B ( ) Casado(a). C ( ) Separado(a) judicialmente/divorciado(a). D ( ) Viúvo(a). E ( ) Outro.
Questão 2. Como você se considera?	A ( ) Branco(a). B ( ) Negro(a). C ( ) Pardo(a)/mulato(a). D ( ) Amarelo(a) (de origem oriental). E ( ) Indígena ou de origem indígena.
***Questão 3. Qual a renda total de sua família, incluindo seus rendimentos?**	A ( ) Até 3 salários mínimos. B ( ) Mais de 3 até 10 salários mínimos. C ( ) Mais de 10 até 30 salários mínimos. D ( ) Acima de 30 salários mínimos. E ( ) Nenhum.
***Questão 4. Qual alternativa a seguir melhor descreve sua situação financeira (incluindo bolsas)?**	A ( ) Não tenho renda e meus gastos são financiados por programas governamentais. B ( ) Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas. C ( ) Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos. D ( ) Tenho renda e não preciso de ajuda para financiar meus gastos. E ( ) Tenho renda e contribuo com o sustento da família. F ( ) Sou o principal responsável pelo sustento da família.
Questão 5. Qual alternativa a seguir melhor descreve sua situação de trabalho (exceto estágio ou bolsas)?	A ( ) Não estou trabalhando. B ( ) Trabalho eventualmente. C ( ) Trabalho até 20 horas semanais. D ( ) Trabalho de 21 a 39 horas semanais. E ( ) Trabalho 40 horas semanais ou mais.
***Questão 6. Seu ingresso no curso de graduação se deu por meio de políticas de ação afirmativa ou inclusão social?**	A ( ) Não. B ( ) Sim, por critério étnico-racial. C ( ) Sim, por recorte social. D ( ) Sim, por sistema que combina dois ou mais critérios anteriores. E ( ) Sim, por sistema diferente dos anteriores.
***Questão 7. Em que tipo de escola você cursou o ensino médio?**	A ( ) Todo em escola pública. B ( ) Todo em escola privada (particular). C ( ) metade em escola pública e particular. D ( ) A maior parte em escola pública. E ( ) A maior parte em escola privada (particular).
Questão 8. Qual modalidade de ensino médio você concluiu?	A ( ) Ensino médio tradicional. B ( ) Profissionalizante técnico (eletrônica, contabilidade, agrícola, outro). C ( ) Profissionalizante magistério (Curso Normal). D ( ) Educação de Jovens e Adultos (EJA) e/ou Supletivo. E ( ) Outra modalidade.
Questão 9. Excetuando-se os livros indicados na bibliografia do seu curso, quantos livros você leu neste ano?	A ( ) Nenhum. B ( ) Um ou dois. C ( ) De três a cinco. D ( ) De seis a oito. E ( ) Mais de oito.
Questão 10. Quantas horas por semana, aproximadamente, você dedicou aos estudos, excetuando as horas de aula?	A ( ) Nenhuma, apenas assisto às aulas. B ( ) De uma a três. C ( ) De quatro a sete. D ( ) De oito a doze. E ( ) Mais de doze.

Fonte: Própria.

A questão três da Tabela 1, faz referência a questão oito da edição de 2014, a cinco em 2011, e a sete em 2008. A questão foi adaptada devido que, as alternativas de renda possuíam sutis diferenças nas três edições. Por exemplo, a primeira alternativa, era até três salários mínimos em 2008, nenhum salário em 2011, e até um salário mínimo e meio em 2014. Ficando a primeira alternativa (questão 3, Tabela 1), até três salários mínimos, englobando as edições de 2008 (alternativa A), 2011 (alternativas B e C), e 2014 (alternativas A e B). Deixando a opção “nenhum salário”, sendo a última alternativa.

O Quadro 5, contém todo o enquadramento geral das alternativas das edições de 2008, 2011, 2014, com as alternativas adaptadas.

Quadro 5: Enquadramento geral das alternativas para a questão três.

Questão 3 da tabela 1.	2008	2011	2014
A ( ) Até 3 salários mínimos.	A.	B e C.	A e B.
B ( ) Mais de 3 até 10 salários mínimos.	B.	D, E, e F.	C, D, e E.
C ( ) Mais de 10 até 30 salários mínimos.	C e D.	G	F
D ( ) Acima de 30 salários mínimos.	E	H	G
E ( ) Nenhum.	-	A	-

Fonte: Própria.

Outra questão da Tabela 1 que passou por adaptações nas alternativas, foi a quatro. A questão quatro, assumiu o modelo de alternativas de 2014, porque possuía uma questão a mais: a alternativa “Não tenho renda e meus gastos são financiados por programas governamentais”, que não estava presente em 2011 e 2008. A questão quatro refere-se a questão nove em 2014, a seis em 2011, e a nove em 2008.

As associações feitas das alternativas da questão quatro, estão presentes no Quadro 6.

A principal adaptação presente na questão seis da Tabela 1, foi a alternativa C com resposta: “Sim, por recorte social”, presente na edição de 2008, e não nas demais. Para contornar a situação, o critério de renda e ter estudado em escola pública, presentes em 2011 e 2014, passaram a ser enquadrados como recorte social. A questão seis refere-se as questões quinze em 2014, doze em 2011 e 2008.

O Quadro 7 traz as associações feitas, para um melhor enquadramento possível das alternativas para a questão seis.

Quadro 6: Enquadramento das alternativas para a questão quatro.

Alternativas da questão 4 da tabela 1.	2008	2011	2014
A ( ) Não tenho renda e meus gastos são financiados por programas governamentais.	-	-	Modelo igual ao da Tabela 1, questão 4.
B ( ) Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas	Não trabalho e meus gastos são financiados pela família.	Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas.	Modelo igual ao da Tabela 1, questão 4.
C ( ) Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos.	Trabalho e recebo ajuda da família	Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos.	Modelo igual ao da Tabela 1, questão 4.
D ( ) Tenho renda e não preciso de ajuda para financiar meus gastos.	Trabalho e me sustento	Tenho renda e me sustento totalmente.	Modelo igual ao da Tabela 1, questão 4.
E ( ) Tenho renda e contribuo com o sustento da família.	Trabalho e contribuo com o sustento da família.	Tenho renda, me sustento e contribuo com o sustento da família.	Modelo igual ao da Tabela 1, questão 4.
F ( ) Sou o principal responsável pelo sustento da família.	Trabalho e sou o principal responsável pelo sustento da família.	Tenho renda, me sustento e sou o principal responsável pelo sustento da família.	Modelo igual ao da Tabela 1, questão 4.

Fonte: INEP.

Quadro 7: Enquadramento das alternativas para a questão seis.

Alternativas da questão 6 da Tabela 1.	2008	2011	2014
A ( ) Não.	D	A	A
B ( ) Sim, por critério étnico-racial.	A	B	B
C ( ) Sim, por recorte social.	B	C e D.	C e D.
D ( ) Sim, por sistema que combina dois ou mais critérios anteriores.	-	E	E
E ( ) Sim, por sistema diferente dos anteriores.	C	F	F

Fonte: INEP.

A última adaptação que foi feita corresponde a questão sete da Tabela 1. Trata-se da adaptação mais simples, onde a alternativa F com a resposta, “Parte no Brasil e parte no exterior”, presente na edição de 2014, foi retirada porque ninguém havia marcado a questão. As demais alternativas de 2014 e dos outros anos foram consideradas já que eram compatíveis, embora a alternativa metade em escola pública e particular, só constava na edição de 2008.

Depois da apresentação das adaptações feitas nas questões do questionário socioeconômico, será a vez de mostrar as questões do questionário de percepção da prova. Na Tabela 2 estão presentes as questões de percepção da prova. As questões não passaram por modificações nas três edições do exame que estão sendo avaliadas. Não contando com alternativas diferentes, nem questões a mais.

Tabela 2: Questões do questionário de percepção da prova.

1- Qual o grau de dificuldade desta prova na parte de Formação Geral?	A = Muito fácil. B = Fácil. C = Médio. D = Difícil. E = Muito difícil.
2 - Qual o grau de dificuldade desta prova na parte do Componente Específico?	A = Muito fácil. B = Fácil. C = Médio. D = Difícil. E = Muito difícil.
3 - Considerando a extensão da prova, em relação ao tempo total, você considera que a prova foi:	A = Muito longa. B = Longa. C = Adequada. D = Curta. E = Muito curta.
4 - Os enunciados das questões da prova na parte de Formação Geral estavam claros e objetivos?	A = Sim, todos. B = Sim, a maioria. C = Apenas cerca da metade. D = Poucos. E = Não, nenhum.
5 - Os enunciados das questões na parte do Componente Específico estavam claros e objetivos?	A = Sim, todos. B = Sim, a maioria. C = Apenas cerca da metade. D = Poucos se apresentam. E = Não, nenhum.
6 - As informações/instruções fornecidas para a resolução das questões foram suficientes para resolvê-las?	A = Sim, até excessivas. B = Sim, em todas elas. C = Sim, na maioria delas. D = Sim, somente em algumas. E = Não, em nenhuma delas.
7 - Você se deparou com alguma dificuldade ao responder à prova. Qual?	A = Desconhecimento do conteúdo. B = Forma diferente de abordagem do conteúdo. C = Espaço insuficiente para responder às questões. D = Falta de motivação para fazer a prova. E = Não tive qualquer tipo de dificuldade para responder à prova.
8 - Considerando apenas as questões objetivas da prova, você percebeu que:	A = Não estudou ainda a maioria desses conteúdos. B = Estudou alguns desses conteúdos, mas não os aprendeu. C = Estudou a maioria desses conteúdos, mas não os aprendeu. D = Estudou e aprendeu muitos desses conteúdos. E = Estudou e aprendeu todos esses conteúdos.
9 - Qual foi o tempo gasto por você para concluir a prova?	A = Menos de uma hora. B = Entre uma e duas horas. C = Entre duas e três horas. D = Entre três e quatro horas. E = Quatro horas e não consegui terminar.

Fonte: INEP.

As questões do questionário de percepção da prova incluem perguntas como: o grau de dificuldade da prova, extensão da prova, aspectos de clareza e objetividade dos enunciados, dificuldades enfrentadas, se os conteúdos estavam de acordo com o que foi estudado pelo aluno, e o tempo gasto para concluir a prova.

Para concluir as descrições a respeitos dos dados, o Quadro 8 contém as demais variáveis selecionadas, com sua descrição respectiva.

Quadro 8: Variáveis selecionadas.

Variável	Descrição
nu_ano	Ano de realização do exame.
nu_idade	Idade do inscrito.
tp_sexo	Sexo do inscrito.
vt_esc_ofg	Escolha de reposta da parte objetiva da formação geral.
vt_esc_oce	Escolha de resposta da parte objetiva do componente específico.
nt_obj_fg	Nota bruta na parte objetiva da formação geral - Convertida para escala de 0 a 100.
nt_dis_fg	Nota bruta na parte discursiva da formação geral - Convertida para escala de 0 a 100.
nt_fg	Nota bruta na formação geral - Média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral (0 a 100).
nt_obj_ce	Nota bruta na parte objetiva do componente específico - Convertida para escala de 0 a 100.
nt_dis_ce	Nota bruta na parte discursiva do componente específico - Convertida para escala de 0 a 100.
nt_ce	Nota bruta no componente específico - Média ponderada da parte objetiva (85%) e discursiva (15%) no componente específico (0 a 100).
nt_ger	Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%) (0 a 100).

Fonte: Própria.

As variáveis do Quadro 8 foram selecionadas com intuito de se obter o perfil do estudante de acordo com as notas da prova, acrescidos dos questionários de percepção da prova e socioeconômico.

Além da questão de um perfil a ser analisado, juntamente com os questionários, outra vertente deste trabalho, corresponde a análise dos gabaritos para a verificação de quais questões os estudantes estão acertando diante da complexidade da questão, que podem ser classificadas em fácil, média, difícil, e muito difícil.

Quanto aos gabaritos das provas específicas, foram criadas vinte e sete variáveis para cada resposta.

3.2 METODOLOGIA

Primeiramente para a realização deste trabalho foi feito um levantamento bibliográfico relacionado aos conceitos e definições descritas no capítulo de fundamentação teórica, com foco na mineração de dados.

Em seguida, mediante a compreensão de quão ampla é a mineração de dados, foi decido seguir a metodologia KDD (apresentada no capítulo dois), como base sólida e tradicional para aplicação da mineração de dados.

O processo KDD é uma das metodologias mais usadas, tendo destaque também, as metodologias *Cross-Industry Standard Process for Data Mining* (CRISP-DM) e *Sample, Explore, Modify, Model, Assess* (SEMMA).

De acordo com Shafique e Qaiser (2014), o processo KDD é mais completo e preciso. Sendo a CRISP-DM e SEMMA específicas para empresas que integram essas metodologias em seus *softwares*.

Dessa maneira, para atender aos objetivos traçados no capítulo 1, as etapas presentes no processo KDD serão descritas nos aspectos deste trabalho nas subseções a seguir, com exceção da primeira parte do processo, que se refere ao tema a ser explorado, já descrito, assim como a compreensão do domínio, e as pretensões a serem alcançadas.

3.2.1 Seleção

Para este trabalho foram selecionadas trinta e uma variáveis para serem exploradas. De antemão, vale destacar que essas variáveis serão usadas diante da necessidade e relevância para o trabalho, diante dos objetivos traçados.

Com as planilhas abertas, começou o processo de seleção dos dados. As planilhas, foram baixadas do Portal Brasileiro De Dados Abertos. As planilhas continham os registros de todos os estudantes presentes nos cursos do grupo II, que fizeram o ENADE nos determinados anos.

A seleção inicialmente aconteceu com delimitação dos dados pela área de enquadramento do curso. Cursos relacionados a computação possuem o código 4004.

Em seguida, a filtração ocorre por meio do código da Instituição De Ensino Superior (IES), no caso da UERN, é o código 71. Possuindo nesse momento da seleção, os cursos de ciência da computação da UERN.

Depois, foi feita uma seleção por meio do código do curso, restringindo ao curso de ciência da computação do campus avançado de Natal. O código utilizado foi o 57606.

Esse procedimento foi feito para as três bases de dados em questão, criando dessa forma três bases separadas, que irão se juntar posteriormente para formar uma única base de dados.

3.2.2 Pré-processamento

Nessa fase, as bases de dados já estão unificadas. Sendo eliminados os estudantes ingressantes e os ausentes. Ficando presente na base, os alunos que preencheram pelo menos um dos questionários.

Os campos vazios foram sinalizados por um ponto. Específicamente para a parte da classificação, alguns filtros presentes na ferramenta WEKA foram utilizados.

O primeiro deles foi o *Replace Missing Values*, que calcula a média dos valores dos atributos para cada campo vazio. Em seguida as saídas das classes que não tinham valores, e continham apenas uma instância associada, foram eliminadas.

Em seguida o filtro *Numeric Transformation* foi aplicado para arredondar possíveis resultados do filtro anterior, com mais uma de casa decimal, para apenas uma casa decimal, ficando dessa forma de acordo com a classificação numérica apresentada na subseção 3.2.3.

O filtro *Numeric to Nominal* foi aplicado apenas na variável “ano” porque ela estava sendo considerada como numeral, e poderia causar comparações condicionais, como “maior que” em alguns classificadores, não restringindo os anos em questão.

Depois o filtro *Discretize* foi aplicado nas variáveis que eram referentes às notas. Criando cinco intervalos de frequência.

Os dados estavam desbalanceados, e testes prévios, a classificação por meio dos algoritmos não era satisfatória. Então, as bases de dados precisaram ser balanceadas.

Para o balanceamento dos dados foram aplicados a combinação de dois filtros: *synthetic minority over-sampling technique* (SMOTE) e *SpreadSubsample*. Primeiramente foi aplicado o SMOTE na *minority class* da base de dados, criando instâncias sintéticas da classe, por meio do algoritmo KNN. Logo depois, o *SpreadSubsample,* que seleciona uma amostra aleatória das demais classes para reduzir a quantidade de instâncias, e assim efetuar o balanceamento dos dados. Vale ressaltar que classes que continham apenas uma instância, foram desconsideradas, devido ao fato de não possuir o número mínimo para a aplicação do SMOTE.

Para finalizar a etapa de pré-processamento, foi aplicado o filtro *Ramdomize* para gerar uma aleatoriedade no conjunto de dados.

Não tendo mais nenhum problema para ser corrigido na fase de pré-processamento, a base de dados pode seguir para a próxima etapa no processo KDD.

3.2.3 Transformação

Nessa fase, a princípio, duas mudanças importantes foram feitas nos dados. A primeira foi a de criar intervalos de frequência para as notas de 0 a 100. Os intervalos estão contidos no Quadro 9. Essa medida servirá principalmente para a parte descritiva no que se refere ao aprendizado não supervisionado, e para as classes na fase da classificação.

Quadro 9: Classificação das notas em intervalos de frequência.

Nota	Classificação
De 0 até 20	A
De 21 até 40	B
De 41 até 60	C
De 61 até 80	D
De 81 até 100	E

Fonte: Própria.

Vale ressaltar que a base original relacionada as notas, irá ser utilizada na parte da mineração de dados voltada para a descrição, onde se utiliza algumas técnicas estatísticas, trazendo melhores resultados com os valores numéricos presentes nas notas.

Dessa forma, os valores categóricos presentes nos questionários foram convertidos para valores numéricos como mostra o Quadro 10. Vale ressaltar que os dois modelos, tanto o categórico quanto o numérico serão utilizados.

A segunda transformação trata-se da divisão das variáveis dos gabaritos objetivos de formação geral (vt_esc_ofg) e formação específica (vt_esc_oce) em várias variáveis, totalizando vinte e sete para as questões objetivas de formação específica e oito para de formação geral.

O Quadro 11, apresenta a classificação para a variável idade de acordo com um intervalo de frequência. E dessa forma, encerra-se a etapa de transformação do processo KDD para este trabalho.

Quadro 10: Conversão de valores categóricos para numéricos.

Valor categórico	Valor numérico
.	
a	1
b	2
c	3
d	4
e	5
f	6

Fonte: Própria.

Quadro 11: Classificação para a variável idade.

Idade	Classificação
21-24	A
25-28	B
29-32	C
>=33	D

Fonte: Própria.

3.2.4 Mineração de dados e análises

O processo metodológico foi descrito anteriormente, de maneira mais enfática, para que no próximo capítulo, a fase da mineração e análise ganhe praticamente todo o destaque. Depois da devida importância para as fases anteriores do processo KDD, chega a vez de aplicar a mineração de dados.

Nessa fase, para este trabalho, a MD será direcionada em duas abordagens: a descritiva, e a preditiva.

A abordagem descritiva conta com técnicas de visualização dos dados. As que serão utilizadas são: visualização em gráficos de duas dimensões, histogramas, gráficos de caixa, e correlações.

O algoritmo *apriori* que usa a técnica de associação, que também faz parte da parte descritiva, é outro recurso que será empregado. Se enquadrando no aprendizado não supervisionado.

Depois de analisados os comportamentos das variáveis na parte descritiva, será empregada a abordagem preditiva utilizando as técnicas de classificação descritas no capítulo de fundamentação teórica. As técnicas de classificação pertencem ao aprendizado supervisionado.

No processo de treinamento e teste, foi escolhida técnica *k fold cross validation*, com K = 10. Essa técnica particiona o conjunto dos dados, e faz o treinamento e o teste de acordo com o K, ou seja, se K = 3, e se têm um conjunto de dados com 1500 instâncias, a técnica divide o conjunto de dados em três partes com 500 instâncias.

Sendo que, a técnica fará o treinamento e o teste, três vezes, diferentemente para cada parte dividida do conjunto de dados, considerando duas das partes para treinamento e uma para teste.

Ao calcular a média do desempenho de cada um desses modelos, consegue-se obter uma estimativa melhor do desempenho do modelo a ser considerado.

Após a aplicação da classificação, o parâmetro da área abaixo da curva ROC será escolhido como fator da escolha de uma técnica sobre as outras.

Escolhido o classificador, são analisados os demais parâmetros que estão descritos na seção 3.3.2.

3.2.5 Conhecimento

O principal conhecimento que se quer extrair deste trabalho, são os possíveis padrões relacionados ao desempenho da nota final, com as variáveis selecionadas dos questionários socioeconômicos e de percepção da prova.

Os pontos chave deste trabalho estão de acordo com possíveis variáveis podem ser consideradas como fatores que podem determinar o desempenho do estudante na nota final da prova. As seguintes variáveis serão consideradas para as análises:

I. A renda familiar e do estudante;
II. A situação de trabalho;
III. A quantidade de horas de estudo e livros lidos durante o ano (não didáticos);
IV. Grau de dificuldade, objetividade e clareza das questões em relação se estão condizentes com as respostas na parte de percepção da prova específica;
V. Tempo gasto na tentativa de resolução da prova.

Apesar dessas questões estarem como centro deste trabalho, outras relações podem ser apresentadas no capítulo dos resultados.

A próxima seção segue com a descrição da ferramenta WEKA, e a configuração de seus classificadores, e a apresentação dos parâmetros que serão considerados para a avaliação dos modelos criados.

3.3 FERRAMENTA WEKA

O *Waikato Environment for Knowledge Analysis* (WEKA), trata-se de um software que utiliza a linguagem Java, disponível para download de forma gratuita, que incorpora muitas implementações que compreendem ao campo da aprendizagem de máquina, com uma variedade de algoritmos voltados para a mineração de dados. (WITTEN et al., 1999).

O WEKA, foi escolhido como o programa principal para a execução da mineração de dados, proposta deste trabalho, por ser um dos *softwares* pioneiros com relação à mineração de dados, amplamente adotado pela comunidade acadêmica, possuindo uma comunidade ativa, com atualizações constantes.

A ferramenta WEKA não é vista apenas como um único programa, mas como uma coleção de programas interdependentes unidos por uma interface de usuário. Possui técnicas de aprendizagem de máquina voltadas para o aprendizado supervisionado e não supervisionado. (GARNER, 2018).

A interface gráfica do WEKA, possui cinco botões na versão 3.8.1. (Versão essa, que será utilizada neste trabalho): *Explorer*, *Experimenter*, *Knowledge Flow*, *Simple CLI*, e *Workbench*.

Em Frank, Hall e Witten (2016), podemos encontrar descrições sobre os botões da interface citados da seguinte maneira:

O Explorer sendo um ambiente para um conjunto de dados serem explorados, contendo as seguintes abas:

Aba de pré-processamento: utilizada para a escolha e modificação dos dados.

Aba de classificação: utilizada para treinamento e teste dos dados. Podendo ser escolhida técnicas de classificação ou de regressão.

Aba de Grupos: utilizada para análises da formação de grupos de dados.

Aba de associação: utilizada para a criação de regras de associação.

Aba seleção de atributos: utilizada para seleção de atributos relevantes do conjunto de dados.

Aba de visualização: Exibe os dados em duas dimensões, possibilitando uma interação.

O *Experimenter* é um ambiente utilizado para comparar as técnicas de aprendizado. Com o *Knowledge Flow* é possível representar algoritmos de aprendizagem através de interfaces *drag-and-drop,* usual na aprendizagem

incremental. Já o *Simple CLI* é a interface de comandos de linhas. O *Workbench* é a unificação das três primeiras interfaces, ou seja, todas funcionalidades em uma única aplicação.

Todos os botões citados estão presentes na Figura 12, que apresenta a interface inicial do WEKA.

Figura 12: Interface inicial da WEKA.

Fonte: Própria.

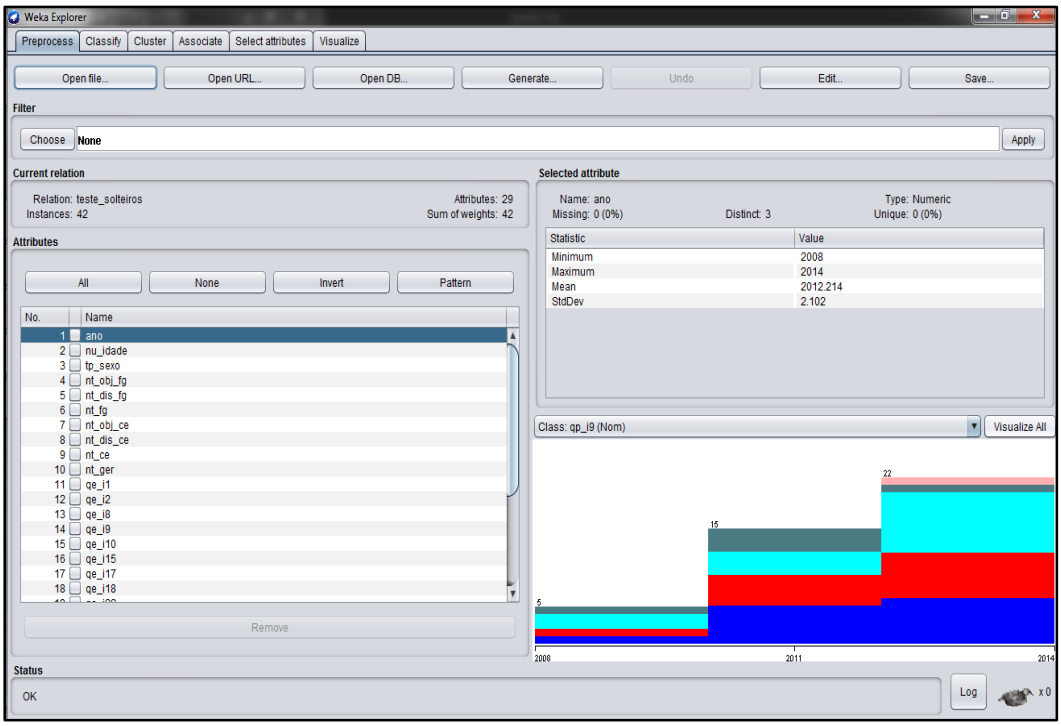
A WEKA possui alguns pacotes padrões instalados, que dão suporte as funcionalidades do programa, como as técnicas de classificação, por exemplo. Contudo novos pacotes podem ser instalados de acordo com a necessidade do usuário, dessa forma, não sobrecarregando o usuário com todas as ferramentas ativas, possibilitando contribuições da comunidade do WEKA, com o desenvolvimento de novos pacotes e atualizações. O *Package Management System* pode ser acessado em *tools* no menu do GUI *Chooser*, da Figura 12, em seguida pesquisa-se o pacote, e depois instala-o.

O formato padrão dos dados no WEKA, é o de arquivo de relacionamento de atributo, do inglês *Attribute Relationship File Format* (ARFF). Muitas bases de dados estão concentradas em planilhas, ou em arquivos de bancos de dados avançados. Contudo é preciso que haja a conversação para o formato padrão. As versões mais atuais do WEKA, também conseguem ler arquivos no formato *Comma-separated value* (CSV). Sendo assim, a base de dados deve estar em um dos formatos suportados, para que possa ser explorada no WEKA.

O presente trabalho se limitará em utilizar algumas funcionalidades presentes no ambiente *Explorer*, como as abas de pré-processamento, classificação e seleção de atributos. Todas as atividades feitas serão detalhadas nos capítulos seguintes.

As abas da interface Explorer estão presentes na Figura 13 na parte superior, contendo as abas citadas anteriormente.

Figura 13: Interface Explorer da WEKA.



Fonte: Própria.

3.3.1 Configurações dos algoritmos na WEKA

Nesta seção serão descritas as principais configurações dos algoritmos que fazem parte deste trabalho. Especificamente para os algoritmos da classificação, a configuração adotada, será a padrão para cada classificador.

Com relação ao algoritmo de aprendizagem não supervisionada, o *Apriori*, busca encontrar padrões frequentes na sua base de dados, ou seja, associações entre os objetos. Como já foi descrito no capítulo de fundamentação teórica, o *Apriori*, utiliza conceitos de probabilidade para fazer essas associações. Os conceitos principais já abordados são: suporte e confiança.

A métrica padrão a ser analisada será a de confiança. E as regras serão ordenas pelo *Lift*, onde o valor maior que 1, significa que a condição do item Y acontece se tiver o item X como antecessor, no formato {X} → {Y}, reforçando ainda mais a importância da regra.

Para este trabalho, o valor de suporte mínimo será de 0.2, o valor mínimo de confiança será de 0.75, e a quantidade de regras serão estipuladas durante a aplicação da técnica, contudo como a base de dados é pequena, no máximo trinta regras poderão ser analisadas.

Quanto aos classificadores:

Algumas das principais configurações do classificador J48, são:

Unpruned tree – Esse parâmetro, corresponde a “poda” da árvore. Utilizado para reduzir árvores grandes. Assumindo os valores *false* ou *true.* O valor padrão é *false*.

minNumObj – Parâmetro que contém o número mínimo de instâncias por nós. O padrão adotado é 2.

Os parâmetros principais para o classificador IBk são: o tipo de busca pelos vizinhos e o valor do K. A busca será realizada pela distância euclidiana entre dois pontos d = $\sqrt{(x2 - x1)^2 + (y2 - y1)^2}$. Já o valor de K utilizado será o valor padrão de 1.

Para as máquinas de vetores, duas configurações merecem destaque. O valor de C é o primeiro deles. Esse parâmetro que tem como padrão o valor 1.0, tem o objetivo de controlar o quanto de margem os vetores de suporte irão criar entre as classificações no hiperplano.

O valor de C mais apropriado depende da base de dados. E geralmente assume logaritmos de 10^{-3} até 10^{3}, ou seja, 0.01, 0.1, 1.0, 10.0, 100.0.

O outro parâmetro, é o tipo de *kernel* que será considerado. *Kernels* são métodos que utilizam funções lineares ou não lineares para calcular o produto de vetores em um espaço. A função de kernel padrão no WEKA é a polinomial.

Na Ferramenta WEKA, o classificador *NaiveBayes* possui poucas configurações. O parâmetro “*num decimal places*”, representa a quantidade de casas decimais que aparecem nas saídas das tabelas de frequência criadas pelo classificador. O número padrão é 2, ou seja, duas casas decimais.

Quanto ao MPL, uma das principais configurações que pode ser feita no classificador, é o número de camadas ocultas.

Na WEKA, a quantidade de camada ocultas correspondem a algumas letras. A letra 'a' = (atributos + classes) / 2, 'i' = atributos, 'o' = classes, e 't' = atributos + classes.

Em concatenação com as letras, pode-se limitar a quantidade de nós por camada, e consequentemente o número de camadas. Por exemplo: a,2,5,6. Onde se tem 4 camadas, com o valor de a, a segunda camada com dois nós, a terceira com cinco e a quarta camada com seis nós.

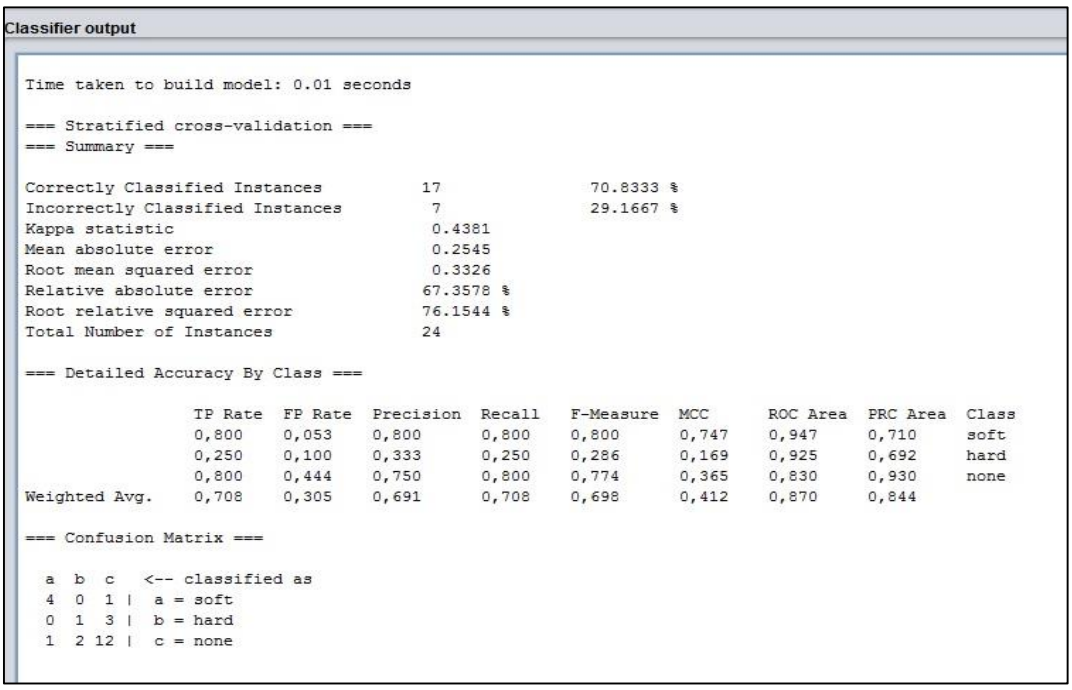
3.3.2 Parâmetros comuns aos classificadores

As métricas comuns aos algoritmos estão presentes na Figura 14. O exemplo utilizado são os das lentes de contato, já citada anteriormente.

O primeiro parâmetro é a da classificação correta das instâncias. O resultado pode ser visto em porcentual ou na quantidade de instâncias classificadas corretamente. É a acurácia do modelo.

Em seguida está disponível a quantidade de instâncias classificadas erroneamente, também em formato de percentual e em número de instâncias.

Figura 14: Parâmetros comuns a todos os algoritmos classificadores.



Classifier output

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.4381
Mean absolute error                      0.2545
Root mean squared error                  0.3326
Relative absolute error                 67.3578 %
Root relative squared error             76.1544 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,800    0,053    0,800      0,800    0,800      0,747    0,947     0,710     soft
               0,250    0,100    0,333      0,250    0,286      0,169    0,925     0,692     hard
               0,800    0,444    0,750      0,800    0,774      0,365    0,830     0,930     none
Weighted Avg.  0,708    0,305    0,691      0,708    0,698      0,412    0,870     0,844

=== Confusion Matrix ===

  a  b  c   <-- classified as
  4  0  1 |  a = soft
  0  1  3 |  b = hard
  1  2 12 |  c = none
```

Fonte: Própria.

A terceira métrica é a da estatística *kappa*, que traz um valor de concordância (precisão) ou não sobre o treinamento produzido pelo classificador. Um *kappa* de 1 indica uma concordância quase perfeita, enquanto um *kappa* menor que 0 indica uma não concordância (VIERA; GARRET, 2005).

Neste caso a métrica *kappa* indica a concordância diante da classificação e das classes classificadas certamente. O Quadro 12 apresenta um índice de utilização dessa concordância.

Quadro 12: Classificação de concordância Kappa.

Índice do Kappa	Concordância
< 0	Praticamente não há concordância.
0.01–0.20	Leve concordância.
0.21– 0.40	Concordância justa.
0.41–0.60	Concordância moderada.
0.61–0.80	Concordância substancial
0.81–0.99	Concordância quase perfeita.

Fonte: adaptado de Viera e Garret (2005).

Em seguida, os parâmetros de erro na seguinte sequência: erro absoluto médio, erro quadrático médio da raiz, erro relativo médio, e erro relativo quadrático da raiz.

Depois vários detalhes sobre a acurácia das classes. Contendo na sequência: TP Rate, taxa de positivos verdadeiros (instâncias classificadas corretamente como uma determinada classe, FP Rate (taxa de falsos positivos, casos falsamente classificados como uma determinada classe).

Precision, proporção de instâncias que são verdadeiramente de uma classe dividida pelas instâncias totais classificadas como essa classe. Recall, proporção de instâncias classificadas como uma determinada classe dividida pelo total real dessa classe (equivalente à taxa TP).

F-Measure, medida combinada de precisão e recall calculada como 2 * *Precision* * *Recall* / (*Precision* + *Recall*). MCC, *Matthews correlation coefficient*, ou coeficiente de correlação de Matthews, que retorna um coeficiente próximo de um ou próximo de menos 1. Quanto mais próximo de um positivo, há uma concordância quanto a classificação das classes. O cálculo leva em conta os falsos positivos e os positivos verdadeiros.

O *Receiver Operating Characteristic curve* (ROC) da área é um valor de 0 a 1 que é utilizado para criar uma curva em um gráfico que contém a taxa de positivos

verdadeiros (eixo y) contra a taxa de falsos positivos (eixo x) para cada classe. Quanto mais próximo de 1, pode-se dizer que a curva é excelente.

O *Precision-Recall Curve* (PRC) é uma métrica parecida com o ROC, mas com uma diferença, leva em consideração o desbalanceamento dos dados, em um conjunto de dados altamente desbalanceados.

Por fim, a matriz de confusão, métrica utilizada para visualizar a quantidade de instancias que foram classificadas corretamente ou não. Essa é uma métrica importante, porque mesmo que acurácia for razoavelmente considerável, uma determinada classe pode ter sido muito mal classificada, afetando o modelo como um todo.

A matriz de confusão, na Figura 14, pode ser interpretada da seguinte maneira, das cinco instâncias que pertenciam a saída *soft,* quatro foram classificadas corretamente, das quatro instancias pertencentes a saída *hard,* apenas uma foi classificada corretamente, dessa forma, apesar da acurácia ter sido de 71%, a classificação dessa saída da classe, pode comprometer o modelo. Para a classe *none,* com quatorze instâncias, apenas duas foram classificadas erradamente.

Apesar de existirem essa variedade de métricas, três serão levadas mais em conta, são: a área abaixo da curva ROC, a acurácia e a matriz de confusão.

Primeiramente, para a escolha do melhor classificador, será utilizada a área abaixo da curva ROC, em seguida as análises começaram da matriz de confusão, precisão do modelo, e concordância Kappa. Seguida, por análises do modelo por meio de visualização da predição, proporcionada por componentes da ferramenta WEKA. Caso o classificador escolhido, apresente uma estrutura do modelo criado, de fácil interpretação, essas informações também serão analisadas.

4 RESULTADOS E DISCUSSÕES

Neste capítulo serão apresentados os resultados para cada parte dos objetivos deste trabalho. A seção 4.1 descreve o perfil do estudante encontrado, e a distribuição dos dados para esse perfil, assim como para as questões de percepção da prova. Na seção 4.2 apresenta as análises feitas entre o desempenho dos estudantes do campus de Natal e o desempenho no cenário nacional, ao longo dos anos escolhidos. Em seguida para a subseção 4.2.1 são apresentadas o desempenho dos estudantes de Natal nas questões objetivas específicas, destacando as disciplinas que tiveram um melhor aproveitamento. Na seção 4.3 é apresentado a correlação das variáveis que foram escolhidas. Nas seções seguintes é aplicada a mineração de dados relacionada ao algoritmo *apriori*, e aos classificadores, respectivamente, seção 4.4, e 4.5. Nessas últimas seções algumas informações podem ter sidos suprimidas para dá mais ênfase na parte das análises e interpretações, ou seja, a extração do conhecimento do processo KDD.

4.1 PERFIL GERAL DO ESTUDANTE

Para construir o perfil do estudante, utilizou-se a base de dados numérica, aplicando-se a moda[1] estatística. Para a extração desse conhecimento dos dados, o uso da tarefa de aprendizagem descritiva foi empregue. Para essa finalidade, aplicou-se o gráfico de caixa, para melhor visualização dos resultados. A aplicação do gráfico de caixa foi por meio da linguagem R utilizando a ferramenta *R Studio*.

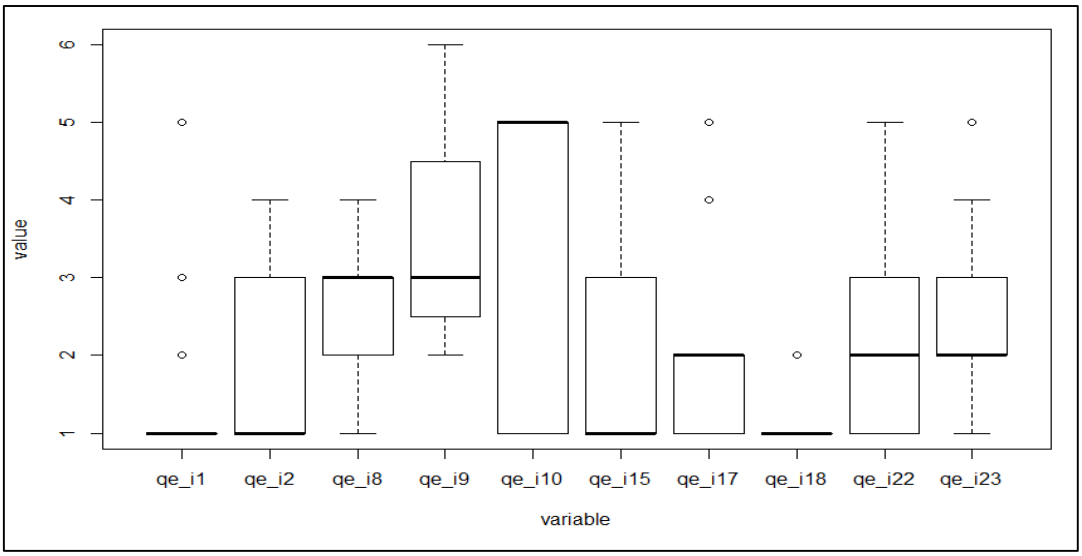
O gráfico de caixa apresenta um conceito específico chamado de quartis. De acordo com Fernandes e Pinto (2018), quartis são:

> Quartis são os valores que dividem um conjunto de dados em quatro partes iguais. Uma vez ordenado o conjunto de dados, o segundo quartil (Q2 - também conhecido como mediana) é o valor que fica a meio dos valores dos elementos do conjunto de dados, isto é, o valor que divide o conjunto de dados em duas partes iguais (metades). Depois o primeiro quartil (Q1) será o valor que fica a meio da primeira metade do conjunto de dados e o terceiro quartil (Q3) será, analogamente, o valor que fica a meio da segunda metade do conjunto de dados. (FERNANDES; PINTO, 2018).

[1] Moda é o valor (ou atributo) que ocorre com maior frequência em uma base de dados.

A aplicação da função chamada *boxplot,* gerou os gráficos para as análises. A Figura 15, reúne um conjunto de gráficos de caixa, que traz consigo um resumo do perfil socioeconômico dos estudantes da UERN, do curso ciência da computação do campus avançado de Natal.

Figura 15: *Boxplots* do perfil socioeconômico dos estudantes analisados.



Fonte: Própria.

A Figura 15, possui dez gráficos de caixa, que representam as questões selecionadas do questionário socioeconômico já apresentadas, ainda na Figura 15, pode-se visualizar que os *outliers*[2], representados em pequenos círculos, foram mantidos, e isso ocorreu por não se tratar de erros de *inputs*, e sim, por ser um conjunto de dados pequeno, e que tem alguns pontos fora da curva, mas que não prejudicam a análise, uma vez que é necessário para manter sua totalidade, e não haver perda de informação.

A análise do *boxplot* da primeira questão (qe_i1), que corresponde ao estado civil dos estudantes, deixa claro que não há amplitude interquartílica, nem valores de máximo e mínimo, e consequentemente uma variação de valores pertinentes, sendo sua mediana[3], ficando no valor 1, correspondendo que os estudantes quase na sua totalidade são solteiros, apresentados alguns *outiers* inexpressíveis.

[2] Em estatística, *outlier*, é um valor que apresenta um grande afastamento das demais da série, ou que é inconsistente.

[3] A mediana é o valor da variável que ocupa a posição central de um conjunto de n dados ordenados.

A segunda questão (qe_i2), retrata a questão da raça ou etnia, e possui uma distribuição positivamente assimétrica, uma vez que que a mediana está sobre o quartil Q1. Apesar de apresentar uma amplitude interquartílica abrangente, o gráfico aponta a mediana para o valor 1, que corresponde a raça ou etnia branco (a).

O terceiro *boxplot*, refere-se a questão sobre a renda familiar incluindo com a do estudante (qe_i8). Com esse gráfico, podemos observar que a amplitude interquartílica, abrange dois valores inteiros. O que faz referência ao valor 2, corresponde a uma renda de até 3 salários mínimos, já o valor 3, à uma renda de mais de 3 até 10 salários mínimos. A mediana, por sua vez, ficou no valor 3, sobre o quartil Q3, significando uma distribuição negativamente assimétrica.

O próximo *boxplot* (qe_i9), representa a distribuição da renda individual do estudante, o que inclui bolsas. Apresentando uma mediana com valor 3, que representa o cenário dos alunos que tem renda, mas que recebem ajuda da família ou de outras pessoas para financiar os seus gastos. Esse *bloxplot*, tem uma tendência próxima ao quartil Q1, então pode-se dizer que se trata de uma distribuição positivamente assimétrica.

A questão sobre a situação de trabalho do estudante (qe_i10), foi a que teve um *boxplot* com uma distribuição bastante ampla, se comparado aos demais, abrangente todas as respostas para a questão. A mediana do *boxplot* para essa questão foi o valor 5, que corresponde aos estudantes que trabalhavam 40 horas semanais ou mais. Como a mediana está sobre o quartil Q3, se trata de uma distribuição negativamente assimétrica.

O *bloxplot* para a questão a cerca se o estudante usou algum tipo de política de inclusão ou ação afirmativa para ingressar no curso (qe_i15), tem como resultado uma mediana de valor 1, estando sobre o quartil Q1, sendo dessa forma positivamente assimétrica. O valor da mediana, mostra que a maior parte dos estudantes, não usaram nenhuma política de cotas.

Quanto ao tipo de escola que o estudante cursou o ensino médio (qe_i17), o *boxplot* para esse caso, apresenta uma amplitude interquartílica de dois valores inteiros, mostrando que as respostas para essa pergunta se concentraram em duas opções, valor 1 para estudantes que cursaram o ensino médio em escolas públicas, e valor 2 para os que cursaram em escolas privadas (particulares), sendo que sua mediana ficou no valor 2, sobre o quartil Q3, implicando em uma distribuição negativamente assimétrica. O *boxplot* ainda teve dados que foram considerados

Outliers, e isso aconteceu por serem casos isolados que não estavam presentes na amplitude interquartílica.

Para a questão da modalidade de ensino médio que o estudante cursou (qe_i18), não apresentou faixa de amplitude interquartílica, sendo sua mediana situada no valor 1, correspondendo ao fato de que os estudantes quase na sua totalidade estudaram em um ensino médio tradicional. Mais uma vez, existe alguns dados foram considerados *outliers* por terem pouca representatividade dos dados, não tendo como ficar em uma faixa de valores pertinentes.

A penúltima questão (qe_i22), traz consigo um *boxplot* uma distribuição simétrica, com sua mediana no valor 2, correspondendo aos estudantes que leram um ou dois livros por ano (excluindo os indicados na bibliografia do curso).

O *boxplot* da última questão (qe_i23) é sobre a distribuição da quantidade de horas que os estudantes estudam por semana, excluindo as horas que passam na aula. Esse *boxplot* apresenta uma amplitude interquartílica que engloba dois valores inteiros, o valor 2, que corresponde de uma a três horas de estudos por semana, e o valor 3, de quatro a sete horas. A mediana ficou em uma a três horas de estudo por semana, sobre o quartil Q1, sendo a distribuição positivamente assimétrica.

Sendo assim, o perfil socioeconômico predominante do estudante de Ciência da Computação do Campus Avançado de Natal é o seguinte: solteiro (a), branco (a), renda da família correspondendo a mais de 3 até 10 salários mínimos, situação financeira pessoal descrita como tendo renda, mas recebe ajuda da família ou de outras pessoas para financiamento dos gastos, trabalha 40 horas semanais ou mais, não ingressou no curso por meio de políticas de ação afirmativa ou inclusão social (cotas), ensino médio em escola pública, modalidade de ensino médio tradicional, lendo um ou dois livros (não didático) por ano, e dedicando-se de uma a três horas de estudo por semana fora da sala de aula.

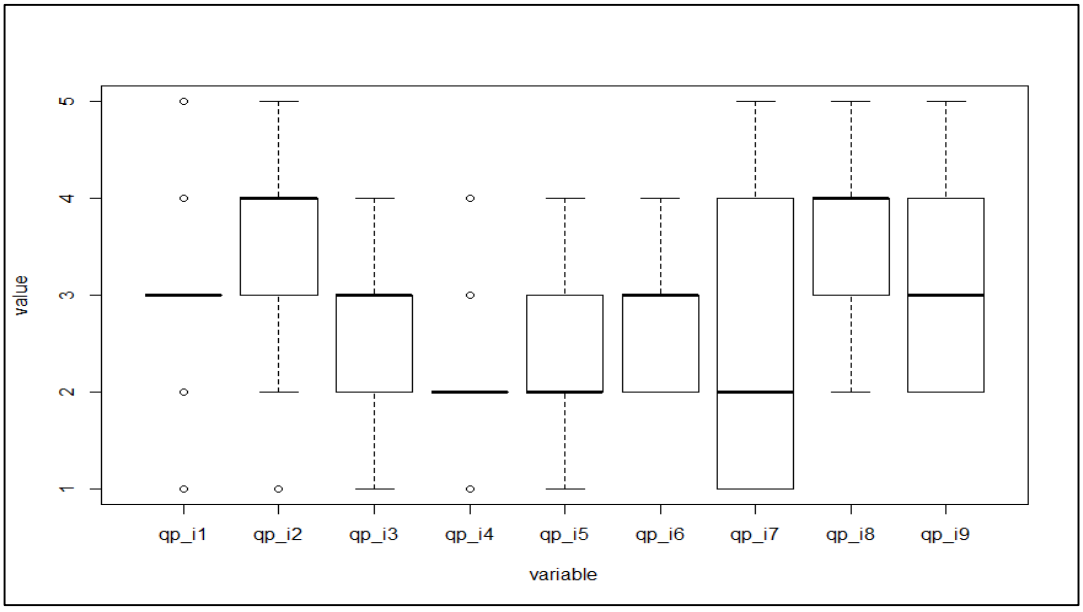
Tabela 3: Métricas retiradas dos *boxplots* do perfil socioeconômico.

	Média	Mediana	Moda	Máximo	Mínimo	Q1	Q3
qe_i1	1,313725	1	1	5	1	1	1
qe_i2	1,921569	1	1	4	1	1	3
qe_i8	2,745098	3	3	4	1	2	3
qe_i9	3,509804	3	3	6	2	2.5	4.5
qe_i10	3,627451	5	5	5	1	1	5
qe_i15	1,803922	1	1	5	1	1	3
qe_i17	1,705882	2	1	5	1	1	2
qe_i18	1,117647	1	1	2	1	1	1
qe_i22	2,18	2	2	5	1	1	3
qe_i23	2,666667	2	2	5	1	2	3

Fonte: Própria.

Depois de ter traçado o perfil do estudante com base nas respostas selecionadas do questionário socioeconômico, o mesmo foi feito com as questões de percepção da prova, com seus gráficos de caixa, estando representados na Figura 16.

Figura 16: *Boxplots* do questionário de percepção da prova.



Fonte: Própria.

O primeiro *boxplot* (qp_i1) da Figura 16, corresponde, a dificuldade da prova na parte da formação geral. Esse *boxplot*, não tem amplitude interquartílica, e possui alguns *Outliers*, que como já discutimos, para esse conjunto de dados, são pontos fora da curva, que se fossem ignorados, uma parte da informação seria perdida. A

mediana ficou no valor 3, que corresponde que os alunos na sua totalidade acharam médio o nível de dificuldade na parte formação geral.

Sobre a dificuldade do componente específico (qp_i2), traz um *boxplot* com uma amplitude interquartílica que varia entre uma dificuldade média e difícil, mas com sua mediana apontando para um nível de dificuldade considerada difícil, que incide sobre o quartil Q3, sendo uma distribuição negativamente assimétrica. O *boxplot* ainda possui um *outiliar*, não estando presente no valor mínimo e máximo, mas que não apresenta nenhuma anomalia, apenas um ponto fora da curva.

Considerando a extensão da prova (qp_i3), aponta um *boxplot* com sua mediana, no valor 3, que corresponde a uma prova com um tempo adequado, possuindo uma distribuição assimétrica negativa, por sua mediana estar sobre o Q3. Outro valor que surge no *boxplot*, é o 2, sinalizando que uma parte dos estudantes acharam a prova longa.

Quanto a objetividade e clareza dos enunciados na parte da formação geral (qp_i4), obteve-se um *boxplot* que aponta sua mediana para o valor 2, que representa a reposta, “Sim, na maioria”. Esse *boxplot* não possui amplitude interquartílica, nem valores de mínimo e máximo e *Outliers*.

Na parte específica sobre a objetividade e clareza dos enunciados (qp_i5), o *boxplot* tem uma amplitude interquartílica que varia entre os valores inteiros, 2 e 3. O valor 2 corresponde a reposta “Sim, a maioria”, e o valor 3 “apenas cerca da metade”. Contudo sua mediana, ficou sobre o quartil Q1, com a resposta “Sim, a maioria”, sendo uma distribuição positivamente assimétrica.

No que se diz a respeito das informações/instruções serem suficientes para a resolução das questões (qp_i6), apresenta um *boxplot* com uma amplitude interquartílica entre dois valores inteiros, o número 2 e o 3. O número 2 representa a resposta “sim, em todas elas”, já o 3, “sim, na maioria delas”. Com a mediana sobre o quartil Q3, a distribuição é negativamente assimétrica, sendo “sim, na maioria delas”, a resposta mais frequente.

A amplitude do *boxplot* sobre a questão de qual seria uma possível dificuldade ao responder a prova (qp_i7) é a mais abrangente da figura tal, contudo a resposta “forma diferente de abordagem do conteúdo”, foi a indicada pela mediana, como sendo a dificuldade mais frequente entre os estudantes, tendo uma tendência mais voltada para uma distribuição positivamente assimétrica.

Com relação ao domínio do conhecimento sobre questões objetivas da prova (qp_i8), o *boxplot* tem uma amplitude interquartílica voltada para dois valores inteiros, o valor 3, que corresponde a resposta “Estudou a maioria desses conteúdos, mas não os aprendeu”, e o valor 4, onde situa-se a mediana, com a resposta mais frequente “Estudou e aprendeu muitos desses conteúdos”. A mediana em questão, está sobre o quartil Q3, tratando-se de uma distribuição negativamente assimétrica.

A última pergunta do questionário de percepção da prova, é sobre o tempo gasto para finalizar a prova (qp_i9). Para essa pergunta, se tem a única distribuição simétrica de todos os *boxplots* analisados até esse momento, com a mediana no centro, sendo a resposta “entre duas e três horas”.

Com relação a moda das respostas dadas no questionário de percepção da prova, médio, foi o grau de dificuldade da prova na parte de formação geral considerado, quanto na parte do componente específico, difícil. Com relação ao tempo total da prova e sua extensão, a prova foi considerada adequada. “Sim, a maioria”, foi a resposta dada quanto a clareza e objetividade dos enunciados das questões na parte de formação geral. Da mesma forma para os da parte do componente específico. Quando perguntado se as informações/instruções para resolução da prova foram suficientes, a resposta foi que sim, na maioria delas. Forma diferente de abordagem do conteúdo, foi a resposta para a dificuldade em responder a prova. Considerando apenas as questões objetivas, os estudantes responderam que estudaram e aprenderam muitos dos conteúdos que foram apresentados na prova. E quanto ao tempo gasto para concluir a prova, entre duas e três horas foi a resposta.

Tabela 4: Métricas retiradas dos *boxplots* do questionário de percepção da prova.

Questão	Média	Mediana	Moda	Máximo	Mínimo	Q1	Q3
qp_i1	3	3	3	5	1	3	3
qp_i2	3,553191	4	4	5	1	3	4
qp_i3	2,531915	3	3	4	1	2	3
qp_i4	2,191489	2	2	4	1	2	2
qp_i5	2,382979	2	2	4	1	2	3
qp_i6	2,893617	3	3	4	2	2	3
qp_i7	2,361702	2	1	5	1	1	4
qp_i8	3,425532	4	4	5	2	3	4
qp_i9	3,06383	3	3	5	2	2	4

Fonte: Própria.

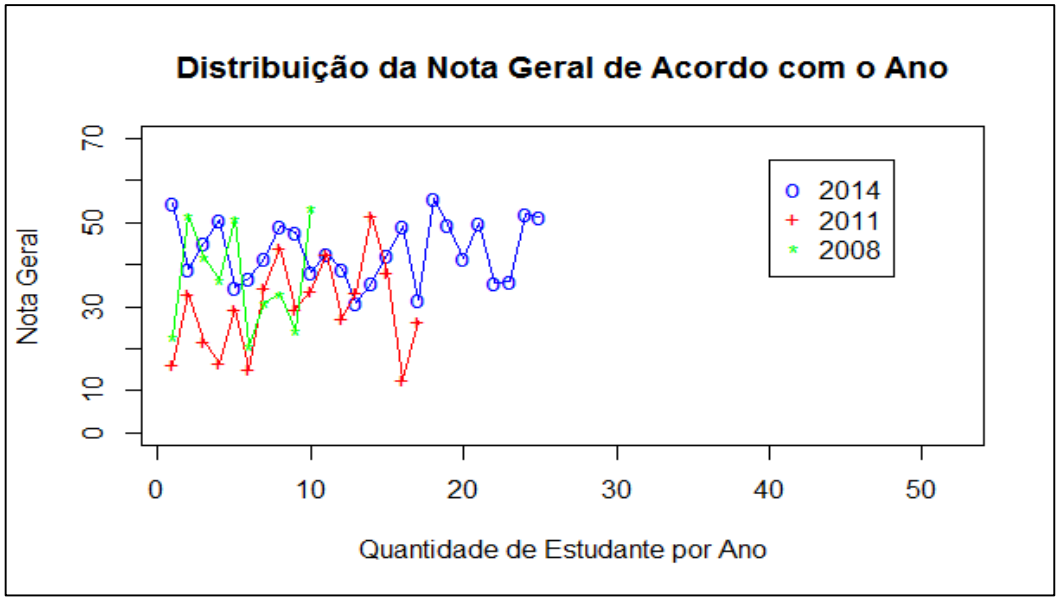
4.2. DESEMPENHO DO ESTUDANTE

O desempenho geral dos estudantes da UERN do curso de ciência da computação (CC), do campus de Natal na prova do ENADE, será apresentado nesta seção, com a motivação de mostrar o panorama geral dos resultados alcançados com a realização da prova nos anos de 2008, 2011, e 2014.

O conceito ENADE do curso em 2008, era de 3, caindo em 2011, para 2, e retornando para 3 em 2014.

No Gráfico 1, está visivelmente o desempenho da nota geral de acordo com os anos estudados. Percebe-se que o pior ano foi exatamente o de 2011, quando o conceito do curso passou para 2. Já os anos de 2008, e 2014, contiveram as melhores notas, sendo as notas de 2014, um pouco mais uniformes, e melhores que as 2008.

Gráfico 1: Desempenho dos estudantes na nota geral.



Fonte: Própria.

Para analisar o desempenho dos estudantes de acordo com o ano, não apenas a nota geral foi utilizada. As demais notas principais que compõem a prova, juntamente com as métricas estatísticas retiradas dos relatórios do ENADE do curso de ciência da computação, também serviram de base para as comparações de desempenho.

As Tabelas 5, 6, e 7, contém as métricas estatísticas para as principais notas das provas, dos anos de 2014, 2011, e 2008 respectivamente, do campus avançado de Natal, da UERN.

Tabela 5: Estatísticas básicas para as notas de 2014.

2014 - UERN	Média	Desvio	Mediana	Mínimo	Máximo
nt_obj_fg	60,5	18,63856	62,5	25	87,5
nt_dis_fg	59,08	26,12896	68	0	87
nt_fg	59,932	13,93299	63,1	33,9	85,1
nt_obj_ce	41,344	6,690483	42,9	28,6	52,4
nt_dis_ce	16,536	16,1976	16,7	0	50
nt_ce	37,644	7,323255	36,5	24,3	51,5
nt_ger	43,232	7,359728	42,3	31	55,7

Fonte: Própria.

Tabela 6: Estatísticas básicas para as notas de 2011.

2011 - UERN	Média	Desvio	Mediana	Mínimo	Máximo
nt_obj_fg	50,73529	21,86187	62,5	0	75
nt_dis_fg	49,41176	35,92015	55	0	100
nt_fg	50,20588	22,07874	52	0	77,5
nt_obj_ce	26,33529	9,250672	23,8	9,5	42,9
nt_dis_ce	4,7	11,28367	0	0	43,3
nt_ce	23,08824	8,762982	20,2	8,1	43
nt_ger	29,88235	10,90735	29,5	12,9	51,6

Fonte: Própria.

Tabela 7: Estatísticas básicas para as notas de 2008.

2008 - UERN	Média	Desvio	Mediana	Mínimo	Máximo
nt_obj_fg	58,75	22,08726	56,25	25	87,5
nt_dis_fg	56,5	17,72475	62,5	25	87,5
nt_fg	57,85	19,5079	59,25	25	80
nt_obj_ce	31,82	11,7264	31,45	11,1	48,1
nt_dis_ce	16,89	20,27256	11,7	0	62,2
nt_ce	29,58	12,22764	27,25	9,4	50,2
nt_ger	36,66	12,40772	34,8	20,8	53,5

Fonte: Própria.

As Tabelas 8 e 9 englobam as métricas para o resultado final, ou seja, a nota geral (NG), respectivamente do ano de 2014 e 2011, no âmbito nacional para os cursos de ciência da computação, no entanto em 2014, os cursos eram agrupados por bacharelado e licenciatura de ciência da computação, enquanto que em 2011, ainda incluíam os cursos de sistemas da informação e engenharia da computação.

A Tabela 10, possui de forma resumida, pela carência de informações mais completas, as métricas estatísticas encontradas no relatório da prova de 2008 de computação.

Os cursos da área da computação em 2008, envolvia apenas bacharelados, formando um grupo de computação e informática, dos cursos de ciência da computação, engenharia da computação, e sistemas de informação.

Tabela 8: Estatísticas básicas da Prova por Grande Região (NG) - 2014.

Estatísticas	Brasil	NO	NE	SE	SUL	CO
Média	45,3	44,1	46,6	44,5	46,9	44,1
Erro padrão da média	0,1	0,5	0,3	0,2	0,3	0,6
Desvio Padrão	14,1	13,7	14,6	13,9	13,4	14,4
Mínima	0	0	0	0	0	0
Mediana	44,8	44,1	46,5	43,5	46,2	43,7
Máxima	96,1	87,3	91,0	91,9	96,1	85,5

Fonte: INEP.

Tabela 9: Estatísticas básicas da Prova, por Grande Região (NG) – 2011.

Estatísticas	Brasil	NO	NE	SE	SUL	CO
Média	31,8	31,0	33,4	31,0	33,4	30,8
Erro padrão da média	0,1	0,3	0,2	0,1	0,2	0,2
Desvio Padrão	11,6	10,4	12,2	11,4	11,7	11,4
Mínima	0,0	0,0	0,0	0,0	0,0	0,0
Mediana	31,0	30,2	32,7	30,2	32,6	30,3
Máxima	86,7	74,9	75,2	86,7	82,8	74,7

Fonte: INEP.

Tabela 10: Estatísticas básicas da prova – 2008.

Brasil	Média	Desvio	Mediana	Mínimo	Máximo
nt_obj_fg	56,5	20,5	62,5	0,0	100,0
nt_dis_fg	38,8	25,3	45	0,0	100,0
nt_fg	49,4	17,4	51	0,0	99
nt_obj_ce	28,1	15,1	33,3	0,0	94,4
nt_dis_ce	13,5	18,2	0,0	0,0	100,0
nt_ce	29,1	13,7	27,7	0,0	86,8
nt_ger	34,8	-	-	-	-

Fonte: INEP.

Em 2014, a média da nota geral dos alunos de ciência da computação do campus de Natal foi de 43,2, bem próximo do resultado nacional de 45,3, mas abaixo do esperado para as demais regiões, como pode ser visto na Tabela 8.

Contudo, vale ressaltar que esse resultado é composto tanto pelos cursos de bacharelado e licenciatura, e que o intervalo modal dos concluintes de bacharelado era de (40; 50], estando os alunos da UERN campus Natal, dentro desse intervalo, portanto um resultado que não foge da realidade dos cursos de bacharelado em ciência da computação.

Enquanto que em 2011, a média geral foi de 29,8, um pouco abaixo da média nacional de 31,8, e das regiões, contudo para o bacharelado em ciência da computação, o intervalo modal era de (20;30], dessa forma, os estudantes da UERN, campus natal, ainda estavam dentro do intervalo aceitável.

Em 2008, a média nacional foi de 34,8, ao mesmo tempo que, a média do campus de Natal foi de 36,6, dessa forma superando as expectativas. O relatório de 2008, não traz consigo o intervalo modal para o curso de ciência da computação, contudo percebe-se que, apesar da média ser inferior ao ano de 2014, o resultado foi superior em todas as regiões 33,8 (NO), 36,0 (NE), 34,4 (SE), 36,2 (SUL), 33,3 (CO).

No que compreende as notas objetivas do componente geral (CG) e específico (CE), as Tabelas 11 e 12, apresentam as métricas estatísticas, para o ano de 2014, e as Tabelas 13 e 14 para o ano de 2011. Ressaltando que os valores para 2008 estão presentes na Tabela 10.

Tabela 11: Estatísticas básicas por Grande Região (CG) – 2014.

Estatísticas	Brasil	NO	NE	SE	SUL	CO
Média	63,1	56,3	62,0	64,0	65,9	62,2
Erro padrão da média	0,2	0,7	0,4	0,3	0,5	0,9
Desvio Padrão	21,0	20,8	21,4	20,8	20,0	22,0
Mínima	0,0	0,0	0,0	0,0	0,0	0,0
Mediana	62,5	62,5	62,5	62,5	62,5	62,5
Máxima	100,0	100,0	100,0	100,0	100,0	100,0

Fonte: INEP.

Tabela 12: Estatísticas básicas por Grande Região (CE) - 2014.

Estatísticas	Brasil	NO	NE	SE	SUL	CO
Média	42,5	37,9	43,6	41,6	45,2	40,8
Erro padrão da média	0,2	0,7	0,4	0,2	0,4	0,7
Desvio Padrão	15,8	14,0	16,4	15,5	15,9	15,9
Mínima	0,0	0,0	0,0	0,0	0,0	0,0
Mediana	42,9	38,1	42,9	42,9	42,9	38,1
Máxima	100,0	85,7	95,2	95,2	100,0	81,0

Fonte: INEP.

Tabela 13: Estatísticas básicas por Grande Região (CG) - 2011.

Estatísticas	Brasil	NO	NE	SE	SUL	CO
Média	50,6	52,5	52,7	49,4	51,4	50,7
Erro padrão da média	0,1	0,6	0,3	0,2	0,3	0,4
Desvio Padrão	19,4	19,5	19,5	19,4	18,6	20,0
Mínima	0,0	0,0	0,0	0,0	0,0	0,0
Mediana	50,0	50,0	50,0	50,0	50,0	50,0
Máxima	100,0	100,0	100,0	100,0	100,0	100,0

Fonte: INEP.

Tabela 14: Estatísticas básicas por Grande Região (CE) - 2011.

Estatísticas	Brasil	NO	NE	SE	SUL	CO
Média	30,8	27,4	33,7	29,3	34,2	30,4
Erro padrão da média	0,2	0,6	0,4	0,2	0,4	0,6
Desvio Padrão	15,4	13,9	16,6	14,8	15,8	15,4
Mínima	0,0	0,0	0,0	0,0	0,0	0,0
Mediana	28,6	23,8	33,3	28,6	33,3	28,6
Máxima	90,5	76,2	85,7	90,5	90,5	81,0

Fonte: INEP.

Com relação ao ano de 2014, a nota objetiva de formação geral ficou bem próxima do cenário nacional, mesmo estando adicionado as licenciaturas. A média do curso de CC do campus de Natal foi de 60,5, enquanto a média nacional foi de 63,1. No entanto, ficando acima da região Norte, que teve um desempenho de 56,3.

Em 2011 a média da nota objetiva de formação dos estudantes era de 50,7, estando praticamente igual ao resultado nacional que foi de 50,6, e acima da região Sudeste com média de 49,4, e igualmente com a região Centro Oeste.

Enquanto que em 2008, a média dos estudantes do campus de Natal era de 58,7, a do Brasil ficou em 56,5. No relatório de 2008, diferentemente de 2011 e 2014, não tinha como comparar com as regiões, por não conter no relatório do ENADE de 2008, especificamente para as questões objetivas de formação geral, as médias das regiões.

Para as questões objetivas específicas da prova, o relatório do ENADE, em todos os anos, trazia separadamente as estatísticas vinculadas ao curso em questão.

Sendo a média nacional para os cursos de bacharelado em 2014 de 42,5, um pouco acima da média dos estudantes do campus de Natal, que foi de 41,3. Em 2011 a média no Brasil era de 30,8, já no campus de Natal, de 26,3. No ano de 2008, a média nacional, e dos estudantes de Natal, eram respectivamente, 28,1 e 31,8.

Apesar de 2014, ano que contém a média geral mais alta, em ambas realidades retratadas, está evidente que as notas objetivas, tanto na parte de formação geral, mas principalmente na parte do componente específico, precisam de uma grande melhora em seu desempenho.

O mesmo ocorre entre as discursivas, principalmente, no componente específico, trazendo um impacto negativo sobre a nota final.

A média do componente específico discursivo em 2014, 2011, e 2008, no âmbito nacional foi de 24,8, 3,3 e 15,0 respectivamente. Ao mesmo tempo que no campus de Natal, as notas para esse quesito foram de, 16,5, 4,7, 16,8.

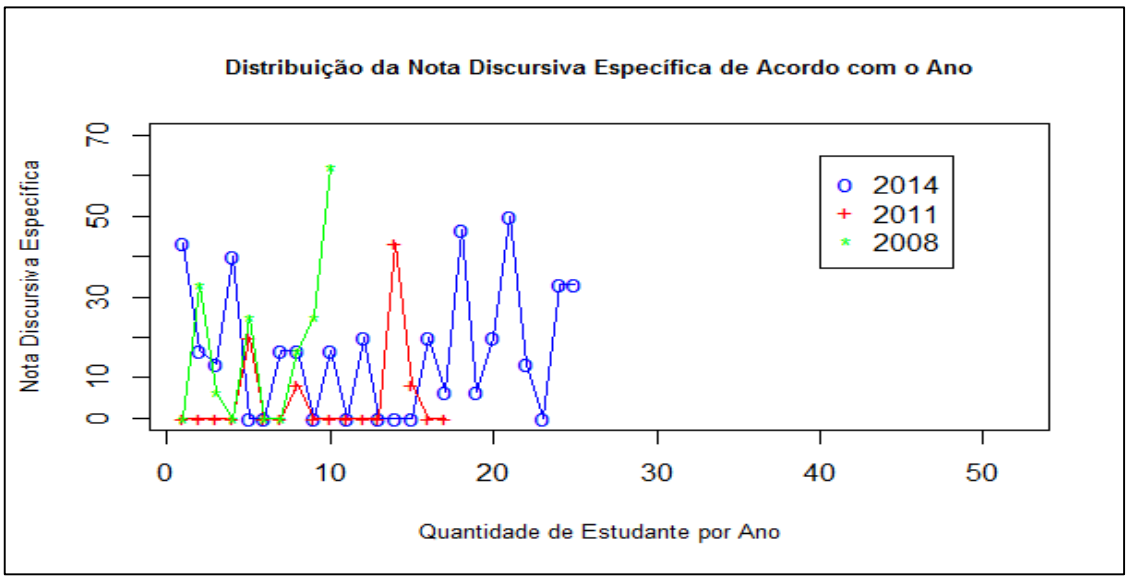
Tendo o intervalo modal para o curso de ciência da computação de [0; 10] em 2014 e 2011. No relatório de 2008, não tem o intervalo modal para o bacharelado em ciência da computação, mas para o grupo em questão (computação e informática), também foi de [0; 10].

Com relação a nota discursiva do componente geral, em 2014, 2011, e 2008, no âmbito nacional foi de 52,1, 48,5 e 38,8 respectivamente. Ao mesmo tempo que no campus Natal, as notas para esse quesito foram de, 59,0, 49,4, 56,5.

Sendo (60; 70] o intervalo modal para o bacharelado em ciência da computação em 2014, e (40;50] em 2011. Em 2008 foi de (51 a 60] para o grupo em questão (computação e informática).

A exibição do cenário das notas discursivas do componente específico dos estudantes de CC do campus de Natal, está presente no Gráfico 2.

Gráfico 2: Desempenho dos estudantes no componente discursivo específico.

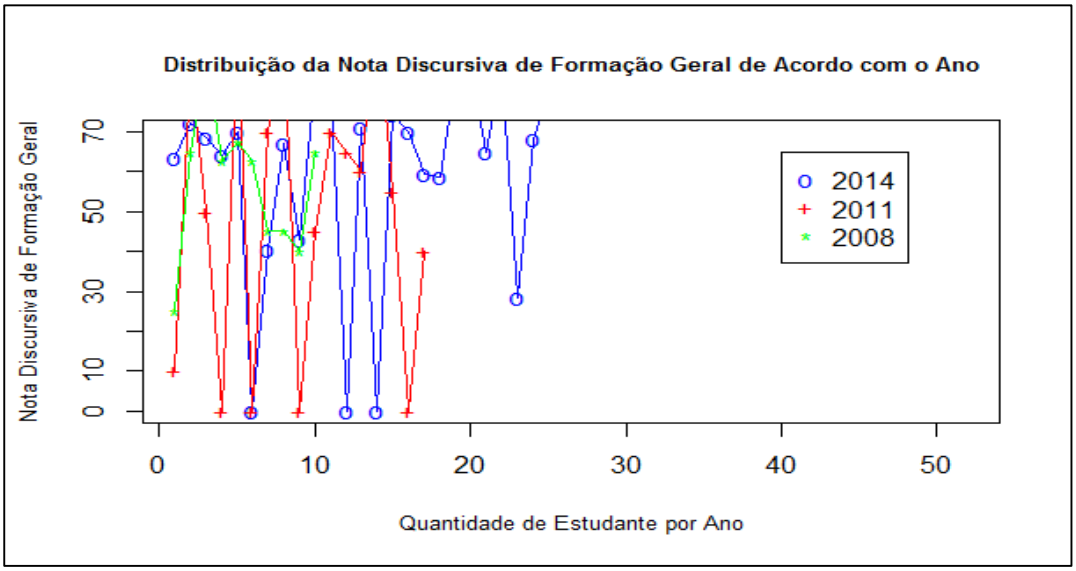


Fonte: Própria.

É no cenário do Gráfico 2, que estão as piores notas da prova, tendo vinte e cinco zeros dos três anos estudados, contra sete na discursiva do componente geral

presentes no Gráfico 3. Pondo em evidência que se a prova tivesse apenas questões específicas do curso, a realidade das notas poderia ser bem pior, uma vez, que a nota final do componente geral, tem impacto de 25% na nota geral final, e nela está contida os melhores resultados.

Gráfico 3: Desempenho dos estudantes no componente geral discursivo.



Fonte: Própria.

4.2.1 Questões específicas objetivas

Nesta seção será apresentado ao leitor, o desempenho dos alunos no componente objetivo específico da prova, de acordo com o ano, com o propósito de permitir a visualização de quais áreas de conhecimento da computação retratadas na prova estão tendo um melhor aproveitamento pelos estudantes, evidenciando o nível de complexidade da questão.

As informações da área de conhecimento, retratada para cada questão nas tabelas desta seção, foram retiradas do anexo IX do relatório do ENADE de 2014, e com base nas referências e comentários dos documentos “ENADE comentado”, dos anos de 2011 e 2008, da Pontifícia Universidade Católica do Rio Grande do Sul, uma vez que não constavam como anexos nos relatórios do ENADE dos anos supracitados.

A Tabela 15, traz consigo o nível de dificuldade e área de conhecimento das vinte e sete questões da prova objetiva específica do ano de 2014.

Tabela 15: Áreas do conhecimento do componente objetivo específico em 2014.

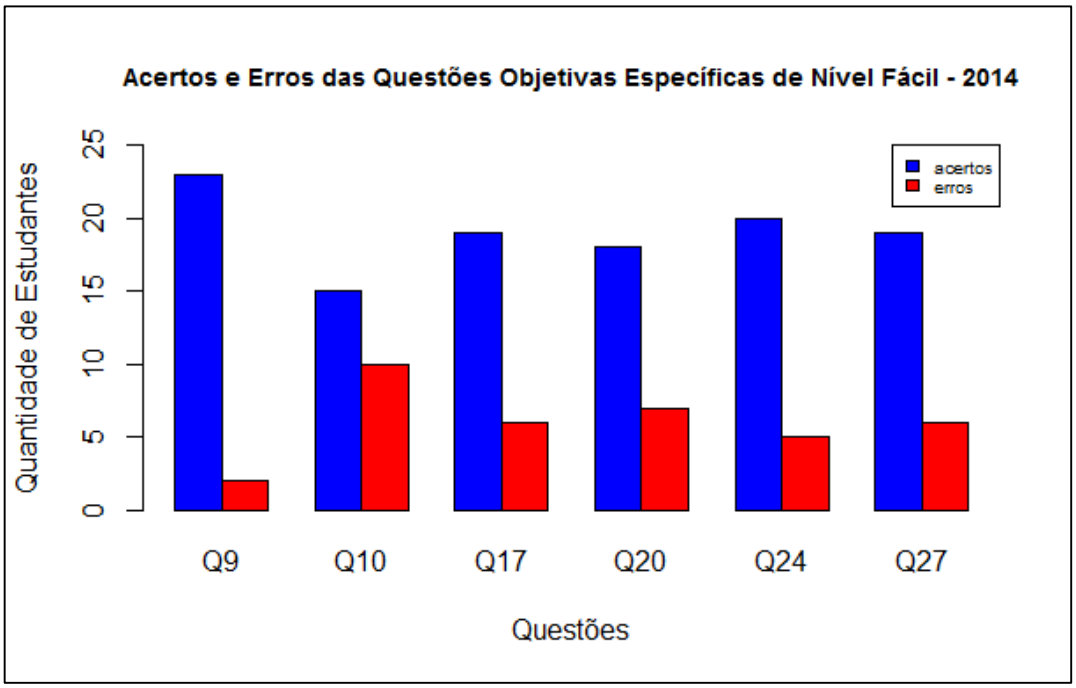
Questão	Nível de Dificuldade	Área de conhecimento
Q9	Fácil	Engenharia de software/interação humano-computador
Q10	Fácil	Engenharia de software/interação humano-computador
Q11	Difícil	Sistemas operacionais e arquitetura de computadores
Q12	Difícil	Teoria da computabilidade e complexidade
Q13	Difícil	Algoritmos e estruturas de dados
Q14	Difícil	Lógica e matemática discreta
Q15	Difícil	Linguagens formais, autômatos e compiladores
Q16	Difícil	Algoritmos e estrutura de dados
Q17	Fácil	Engenharia de software/interação humano-computador
Q18	Difícil	Inteligência artificial e computacional
Q19	Muito difícil	Computação gráfica e processamento de imagem
Q20	Fácil	Sistemas operacionais e arquitetura de computadores
Q21	Difícil	Sistemas operacionais e arquitetura de computadores
Q22	Difícil	Teoria dos Grafos
Q23	Difícil	Fundamentos e técnicas de programação
Q24	Fácil	Banco de dados
Q25	Médio	Linguagens formais, autômatos e compiladores
Q26	Difícil	Lógica e matemática discreta
Q27	Fácil	Lógica e matemática discreta
Q28	Difícil	Teoria da computabilidade e complexidade
Q29	Médio	Paradigmas de linguagens de programação; Lógica e matemática discreta
Q30	Difícil	Teoria dos grafos, teoria da computabilidade e complexidade
Q31	Muito difícil	Redes de computadores e sistemas distribuídos; probabilidade e estatística
Q32	Médio	Lógica e matemática discreta
Q33	Difícil	Sistemas operacionais e arquitetura de computadores
Q34	Muito difícil	Fundamentos e técnicas de programação; sistemas operacionais e arquitetura de computadores
Q35	Difícil	Sistemas digitais

Fonte: Própria

De acordo com as análises para o ano de 2014 os estudantes tiveram mais acertos nas questões de complexidade fácil. São seis questões que compartilham o esse mesmo nível de dificuldade. Os acertos e erros para esse grupo de questões estão presentes no Gráfico 4.

As áreas presentes no grupo de dificuldade fácil são: Engenharia de software/interação humano-computador, com três questões (Q9, Q10, Q17), Sistemas operacionais e arquitetura de computadores, com uma questão (Q20), assim como, Banco de dados (Q24), e Lógica e matemática discreta (Q27).

Gráfico 4: Acertos e erros das questões de nível fácil em 2014.



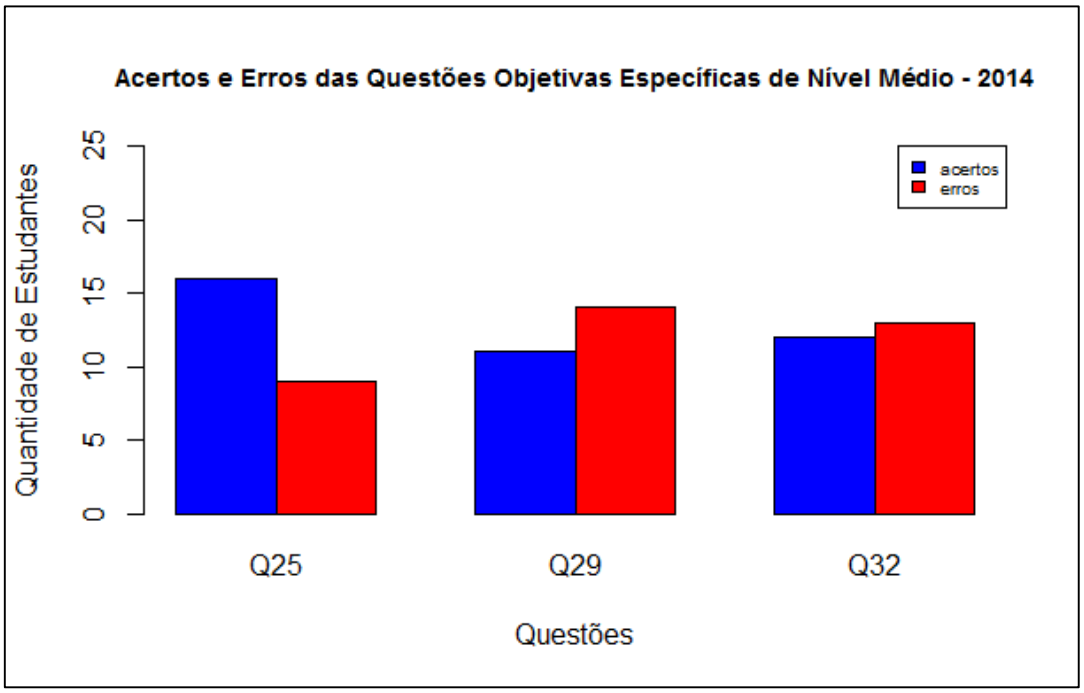
Fonte: Própria.

Com relação a quantidade de acertos para as questões do Gráfico 4, a questão Q9 foi a que teve o maior número de acertos, 92% dos estudantes acertaram, em seguida com 80% para a questão Q24, 76% para as questões Q17 e Q27, 72% para a questão Q20, e 60% para a questão Q10. Em média, a taxa de acerto para as questões com um grau de dificuldade fácil, foi de 76%.

O número de acertos vai diminuindo conforme o grau de dificuldade das questões aumentam. O próximo conjunto de questões que serão retratadas, são as que compreendem o nível de dificuldade médio, e estão presentes no Gráfico 5.

Nesse grupo de questões estão presentes as seguintes áreas de conhecimento: Linguagens formais, autômatos e compiladores, com uma questão (Q25), Paradigmas de linguagens de programação, também com uma questão (Q29), e Lógica e matemática discreta, presente na questão (Q23), assim como na questão (Q29). Vale ressaltar que na prova do ENADE, algumas questões podem conter mais de uma área de conhecimento.

Gráfico 5: Acertos e erros das questões de nível médio em 2014.



Fonte: Própria.

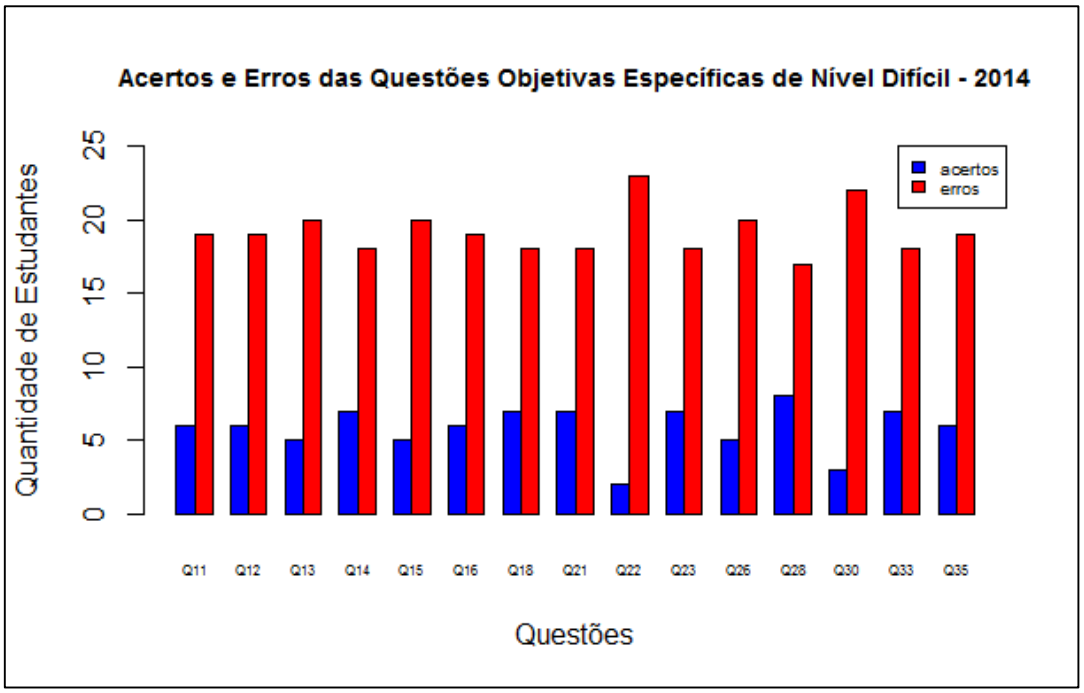
No que se diz respeito ao número de acertos para o grupo das questões de nível de dificuldade médio, 64% dos estudantes acertaram a questão Q25, 44% a questão Q29, e 48% a questão Q32. Em média, a taxa de acerto para as questões com um grau de dificuldade médio, foi de 52%.

A prova objetiva específica do ENADE costuma reunir mais de 50% das questões no nível de dificuldade difícil, e são nessas questões que os acertos dos estudantes não estão sendo satisfatórios, mas está de acordo com da realidade atual dos cursos de bacharelado em ciência da computação.

Sobre as questões específicas de nível difícil do ano de 2014, as áreas de conhecimento são: Sistemas operacionais e arquitetura de computadores, com três questões (Q11, Q21, Q33), Teoria da computabilidade e complexidade, com três questões (Q12, Q28, Q30), Algoritmos e estruturas de dados, com duas questões (Q13, Q16), Teoria dos Grafos, com duas questões (Q22, Q30), Inteligência artificial e computacional, com uma questão (Q18), assim como, Fundamentos e técnicas de programação (Q23), Lógica e matemática discreta(Q26), e Sistemas digitais (Q35).

As quinze questões de dificuldade difícil estão presentes no Gráfico 6.

Gráfico 6: Acertos e erros das questões de nível difícil em 2014.

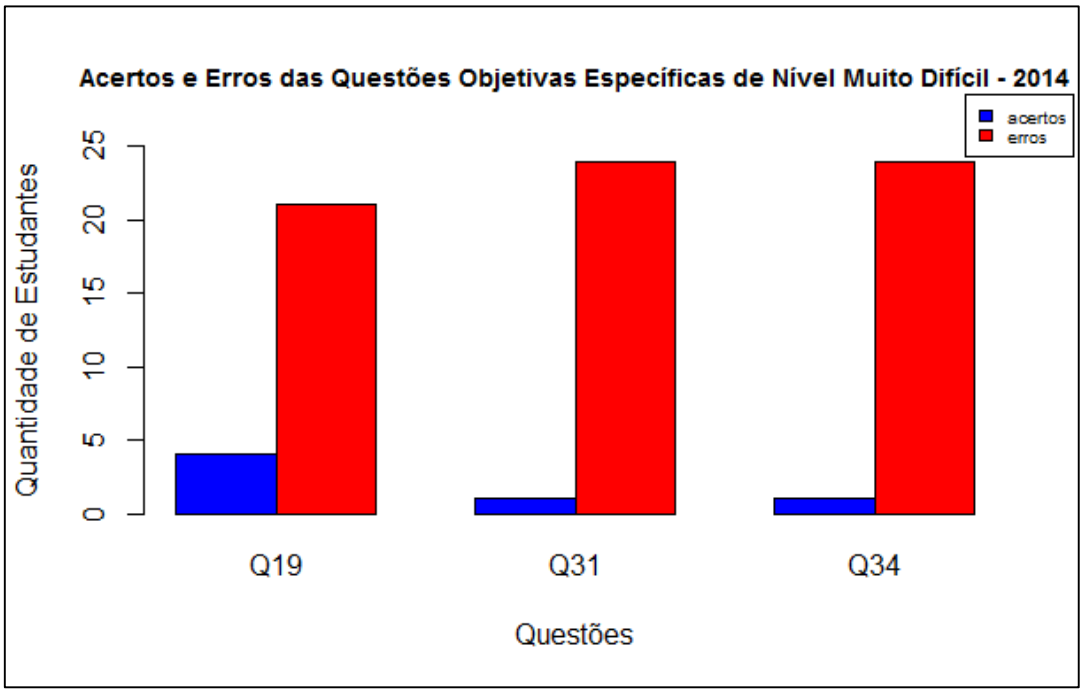


Fonte: Própria.

Os acertos sobre as questões específicas do Gráfico 6 estão abaixo de 35% para cada questão. A taxa maior de acerto foi referente a questão Q28, de Teoria da computabilidade e complexidade, com 32%. Em contrapartida a questão que teve o maior número de erros foi a questão Q22, de Teoria dos Grafos, com 92%. A taxa média de acertos para as questões de complexidade difícil foi de 23,2%.

A prova ainda traz questões com um grau de dificuldade ditas muito difíceis. No caso para a edição de 2014, compõem o quadro, três questões, com as seguintes áreas de conhecimento: Computação gráfica e processamento de imagem, com uma questão (Q19), assim como Redes de computadores e sistemas distribuídos e probabilidade e estatística (Q31), e Fundamentos e técnicas de programação; sistemas operacionais e arquitetura de computadores (Q34). As questões mencionadas, estão presentes no Gráfico 7.

Gráfico 7: Acertos e erros das questões de nível muito difícil em 2014.



Fonte: Própria.

A média de acertos para o grupo de questões do Gráfico 7, foi de apenas 8%, sendo o pior resultado para o grupo de questões de acordo com nível de dificuldade.

As análises seguem com a base de dados do ano de 2011, onde há uma mudança com relação ao grau de dificuldade das questões. O ponto mais acentuado é que não existe questões classificadas como fácies, aumentando o número de questões de dificuldade média, e difícil.

Em comparação com 2014, duas áreas de conhecimento não estão presentes nas questões objetivas específicas de 2011, são as seguintes: banco de dados e Computação gráfica e processamento de imagem.

A Tabela 16 contém o cenário do grau de dificuldades de cada questão, e sua área de conhecimento. Na edição de 2011, uma questão foi anulada.

Tabela 16: Áreas do conhecimento do componente objetivo específico em 2011.

Questão	Nível de Dificuldade	Área de conhecimento
Q9	Difícil	Lógica e matemática discreta
Q10	Médio	Probabilidade e estatística
Q11	Difícil	Linguagens formais, autômatos e compiladores
Q12	Médio	Linguagens formais, autômatos e compiladores
Q13	Anulada	Anulada
Q14	Difícil	Lógica e matemática discreta
Q15	Difícil	Redes de computadores e sistemas distribuídos
Q16	Difícil	Redes de computadores e sistemas distribuídos
Q17	Difícil	Sistemas Digitais
Q18	Médio	Sistemas operacionais e arquitetura de computadores
Q19	Médio	Engenharia de software/interação humano-computador
Q20	Difícil	Teoria dos Grafos
Q21	Médio	Algoritmos e estrutura de dados
Q22	Difícil	Sistemas Digitais; Sistemas operacionais e arquitetura de computadores
Q23	Difícil	Linguagens formais, autômatos e compiladores
Q24	Difícil	Algoritmos e estrutura de dados
Q25	Difícil	Paradigmas de linguagens de programação
Q26	Difícil	Probabilidade e estatística
Q27	Difícil	Fundamentos e técnicas de programação; Paradigmas de linguagens de programação
Q28	Difícil	Paradigmas de linguagens de programação; Algoritmos e estrutura de dados
Q29	Muito difícil	Sistemas operacionais e arquitetura de computadores
Q30	Difícil	Fundamentos e técnicas de programação; Algoritmos e estrutura de dados
Q36	Difícil	Teoria da computabilidade e complexidade
Q37	Difícil	Fundamentos e técnicas de programação
Q38	Muito difícil	Linguagens formais, autômatos e compiladores
Q39	Difícil	Algoritmos e estrutura de dados; Paradigmas de linguagens de programação
Q40	Muito difícil	Inteligência artificial e computacional

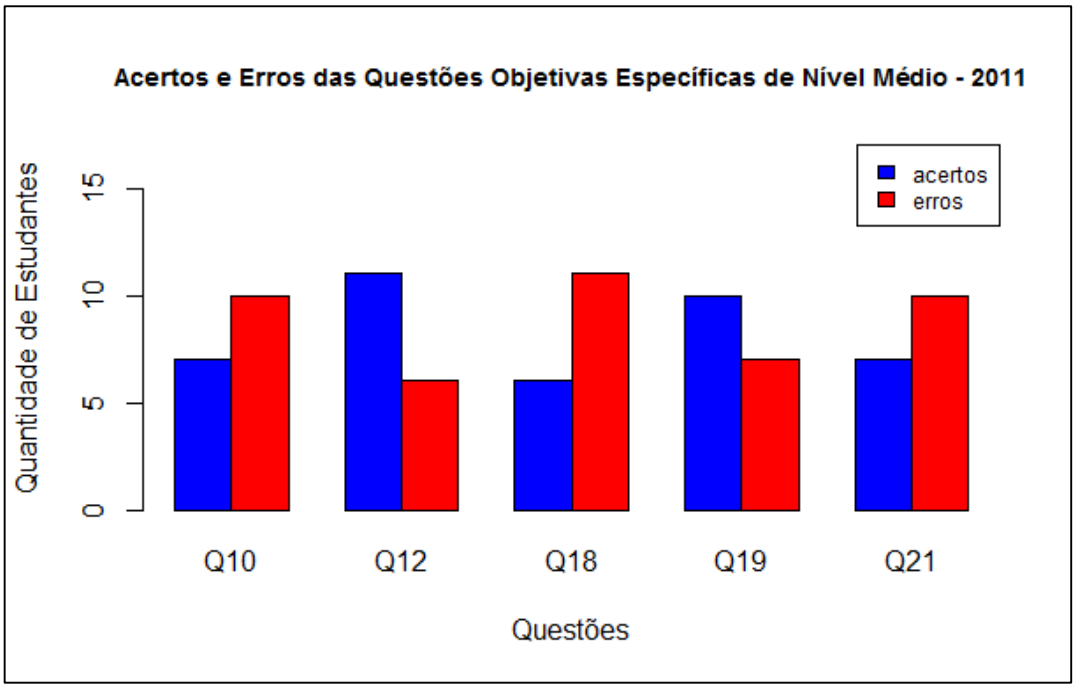
Fonte: Própria.

Diante das análises para as questões específicas, o ano de 2011, foi ano que teve pior desempenho dos estudantes. Sem contar com questões de grau de dificuldade fácil, os estudantes responderam questões de complexidade média, difícil, e muito difícil.

Cinco questões fazem parte do nível de dificuldade médio, que incluem as seguintes áreas de conhecimento, com uma questão cada: Probabilidade e estatística (Q10), Linguagens formais, autômatos e compiladores (Q12), Sistemas operacionais e arquitetura de computadores (Q18), Engenharia de software/interação humano-computador (Q19), e Algoritmos e estrutura de dados (Q21).

Os erros e acertos para as questões objetivas específicas de nível médio estão presentes no Gráfico 8.

Gráfico 8: Acertos e erros das questões de nível médio em 2011.



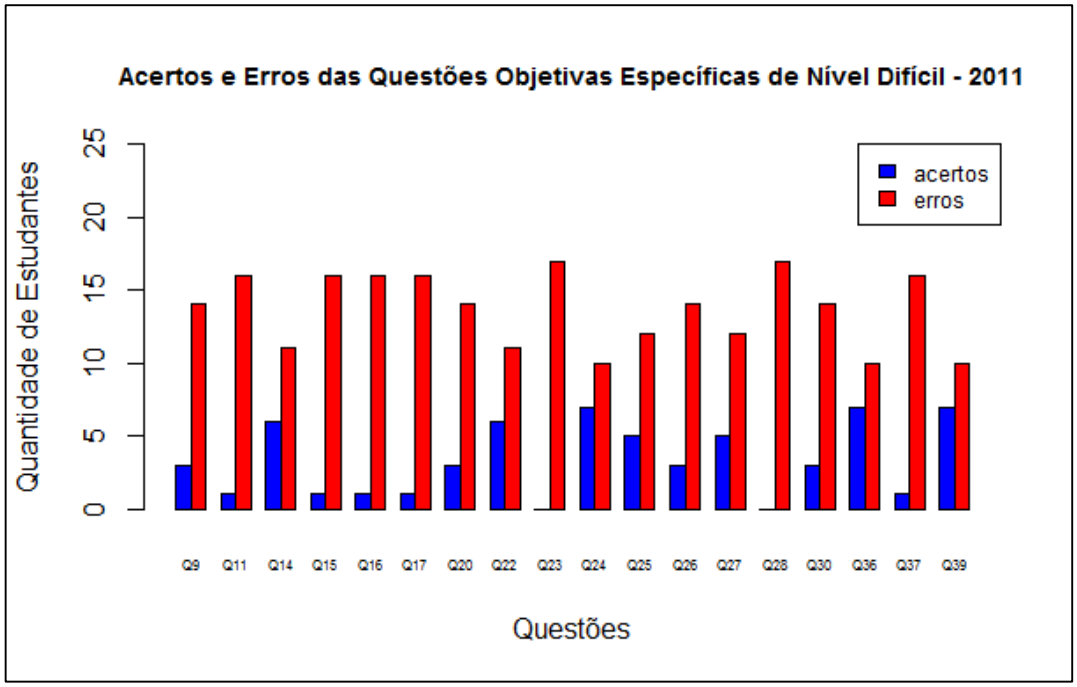
Fonte: Própria.

No que se refere aos acertos e erros expostos no Gráfico 8, das cinco questões, em duas os estudantes conseguiram acertar mais do que errar. A questão com mais acertos foi a questão Q12, com 64,7%, seguida pelas questões Q19, com 58,8%, Q10 e Q21, ambas com 41,1%, e a Q18 com 35,3%. Sendo a média de acertos para esse grupo de questões de 48,2%, um pouco abaixo do resultado de 52% em 2014.

Sobre as questões específicas de nível difícil do ano de 2011, as áreas de conhecimento são: Algoritmos e estrutura de dados, com quatro questões (Q24, Q28, Q30, Q39), assim como, Paradigmas de linguagens de programação (Q25, Q27, Q28, Q39). Fundamentos e técnicas de programação, com três questões (Q27, Q30, Q37). Lógica e matemática discreta, com duas questões (Q9, Q14), da mesma maneira que, Linguagens formais, autômatos e compiladores (Q11, Q23), Redes de computadores e sistemas distribuídos (Q15, Q16), e Sistemas Digitais (Q17, Q22). Teoria dos Grafos, com uma questão (Q20), do mesmo modo que, Sistemas operacionais e arquitetura de computadores (Q22), Probabilidade e estatística (Q26), e Teoria da computabilidade e complexidade (Q36).

Ao todo, na edição de 2011, dezoito questões objetivas específicas foram consideradas de nível difícil. Os acertos e erros para essas questões estão no Gráfico 9.

Gráfico 9: Acertos e erros das questões de nível difícil em 2011.



Fonte: Própria.

O pior resultado encontrado no Gráfico 9, foram das questões Q23 e Q28, referente às Linguagens formais, autômatos e compiladores, e Paradigmas de linguagens de programação e Algoritmos e estrutura de dados, respectivamente, com nenhum acerto.

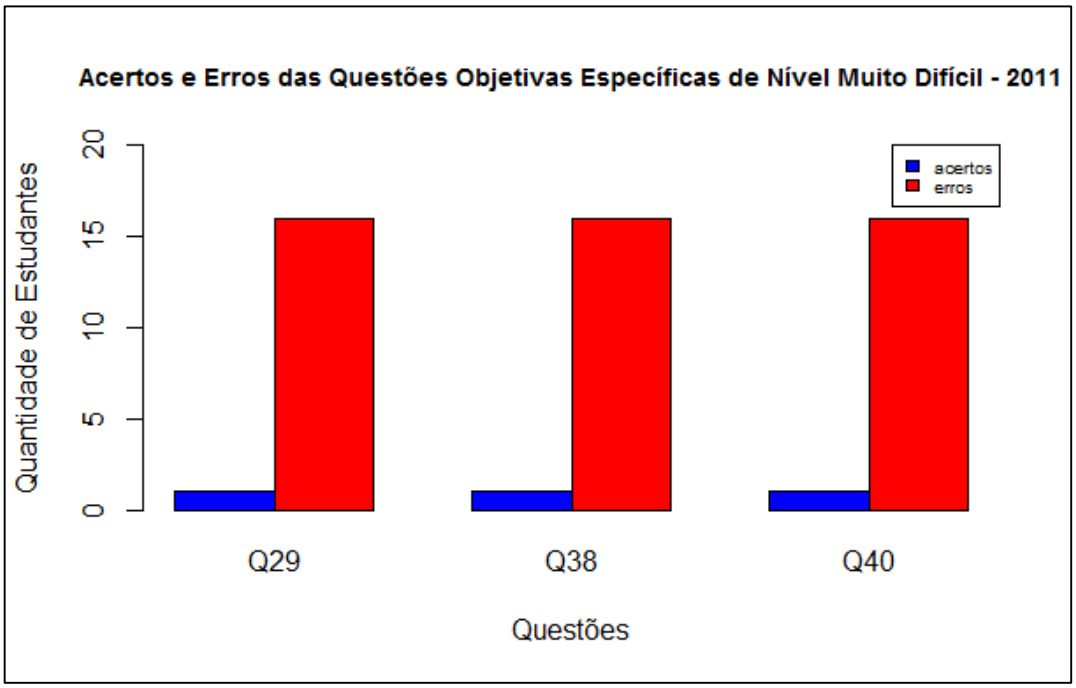
Enquanto que os maiores acertos estão nas questões Q24, Q36, Q39, com 41,2%. A taxa média de acertos para as questões de complexidade difícil foi de 19,6%, um pouco abaixo do ano de 2014, com 23,2%.

Em relação as questões de nível muito difícil, as áreas de conhecimento são: Sistemas operacionais e arquitetura de computadores, com uma questão (Q29), da mesma forma que, Linguagens formais, autômatos e compiladores (Q38), e Inteligência artificial e computacional (Q40).

A taxa média de acertos para as questões de nível muito difícil foi de 5,8%, um pouco abaixo do ano de 2014, que foi de 8%.

Os acertos e erros do nível de dificuldade muito difícil, podem ser visualizados no Gráfico 10.

Gráfico 10: Acertos e erros das questões de nível muito difícil em 2011.



Fonte: Própria.

Dando continuidade com as análises, o próximo conjunto de dados são referentes ao ano de 2008. Nesse ano estão presentes as áreas de conhecimento em banco de dados, e Computação gráfica e processamento de imagem, que não tinham sido inseridas no ano de 2011.

No ano de 2008, não foram identificadas perguntas objetivas específicas das seguintes áreas de conhecimentos: Inteligência artificial e computacional, Paradigmas de linguagens de programação, Teoria dos Grafos, e Teoria da computabilidade e complexidade. Estas áreas de conhecimento estavam nas objetivas específicas de 2011, e 2014. A Tabela 17 contém o cenário do grau de dificuldades de cada questão, e sua área de conhecimento.

Tabela 17: Áreas do conhecimento do componente objetivo específico em 2008.

Questão	Nível de Dificuldade	Área de conhecimento
Q11	Médio	Sistemas operacionais e arquitetura de computadores
Q12	Difícil	Engenharia de software/interação humano-computador
Q13	Médio	Lógica e matemática discreta; Métodos Formais
Q14	Muito difícil	Algoritmos e estrutura de dados; Fundamentos e técnicas de programação
Q15	Médio	Redes de computadores e sistemas distribuídos
Q16	Difícil	Engenharia de software/interação humano-computador
Q17	Difícil	Lógica e matemática discreta
Q18	Difícil	Fundamentos e técnicas de programação; Algoritmos e estrutura de dados
Q19	Médio	Sistemas operacionais e arquitetura de computadores
Q21	Difícil	Algoritmos e estrutura de dados
Q22	Médio	Linguagens formais, autômatos e compiladores
Q23	Muito difícil	Banco de Dados
Q24	Muito difícil	Sistemas Digitais
Q25	Difícil	Computação gráfica e processamento de imagem
Q26	Difícil	Computação gráfica e processamento de imagem
Q27	Difícil	Redes de computadores e sistemas distribuídos
Q28	Difícil	Algoritmos e estrutura de dados
Q29	Difícil	Linguagens formais, autômatos e compiladores
Q30	Difícil	Redes de computadores e sistemas distribuídos
Q31	Difícil	Algoritmos e estrutura de dados
Q32	Difícil	Probabilidade e Estatística
Q33	Difícil	Linguagens formais, autômatos e compiladores
Q34	Difícil	Redes de computadores e sistemas distribuídos
Q35	Difícil	Redes de computadores e sistemas distribuídos
Q36	Difícil	Redes de computadores e sistemas distribuídos
Q37	Difícil	Probabilidade e Estatística
Q38	Difícil	Sistemas Digitais

Fonte: Própria.

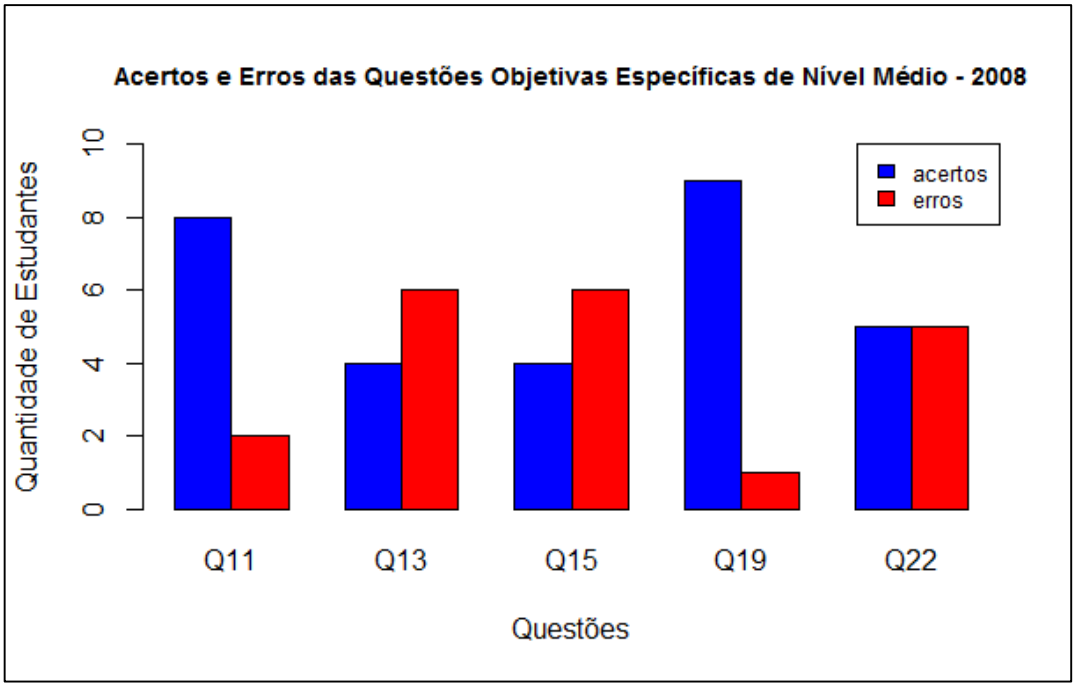
Em 2008, não teve questões de nível de dificuldade fácil, assim como em 2011, a prova foi constituída de cinco questões de nível de dificuldade, dezenove de nível difícil (uma questão foi anulada em 2011), e três de nível muito difícil.

As áreas de conhecimento que nortearam as questões de nível de dificuldade médio, foram: Sistemas operacionais e arquitetura de computadores com duas questões (Q11, Q19), as demais com uma questão, Lógica e matemática discreta e Métodos Formais (Q13), Redes de computadores e sistemas distribuídos (Q15), e Linguagens formais, autômatos e compiladores (Q22).

A taxa média de acertos para as questões de nível de dificuldade médio foi de 60%, ficando acima dos anos de 2011 e 2014, respectivamente com 48,2%, e 52%. Isso aconteceu devido aos acertos das questões Q19 e Q11, ambas da área de conhecimento em Sistemas operacionais e arquitetura de computadores, com 90% e 80% de acertos, respectivamente.

Os acertos e erros das questões supracitadas estão presentes no Gráfico 11.

Gráfico 11: Acertos e erros das questões de nível médio em 2008.



Fonte: Própria.

Com relação às questões de grau de dificuldade difícil, observar-se que no Gráfico 12, existem três questões que todos os alunos erraram. São elas: a questão Q21, Q30, e Q37, com as respectivas áreas de conhecimento, Algoritmos e estrutura de dados, Redes de computadores e sistemas distribuídos, e Probabilidade e Estatística.

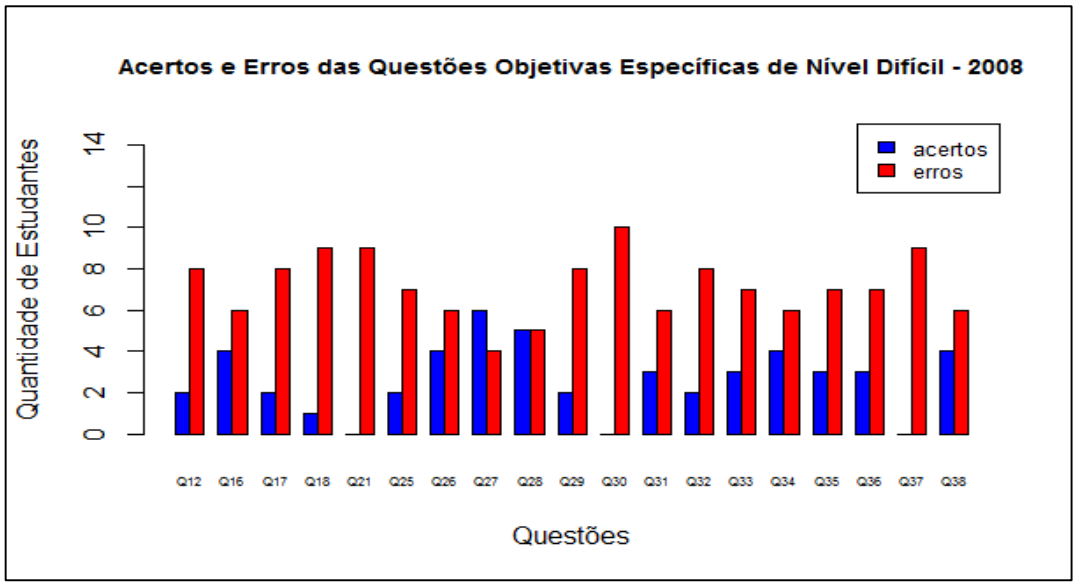
Em compensação, com 60% de acertos, a questão Q27, da área de em conhecimento em Redes de computadores e sistemas distribuídos, seguida por 50% da área de conhecimento em Algoritmos e estrutura de dados, obtiveram os melhores resultados por parte dos estudantes.

Sendo 2008, o ano que teve o melhor aproveitamento por parte dos estudantes nas questões objetivas específicas de nível de dificuldade difícil, com taxa média de acertos de 26,3, ficando à frente dos anos de 2011 e 2014, com respectivamente, 19,6%, e 23,2%.

Em relação as questões de nível muito difícil, as áreas de conhecimento são: Algoritmos e estrutura de dados; Fundamentos e técnicas de programação, com uma questão (Q14), da mesma forma que, Banco de Dados (Q23), e Sistemas Digitais (Q24).

A taxa média de acertos para as questões de nível muito difícil foi de 23,3%, muito acima dos anos de 2011 e 2014 que tiveram respectivamente, 5,8%, e 8% de acertos. Os acertos e erros do nível de dificuldade muito difícil, podem ser visualizados no Gráfico 13.

Gráfico 12: Acertos e erros das questões de nível difícil em 2014.



Fonte: Própria.

Gráfico 13: Acertos e erros das questões de nível muito difícil em 2008.

Fonte: Própria.

A Tabela 18, apresenta o resumo dessa seção, com o desempenho dos estudantes para cada área de conhecimento do curso.

Tabela 18: Desempenho dos estudantes nas áreas de conhecimento.

Ciência da computação	Ano			Nível de Dificuldade				Média de acertos	Aproveitamento
Área de conhecimento	2008	2011	2014	Fácil	Médio	Difícil	Muito difícil	%	
Algoritmos e estruturas de dados			x			x		20%	26,1 % 22,3%*
			x			x		24%	
		x			x			41,2 %	
		x				x		41,2%	
		x				x		0	
		x				x		17,6%	
		x				x		41,2%	
	x						x	20%	
	x					x		10%	
	x					x		0	
	x					x		30%	
Banco de dados			x	x				80%	50%
	x						x	20%	
Computação gráfica e processamento de imagem			x				x	16%	25,3%
	x					x		20%	
	x					x		40%	
Engenharia de software/interação humano-computador			x	x				92%	57,8%
			x	x				60%	
			x	x				76%	
		x			x			58,8%	
	x					x		20%	
	x					x		40%	
Fundamentos e técnicas de programação			x			x		28%	16,9% 12,4%*
			x				x	4%	
		x				x		29,4%	
		x				x		17,6%	
		x				x		5,9%	
	x						x	20%	
	x					x		10%	
Inteligência artificial e computacional			x			x		28%	16,9%
		x					x	5,9%	
Linguagens formais, autômatos e compiladores			x			x		20%	31,8% 28,9%*
			x		x			64%	
		x				x		5,9%	
		x			x			64,7%	
		x				x		0	
		x					x	5,9%	
	x				x			50%	
	x					x		20%	
	x					x		30%	
Lógica e matemática discreta			x			x		28%	35% 36,5%*
			x			x		20%	
			x	x				76%	
			x		x			44%	
			x		x			48%	
		x				x		17,6%	
		x				x		35,3%	
	x				x			40%	
	x					x		20%	
Paradigmas de linguagens de programação			x		x			44%	29,4% 28,8*
		x				x		29,4%	
		x				x		29,4%	
		x				x		0	
		x				x		41,2%	

Probabilidade e estatística			x			x	4%	19,7% 16,6%*
		x			x		41,2%	
		x				x	17,6%	
	x					x	20%	
	x					x	0	
Redes de computadores e sistemas distribuídos			x			x	4%	26,5% 24%*
		x				x	5,9%	
		x				x	5,9%	
	x				x		40%	
	x					x	60%	
	x					x	0	
	x					x	40%	
	x					x	30%	
	x					x	30%	
Sistemas digitais			x			x	24%	25% 27%*
		x				x	5,9%	
		x				x	35,3%	
	x					x	30%	
	x					x	40%	
Sistemas operacionais e arquitetura de computadores			x			x	24%	45,4% 40,2%*
			x	x			72%	
			x			x	28%	
			x			x	28%	
			x			x	4%	
		x			x		35,3%	
		x				x	35,3%	
		x				x	5,9%	
	x				x		80%	
	x				x		90%	
Teoria da computabilidade e complexidade			x			x	24%	32.4% 23%*
			x			x	32%	
			x			x	12%	
		x				x	5,9%	
		x				x	41,2	
Teoria dos grafos			x			x	8%	12,8% 12,5%*
			x			x	12%	
		x				x	17,6%	

Fonte: Própria. *Quando a questão é interdisciplinar.

4.3 CORRELAÇÃO ENTRE AS VARIÁVEIS

A correlação foi aplicada utilizando a função *cor* da linguagem R, presente na ferramenta *R Studio*. A função *cor* cria uma matriz de correlações entre as variáveis presentes, que foi armazenada em um arquivo com extensão csv. Para a criação dos gráficos, utilizou-se a função *corrplot*, presente no pacote de mesmo nome.

A correlação somente pode concretizar-se quando as células da planilha que continham um ponto, que era associado aos estudantes que não responderam a alguma pergunta dos questionários, fossem retirados, e deixadas sem valores.

Para lidar com a falta de valores, foi utilizado o método *complete.obs*, que trata esse problema por meio da eliminação *casewise,* na qual, as linhas que contém algum valor “N/A”, são eliminadas, permanecendo as demais para a correlação. O método adotado foi o de *Pearson*, padrão da função cor.

Os seguintes comandos foram utilizados:

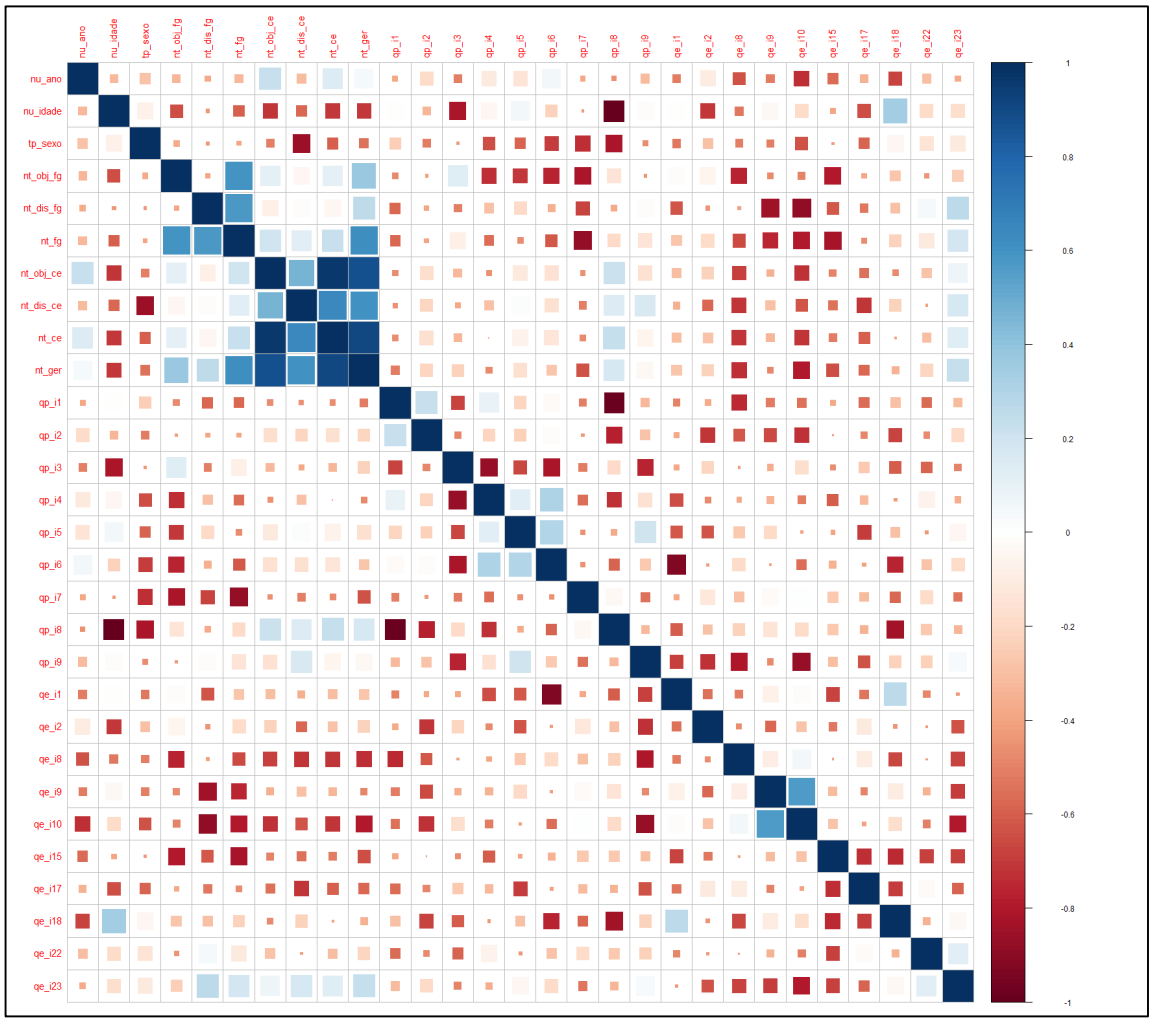
1. correlacao = cor(data, use = "complete.obs")
2. corrplot(correlacao, method = "square")
3. corrplot(correlacao, method = "number", number.cex=0.5)
4. write.csv(correlacao,'correlationmatrix.csv')

O primeiro comando cria a matriz de correlações, eliminado as linhas que contém pelo menos uma célula sem valor, salvando a matriz na variável *correlacao*. O segundo e o terceiro comando, criam os gráficos. O método *square*, cria um gráfico com quadrados, indicando pelo tamanho do quadrado, as correlações. Quadrados maiores têm uma forte correlação, enquanto pequenos ou quase imperceptíveis, uma correlação fraca. O método *number* traz o valor númerico das correlações de acordo com as cores azul e vermelha. Azul para as variáveis fortemente correlacionadas diretamente proporcional, com valor positivo próximo de 1, e vermelho para variáveis fortemente correlacionadas indiretamente proporcional, com valor negativo próximo de -1.

Em seguida, para analisar os valores da matriz de maneira mais detalhada, os valores das correlações foram armazenados em um arquivo csv, e depois transformado para xlsx.

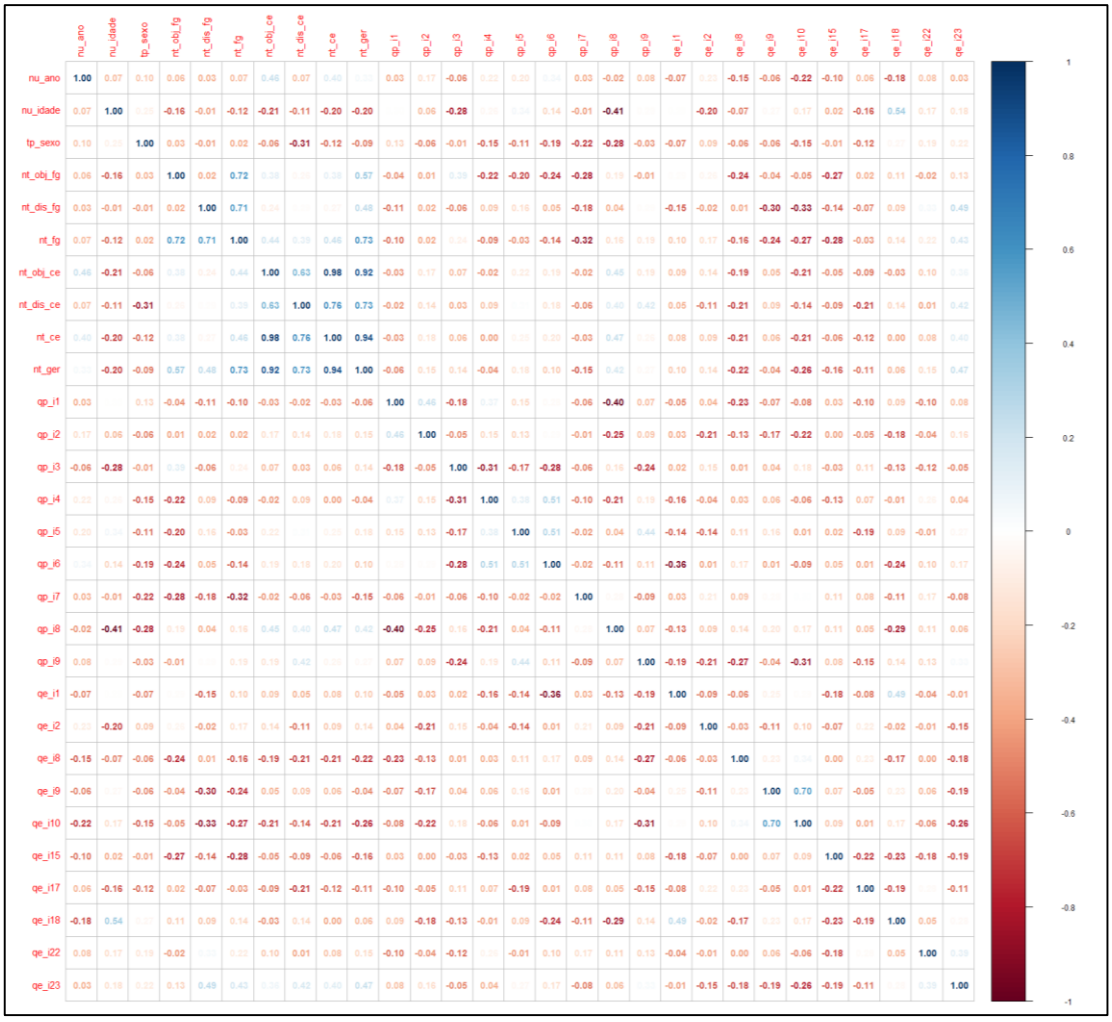
Os Gráficos 14 e 15 trazem a representação das correlações em formato *Square* e numérico, respectivamente

Gráfico 14: Representação geométrica para as correlações entre as variáveis.



Fonte: Própria.

Gráfico 15: Representação numérica para as correlações entre as variáveis.



Fonte: Própria.

Os resultados das correlações revelam que existe uma forte correlação positiva entre as notas, o que era esperado, visto que elas possuem diretamente relações entre si, por meio dos pesos atribuídos, contribuindo na composição final de determinadas notas.

Dentre as variáveis de nota, a nota específica (nt_ce) em conjunto com a nota objetiva específica (nt_obj_ce), possuem a maior correlação, sendo de 0.98. Em seguida a correlação da nota específica (nt_ce) e nota geral (nt_ger), tiveram um valor de 0.94. Outra forte correlação é da nota objetiva específica (nt_obj_ce) com a nota geral (nt_ger), sendo de 0,92.

Dessa forma, constata-se que tanto a nota objetiva específica, quanto nota específica, têm uma maior influência na nota geral da prova. O Quadro 13, apresenta os maiores valores das correlações entre as notas.

Quadro 13: Correlações entre as notas.

Primeira variável	Segunda variável	Correlação
nt_obj_ce	nt_ce	0,98
nt_ce	nt_ger	0,94
nt_obj_ce	nt_ger	0,92
nt_ce	nt_dis_ce	0,75
nt_fg	nt_ger	0,73
nt_dis_ce	nt_ger	0,73
nt_obj_fg	nt_fg	0,72
nt_dis_fg	nt_fg	0,71
nt_obj_ce	nt_dis_ce	0,62
nt_obj_fg	nt_ger	0,57

Fonte: Própria.

As relações entre as demais variáveis demonstram correlações fracas, tanto positivamente quanto negativamente, salvo, a correlação de 0.7 entre as variáveis qe_i9 (situação financeira) e qe_i10 (situação de trabalho), podendo ser explicada da seguinte maneira, quem tem renda, geralmente trabalha, quem não, depende de alguma ajuda financeira, por meio da família ou de outras pessoas. Outras correlações menos pertinentes são apresentadas no Quadro 14.

Quadro 14: Principais correlações entre as demais variáveis.

Primeira variável	Segunda variável	Correlação
nu_idade	qe_i18	0,54
qp_i6	qp_i4	0,51
qp_i6	qp_i5	0,51

Fonte: Própria.

Diante das perguntas que cercam esse trabalho, as questões que diretamente se destacaram na correlação, foram as da situação financeira e de trabalho, contudo, para uma melhor compreensão dessa correlação e das outras perguntas propostas, serão analisadas nas seções e subseções posteriores, no aprendizado não supervisionado (*apriori*) no *R studio*, e no aprendizado supervisionado (classificadores) novamente na ferramenta *WEKA*, na tentativa de se obter um panorama mais completo sobre as questões levantadas.

4.5 APLICAÇÃO DO ALGORITMO APRIORI

Esta seção refere-se à aplicação do algoritmo *Apriori*, dando continuidade com o cenário descritivo presente na primeira parte deste trabalho. O *Apriori* utiliza o conceito de aprendizado não supervisionado, ou seja, o aprendizado ocorre de forma autônoma, fazendo a detecção dos padrões presentes na base de dados, sem a avaliação prévia de saídas.

O *Apriori* apresenta o conhecimento por meio regras de associação, estando nas mãos de quem conduz as análises, o critério de escolha das regras mais relevantes para seu estudo. O suporte, a confiança, e o *lift* serão as métricas utilizada, apresentadas no capítulo de fundamentação teórica.

Para a utilização do *Apriori* foi necessário a utilização do *R studio*, ferramenta que incorpora a linguagem R, sendo escolhida em vez da *WEKA*, por dá a possibilidade de usar *scripts* que conseguem retirar a redundância de várias regras de associação, eliminando-as, sendo o resultado mais efetivo para as análises.

Para a criação do conjunto de regras pelo *Apriori* no R studio, foram utilizados os seguintes comandos: *libray(arules)*, utilizado para inicializar os recursos que estão presentes no pacote *arules*, e *rules = apriori(data),* para a criação das regras por meio da função *apriori.* Para esse estudo, o algoritmo retornou um total de 138 460 regras, com o suporte padrão de 0,1.

Figura 17: Quantidade de regras geradas pelo *Apriori*.

rules	Formal class rules
rules_teste	Large rules (138460 elements, 8.4 Mb)

Fonte: Própria.

A base de dados utilizada foi a que continha os valores categóricos, já que o *Apriori* necessita que as variáveis estivessem nesse formato, a base de dados, mesmo já estando no formato adequado, foi novamente transformada em *factor*. A seguinte linha de comando foi a responsável por essa ação: *data = data.frame(sapply(data,as.factor))*.

A quantidade de regras geradas pelo algoritmo foi enorme, visto que estava atendo ao suporte mínimo de 0.1. A Figura 17, apresenta a variável chamada *rules_teste* com uma enorme quantidade de regras. Para diminuir a quantidade de regras, o suporte (frequências dos *itemsets*), foi aumentado para 0.2, sendo adotado o nível de confiança de 0.75. O comando seguinte traz consigo essas correções, e ainda coloca as regras em uma listagem decrescente de acordo com o nível confiança: *rules = apriori(data, parameter=list(support=0.2, confidence=0.75))*.

O número de regras caiu drasticamente para 5019. Porém ainda, é considerado uma grande quantidade de regras para serem analisadas. Sabendo que existia uma grande quantidade de regras redundantes, as redundâncias precisavam ser eliminadas.

Para isso, foi utilizado um trecho de código, retirado de Yu-Wei, Chiu (2016). O código está presente na Figura 18, onde pode ser explicado da seguinte maneira, primeiramente, as regras são ordenadas pela métrica *lift*, depois é criado um subconjunto de regras associadas por meio da função *is.subject,* gerando uma matriz de *itemsets*, depois com a função *lower.tri*, coloca-se nas entradas, na parte inferior da matriz, o valor NA, inclusive na diagonal, em seguida com a função *colSums*, é criada uma matriz chamada *redundant*, que contará com os itemsets redundantes, sendo verificados pelos NA's contidos no subconjunto de regras criado anteriormente, depois é só fazer a subtração desses valores que estão contidos na matriz rules.sorted, salvando a lista com as regras melhores regras em rules.pruned, e em seguida essas regras são apresentadas pelo função inspect.

Figura 18: Código utilizado para remover as regras redundantes.

```
rules.sorted = sort(rules, by="lift")
subset.matrix = is.subset(rules.sorted, rules.sorted, sparse = FALSE)
subset.matrix[lower.tri(subset.matrix, diag = T)] = NA
redundant = colSums(subset.matrix, na.rm = T) >=1
rules.pruned = rules.sorted[!redundant]
inspect(rules.pruned)
```

Fonte: Yu-Wei e Chiu (2016).

Em posse das melhores regras selecionadas, elas precisaram ser salvas em uma planilha, e isso foi feito por meio do comando *write(rules.pruned, file = "data.csv", sep = ",")*. Inicialmente o tipo do arquivo da planilha era csv, depois convertido para xlsx, para melhor manipulá-las.

Por mais que as regras de associações obtidas sejam as melhores, elas precisam ter sentido com o contexto das perguntas apresentadas, e não serem tão óbvias, como pode ocorrer, então em seguida as análises foram feitas diante das sessenta e quatro regras.

As análises foram feitas, de acordo com as métricas e relevância para o trabalho. Vale ressaltar que o *lift*, em sessenta regras, é maior que 1, e isso significa que as regras são altamente associadas, no entanto, foram divididas em grupos de relevância diferentes, de acordo com o que já podia ser deduzido previamente.

As regras foram divididas nas seguintes categorias: não relevantes, relevantes, mas já era esperado, pouquíssima relevância, média relevância, e por fim altíssima relevância. As regras foram divididas para uma melhor compreensão e estão presentes nas Figuras 20, 21, e 22.

A Figura 20, apresenta três regras que foram consideradas de altíssima relevância. A primeira delas é a {nt_ce=C} → {net_ger=C}, evidenciando que para o estudante alcançar uma nota final igual a C, que compreende o intervalo de [41,60], será necessário obter uma nota equivalente na nota específica. Essa regra foi considerada de altíssima relevância por destacar a importância da nota específica na composição da nota final, com um nível de confiança em 1 e *lift* de 2,36.

A segunda regra em destaque é a {qe_i15=C} → {qe_I17=A}, onde apresenta que alunos que estudaram o ensino médio todo em escola pública, utilizam cota do recorte social. Por mais que esse cenário seja esperado, foi considerado de altíssima importância por seu grau de confiança ser de 0,92, ou seja, 92% dos estudantes concluintes da UERN do Campus Avançado de Natal, curso Ciência da Computação, que utilizaram o recurso social para iniciar o curso na universidade, estudaram o ensino médio em escola pública, do mesmo modo, o *lift* de 1,91, reafirma que o *itemset* {qe_I17=A} está altamente associado ao {qe_i15=C}.

Figura 19: Primeiro grupo de regras de associação geradas pelo *Apriori*.

Não relevantes | Relevantes, mas já era esperado | Pouquíssima relevância | Média relevância | Altíssima relevância

Regras	Suporte	Confiança	Lift
{nt_ce=C} => {nt_obj_ce=C}	0,21	1,00	2,60
{nt_ce=C} => {nt_ger=C}	0,21	1,00	2,36
{qE_i15=C} => {qE_i17=A}	0,21	0,92	1,91
{nt_ger=C} => {nt_fg=D}	0,35	0,82	1,70
{qp_i3=C,qE_i23=B} => {qp_i1=C}	0,21	1,00	1,68
{qp_i1=C,qE_i10=E} => {qE_i23=B}	0,27	0,82	1,65
{qE_i22=C} => {nt_ce=B}	0,25	0,93	1,61
{qp_i6=C,qE_i8=B} => {nt_ce=B}	0,21	0,92	1,59
{qp_i6=C,qE_i10=E} => {qp_i3=C}	0,21	0,79	1,57
{nt_obj_fg=D,qp_i5=B} => {qp_i3=C}	0,21	0,79	1,57
{qp_i8=D,qE_i23=B} => {qE_i10=E}	0,23	0,86	1,54
{nt_obj_ce=B,qp_i8=D} => {qE_i10=E}	0,21	0,85	1,52
{qE_i9=C} => {nu_idade=B}	0,21	0,79	1,51
{qE_i8=A} => {nu_idade=B}	0,21	0,79	1,51
{qp_i5=B,qE_i17=B} => {qp_i4=B}	0,21	0,79	1,46
{nt_ce=B,qE_i17=B} => {qp_i4=B}	0,21	0,79	1,46
{qp_i7=A} => {qp_i5=B}	0,25	0,81	1,46
{nt_ce=B,qE_i23=B} => {qE_i10=E}	0,25	0,81	1,46
{qp_i9=B} => {qE_i10=E}	0,23	0,80	1,43
{qp_i8=D,qE_i8=B} => {qE_i10=E}	0,23	0,80	1,43
{nt_ce=B,qp_i3=C} => {qp_i5=B}	0,23	0,80	1,43
{qp_i1=C,qp_i5=B} => {nt_ce=B}	0,27	0,82	1,43
{qp_i8=D,qE_i23=B} => {qp_i5=B}	0,21	0,79	1,41

Fonte: Própria.

A terceira regra, {qp_I9=B} → {qe_I10=E}, indica que o estudante que demorou entre uma e duas horas para concluir a prova, trabalhava 40 horas semanais ou mais, com grau de confiança de 0,8, e *lift* de 1,43.

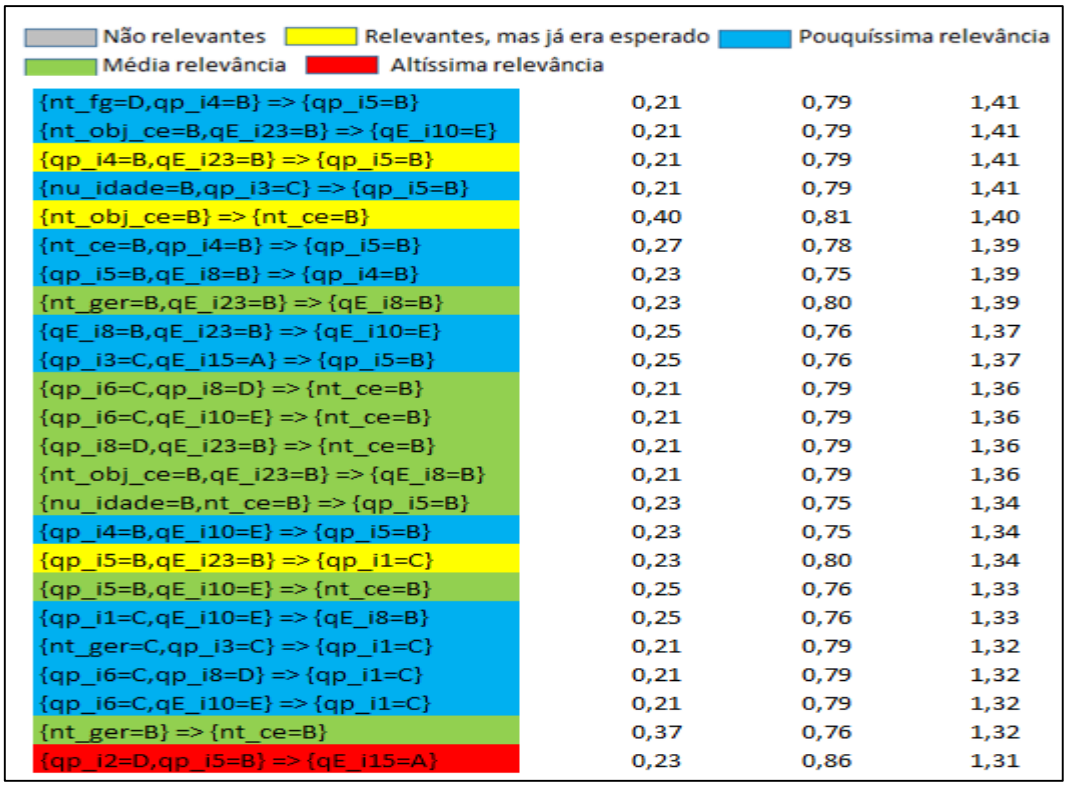
Na Figura 20, ainda vale destacar a regra {nt_ce=C} → {nt_obj_ce=C} que foi colocada como média relevância porque quem tira uma nota final do componente específico classificada como C, tem que pelo menos ter conseguido obter o mesmo valor na parte objetiva, já que as notas discursivas para o componente específico são baixas. O mesmo raciocínio segue para a regra {nt_ger=C} → {nt_fg= D}.

As demais regras que estão classificadas como de média relevância, pouquíssima, e já esperadas, usam interpretações parecidas, apoiando-se também

em análises anteriores na fase descritiva, ou que simplesmente não interessam, ou, tem menor contribuição para o trabalho. Essas interpretações seguem para os demais grupos das figuras seguintes que se encaixam nessas modalidades.

A Figura 21 apresenta o segundo grupo de regras. Nela contém apenas uma regra considerada de altíssima relevância.

Figura 20: Segundo grupo de regras de associação geradas pelo *Apriori*.



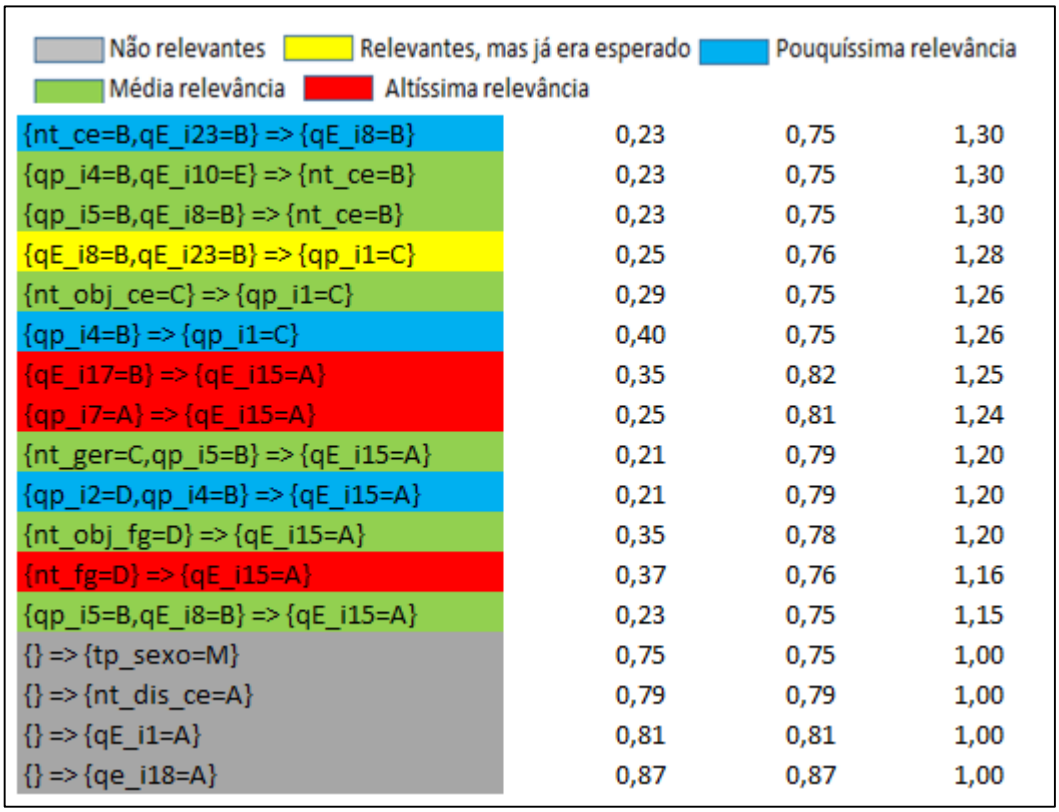
{nt_fg=D,qp_i4=B} => {qp_i5=B}	0,21	0,79	1,41
{nt_obj_ce=B,qE_i23=B} => {qE_i10=E}	0,21	0,79	1,41
{qp_i4=B,qE_i23=B} => {qp_i5=B}	0,21	0,79	1,41
{nu_idade=B,qp_i3=C} => {qp_i5=B}	0,21	0,79	1,41
{nt_obj_ce=B} => {nt_ce=B}	0,40	0,81	1,40
{nt_ce=B,qp_i4=B} => {qp_i5=B}	0,27	0,78	1,39
{qp_i5=B,qE_i8=B} => {qp_i4=B}	0,23	0,75	1,39
{nt_ger=B,qE_i23=B} => {qE_i8=B}	0,23	0,80	1,39
{qE_i8=B,qE_i23=B} => {qE_i10=E}	0,25	0,76	1,37
{qp_i3=C,qE_i15=A} => {qp_i5=B}	0,25	0,76	1,37
{qp_i6=C,qp_i8=D} => {nt_ce=B}	0,21	0,79	1,36
{qp_i6=C,qE_i10=E} => {nt_ce=B}	0,21	0,79	1,36
{qp_i8=D,qE_i23=B} => {nt_ce=B}	0,21	0,79	1,36
{nt_obj_ce=B,qE_i23=B} => {qE_i8=B}	0,21	0,79	1,36
{nu_idade=B,nt_ce=B} => {qp_i5=B}	0,23	0,75	1,34
{qp_i4=B,qE_i10=E} => {qp_i5=B}	0,23	0,75	1,34
{qp_i5=B,qE_i23=B} => {qp_i1=C}	0,23	0,80	1,34
{qp_i5=B,qE_i10=E} => {nt_ce=B}	0,25	0,76	1,33
{qp_i1=C,qE_i10=E} => {qE_i8=B}	0,25	0,76	1,33
{nt_ger=C,qp_i3=C} => {qp_i1=C}	0,21	0,79	1,32
{qp_i6=C,qp_i8=D} => {qp_i1=C}	0,21	0,79	1,32
{qp_i6=C,qE_i10=E} => {qp_i1=C}	0,21	0,79	1,32
{nt_ger=B} => {nt_ce=B}	0,37	0,76	1,32
{qp_i2=D,qp_i5=B} => {qE_i15=A}	0,23	0,86	1,31

Fonte: Própria.

A regra {qp_I2= D, qp_I5=B} → {qe_I15=A} traz consigo a informação de que os estudantes que não utilizaram cotas, marcaram como difícil o grau de dificuldade no componente específico da prova, e disseram que a maioria dos enunciados estavam claros e objetivos. Essa regra foi considerada importante por estabelecer que esse grupo de estudantes, têm um padrão mais acentuado de respostas para as questões qp_I2, qp_I5, dos que utilizam as cotas. Com o grau de confiança de 0,86 e *lift* de 1,31.

O terceiro grupo de regras estão presentes na Figura 22, com destaque para {qE_I17=B} → {qE_I15=A}, {qp_I7} → {qE_I15=A}, e {nt_fg=D} → {qE_I15=A}.

Figura 21: Terceiro grupo de regras de associação geradas pelo *Apriori*.



{nt_ce=B,qE_i23=B} => {qE_i8=B}	0,23	0,75	1,30
{qp_i4=B,qE_i10=E} => {nt_ce=B}	0,23	0,75	1,30
{qp_i5=B,qE_i8=B} => {nt_ce=B}	0,23	0,75	1,30
{qE_i8=B,qE_i23=B} => {qp_i1=C}	0,25	0,76	1,28
{nt_obj_ce=C} => {qp_i1=C}	0,29	0,75	1,26
{qp_i4=B} => {qp_i1=C}	0,40	0,75	1,26
{qE_i17=B} => {qE_i15=A}	0,35	0,82	1,25
{qp_i7=A} => {qE_i15=A}	0,25	0,81	1,24
{nt_ger=C,qp_i5=B} => {qE_i15=A}	0,21	0,79	1,20
{qp_i2=D,qp_i4=B} => {qE_i15=A}	0,21	0,79	1,20
{nt_obj_fg=D} => {qE_i15=A}	0,35	0,78	1,20
{nt_fg=D} => {qE_i15=A}	0,37	0,76	1,16
{qp_i5=B,qE_i8=B} => {qE_i15=A}	0,23	0,75	1,15
{} => {tp_sexo=M}	0,75	0,75	1,00
{} => {nt_dis_ce=A}	0,79	0,79	1,00
{} => {qE_i1=A}	0,81	0,81	1,00
{} => {qe_i18=A}	0,87	0,87	1,00

Fonte: Própria.

A regra {qE_I17=B} → {qE_I15=A} coloca em evidência que estudantes que estudaram o ensino médio todo em escola privada (particular), estão altamente associados com o fato de não usarem diretamente as cotas para o ingresso na universidade, com grau de confiança de 0,82, e *lift* de 1,25.

As últimas regras que foram consideradas de altíssima importância, revelam que os estudantes que não utilizaram cotas no seu ingresso no curso, tem as melhores notas na formação geral, mas consideram o desconhecimento do conteúdo como dificuldade para responder à prova, sendo que a média para essa questão é a forma diferente de abordagem do conteúdo, como forma de dificuldade.

As regras de associações foram importantes para confirmar uma suspeita sobre o perfil de estudantes de escola pública e escola privada, no que se refere a política de ingresso no curso, e trazer alguns padrões imprevisíveis como em {qp_I7} → {qE_I15=A}, mas, com exceção de uma regra, {nt_fg=D} → {qE_I15=A}, o Apriori não extraiu outras informações padronizadas sobre o questionário socioeconômico e de

percepção da prova dos estudantes, que tivessem um impacto negativo ou positivo no desempenho da nota final, levantadas nas questões norte do trabalho.

Dessa maneira, na próxima seção, no que concerne a etapa da classificação, esperasse com a utilização dos métodos de classificação, alcançar conclusões se existe ou não padrões para responder as questões propostas.

4.6 APLICAÇÃO DOS CLASSIFICADORES

Nesta seção será aplicada a técnica de classificação de dados. Essa técnica de mineração de dados, utilizada o aprendizado supervisionado, verificando as saídas, que nesse caso são chamadas de classe, com intuito de obter uma generalização do conhecimento presente nos dados, o que não significa decorar todo o conjunto de dados.

A generalização será feita a partir de uma base de treinamento e outra de teste do conjunto de dados utilizados, essa divisão de bases, neste caso, será obtida através da técnica *cross-validation*, mencionada no capítulo de metodologia. Dessa maneira os classificadores utilizam suas técnicas para obter um aprendizado sobre os dados.

A ferramenta utilizada para aplicar as técnicas de classificação, foi a WEKA. Nessa ferramenta, depois do treinamento e teste, gera um modelo que vai conter métricas de o quão bom foi o processo de aprendizagem.

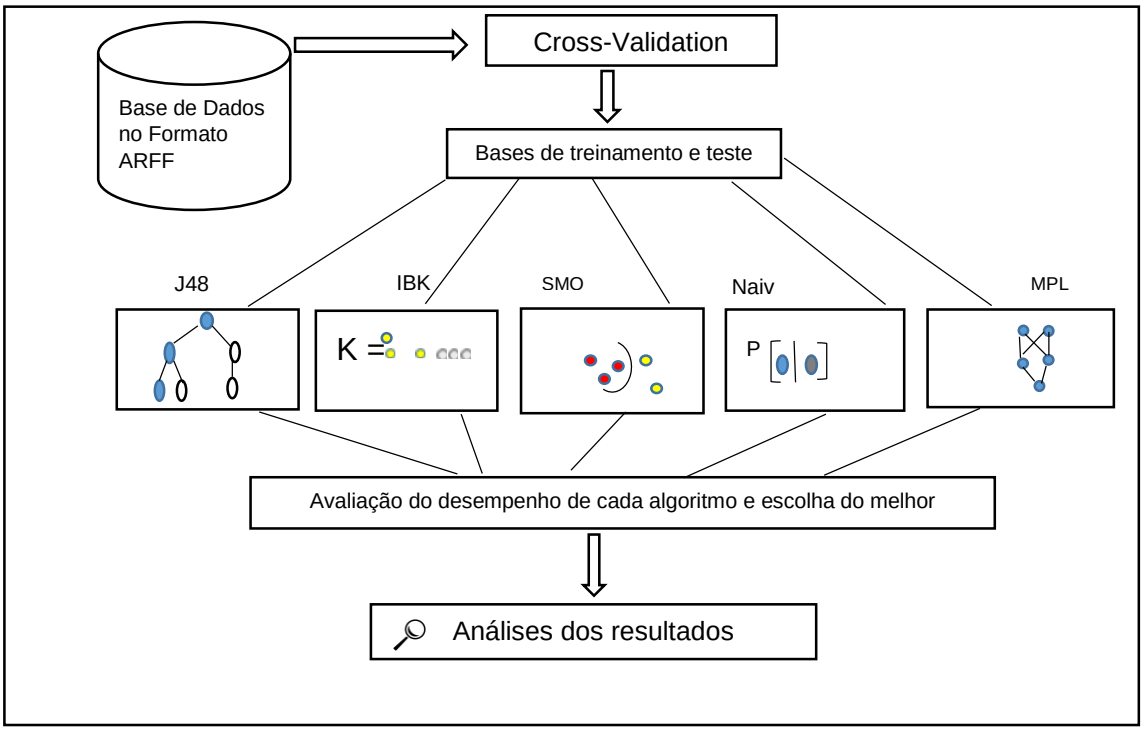
Além das informações que mensuram o desempenho da aprendizagem, a WEKA, ainda possui uma variedade de opções de visualização dos dados, que servem para analisar possíveis padrões no conjunto de dados utilizados.

A WEKA ainda exibe a configuração da estrutura dos classificadores utilizados. Em alguns casos, dependo da abordagem do classificador, fica fácil fazer a análise através da estrutura, como é o caso das árvores de decisão, outros como máquinas de vetores de suporte e redes neurais de multicamadas, não são tão triviais para de se fazer as análises.

Para este trabalho, serão utilizados os classificadores J48 (árvore de decisão), Naive Bayes (redes bayesianas), IBk (abordagem KNN vizinhos), SMO (máquinas de vetores de suporte), e Mutilayer Perceptron (rede neural multicamadas).

A Figura 22, apresenta um resumo do que dos passos que acompanham a classificação para este trabalho.

Figura 22: Resumo dos passos utilizados na classificação.



Fonte: Própria.

A Figura 22, por ser um resumo, omite algumas etapas intermediárias que constam no trabalho para a classificação, no entanto, serão apresentadas no decorrer dessa seção.

A primeira etapa, foi a escolha das variáveis que serão utilizadas na classificação. Com o propósito de responder as questões que norteiam este trabalho, as seguintes variáveis foram utilizadas como classes: qE_I8, qE_I9, qE_I10, qE_I22, qE_I23, qp_I2, qp_I5, e qp_I9. Cinco das variáveis do questionário socioeconômico, e três do questionário de percepção da prova.

Com as variáveis selecionadas, aplicou-se a classificação, e por meio da área abaixo da curva ROC (AROC), apresentada no capítulo de metodologia, começaram as primeiras análises.

Como os dados não estavam balanceados, os classificadores apresentaram um nível muito insatisfatório de predição. Nas Tabelas 19, e 20, estão os valores das áreas em questão. Quanto mais próximo do valor 1, melhor a classificação de verdadeiros positivos.

Tabela 19: Valores da AROC para os dados não balanceados.

Classificador	QE_I8					QE_I9					
	A	B	C	E	Média	B	C	D	E	F	Média
J48	0,43	0,34	0,66	0,34	0,44	0,83	0,48	0,29	0,42	0,43	0,49
IBk	0,66	0,47	0,44	0,40	0,49	0,69	0,50	0,56	0,54	0,44	0,55
NaiveBayes	0,38	0,44	0,67	0,10	0,40	0,86	0,41	0,47	0,79	0,44	0,59
MPL	0,52	0,45	0,41	0,74	0,53	0,94	0,46	0,38	0,60	0,94	0,66
SMO	0,46	0,44	0,34	0,45	0,42	0,92	0,47	0,36	0,62	0,48	0,57
	QE_I22					QE_I10					
	A	B	C	E	Média	A	B	C	D	E	Média
J48	0,44	0,45	0,48	0,63	0,50	0,78	0,12	0,10	0,50	0,71	0,44
IBk	0,35	0,42	0,67	0,66	0,52	0,61	0,64	0,41	0,57	0,55	0,56
NaiveBayes	0,41	0,50	0,63	0,25	0,44	0,87	0,06	0,02	0,26	0,78	0,40
MPL	0,25	0,45	0,73	0,73	0,54	0,81	0,53	0,34	0,42	0,77	0,57
SMO	0,76	0,66	0,80	0,98	0,80	0,74	0,17	0,20	0,25	0,57	0,39

Fonte: Própria.

A intenção de usar essa técnica, é que ela apresenta um panorama mais amplo comparado com as outras métricas, por exibir como está a margem de acertos de predição para cada saída da classe.

Tabela 20: Valores da AROC para os dados não balanceados.

Classificador	QE_I23						QP_I2						
	A	B	C	D	E	Média	A	B	C	D	E	.	Média
J48	0,40	0,51	0,49	0,51	0,38	0,46	0,46	0,47	0,50	0,49	0,40	0,66	0,50
IBk	0,68	0,47	0,47	0,43	0,48	0,51	0,89	0,83	0,63	0,44	0,33	0,48	0,60
NaiveBayes	0,08	0,62	0,53	0,13	0,37	0,34	0,23	0,13	0,56	0,41	0,58	0,85	0,42
MPL	0,94	0,56	0,58	0,33	0,24	0,53	0,47	0,75	0,57	0,35	0,44	0,86	0,56
SMO	0,97	0,52	0,48	0,36	0,43	0,55	0,29	0,16	0,61	0,36	0,34	0,78	0,42
	QP_I5						QP_I9						
	A	B	C	D	.	Média	B	C	D	E	.	-	Média
J48	0,62	0,52	0,57	0,66	0,94	0,66	0,63	0,51	0,58	0,43	0,69	-	0,57
IBk	0,70	0,49	0,42	0,42	0,55	0,52	0,42	0,56	0,56	0,88	0,60	-	0,60
NaiveBayes	0,20	0,63	0,47	0,60	0,84	0,55	0,50	0,47	0,52	0,08	0,84	-	0,48
MPL	0,70	0,59	0,37	0,75	0,79	0,64	0,49	0,60	0,49	0,22	0,90	-	0,56
SMO	0,86	0,57	0,26	0,48	0,78	0,59	0,51	0,50	0,49	0,29	0,68	-	0,49

Fonte: Própria.

Em algumas saídas de certas classes, alguns algoritmos obtiveram uma predição considerada boa, mas difere das outras saídas que compõem a classe, dessa maneira na média final do desempenho do classificador, todos os testes, com exceção da classe QE_I22 (com o classificador SMO), os dados desbalanceados demostraram um rendimento final insatisfatório.

Com base nesse rendimento, os dados foram balanceados, e se obteve melhores desempenhos. Os filtros utilizados foram o SMOTE para aumentar sinteticamente as instâncias da *“Minority class”*, e depois o filtro *SpreadSubsample* para diminuir e igualar as demais saídas da classe, para que haja o balanceamento. Em seguida o filtro *Ramdomize* foi aplicado para embaralhar os registros do novo conjunto de dados, e isso foi feito porque quando aplicado esses filtros, geralmente os dados ficam organizados ordenados de acordo com a saída da classe, dessa forma, poderia prejudicar na distribuição das bases de treinamento e teste.

Algumas saídas de determinadas classes foram removidas, devido ao fato, de possuir apenas uma instância representante, sendo assim, o SMOTE não pode ser aplicado, devido à restrição de pelo menos uma instância vizinha, já que utiliza o algoritmo KNN.

Depois de balancear os dados, novos valores para áreas abaixo da curva ROC foram recalculados e estão presentes nas Tabelas 21 e 22.

Tabela 21: Valores da AROC para os dados balanceados.

Classificador	QE_I8					QE_I9					
	A	B	C	E	Média	B	C	D	E	F	Média
J48	0,53	0,49	0,64	0,77	0,60	0,76	0,51	0,70	0,65	0,88	0,70
IBk	0,56	0,47	0,70	0,93	0,66	0,77	0,58	0,57	0,53	0,93	0,68
NaiveBayes	0,65	0,69	0,98	1.0	0,83	0,88	0,74	0,67	0,78	1.0	0,81
MPL	0,70	0,76	0,86	1.0	0,83	0,94	0,72	0,48	0,75	1.0	0,78
SMO	0,74	0,69	0,67	0,92	0,75	0,89	0,68	0,47	0,71	0,93	0,74
	QE_I22					QE_I10					
	A	B	C	E	Média	A	B	C	D	E	Média
J48	0,82	0,64	0,68	0,77	0,72	0,67	0,87	0,87	0,76	0,66	0,77
IBk	0,77	0,65	0,67	0,98	0,77	0,71	0,95	0,90	0,85	0,65	0,81
NaiveBayes	0,80	0,61	0,56	0,99	0,74	0,90	1.0	1.0	1.0	0,83	0,95
MPL	0,76	0,66	0,80	0,98	0,80	0,95	1.0	0,98	0,92	0,94	0,96
SMO	0,76	0,63	0,82	0,93	0,78	0,88	0,87	0,90	0,82	0,90	0,87

Fonte: Própria.

Tabela 22: Valores da AROC para os dados balanceados.

	QE_I23						QP_I2						
Classificador	A	B	C	D	E	Média	A	B	C	D	E	.	Média
J48	0,93	0,65	0,55	0,93	0,81	0,77	-	-	0,69	0,59	0,86	0,90	0,76
IBk	1.0	0,50	0,64	0,92	0,81	0,77	-	-	0,60	0,60	0,83	0,83	0,71
NaiveBayes	1.0	0,50	0,76	0,99	0,87	0,82	-	-	0,78	0,69	0,99	1.0	0,86
MPL	1.0	0,62	0,71	0,91	0,72	0,79	-	-	0,69	0,59	0,94	0,95	0,79
SMO	1.0	0,54	0,58	0,84	0,69	0,73	-	-	0,52	0,58	0,88	0,89	0,72
	QP_I5						QP_I9						
	A	B	C	D	.	Média	B	C	D	E	.	-	Média
J48	0,86	0,48	0,72	0,72	0,96	0,75	0,72	0,65	0,53	0,85	0,94	-	0,74
IBk	0,97	0,45	0,59	0,80	0,94	0,75	0,60	0,62	0,56	0,92	0,85	-	0,71
NaiveBayes	1.0	0,70	0,66	0,71	1.0	0,83	0,60	0,66	0,63	1.0	1.0	-	0,78
MPL	0,99	0,63	0,43	0,91	0,99	0,79	0,55	0,65	0,55	0,97	0,97	-	0,74
SMO	0,99	0,50	0,39	0,83	0,94	0,73	0,52	0,59	0,66	0,93	0,87	-	0,71

Fonte: Própria.

Com os novos valores da área abaixo da curva ROC, os melhores classificadores foram escolhidos para dar continuação com as análises, de acordo com as médias finais de cada classificador. Em caso de empate, foi escolhido, o classificador uma menor variação para cada saída da classe.

Os classificadores NaiveBayes e Multilayer Perceptron tiveram os melhores resultados para os dados balanceados. O SMO se manteve em algumas ocasiões, como um bom classificador, mas com um desempenho menor, comparado ao Naive e ao MPL. O IBk, assim com o J48, fica um pouco abaixo dos demais já citados, sendo classificadores com menos resultados favoráveis.

Para as questões QE_I9, QE_I23, QP_I2, QP_I5, e QP_I9 o classificador escolhido para seguir com as análises foi o *NaiveBayes*. Já o MPL foi o classificador escolhido para as questões QE_I8, QE_I10, e QE_I22.

A questão QE_I8 é sobre a renda familiar. E uma das perguntas levantadas, foi se a estudantes com renda familiar baixa, que no caso seria até três salários mínimos, tiveram um desempenho menor no resultado final da prova, do que os demais.

A intenção dessa pergunta é verificar se a situação econômica do estudante possa ser um dos fatores que influenciam negativamente no seu desempenho.

Os resultados do modelo gerado pelo MPL estão presentes na Figura 23. O modelo conseguiu acertar 65% das instâncias, quer dizer que, de acordo com a estatística Kappa, o modelo gerado está em um nível de concordância moderada para

com o treinamento do classificador. A matriz de confusão, exibe como foi a distribuição da classificação.

Figura 23: Primeiro modelo gerado para a classe QE_I8.

```
Correctly Classified Instances          26               65      %
Incorrectly Classified Instances        14               35      %
Kappa statistic                          0.5333
Mean absolute error                      0.1851
Root mean squared error                  0.3727
Relative absolute error                 49.3717 %
Root relative squared error             86.0602 %
Total Number of Instances               40

=== Confusion Matrix ===

 a b c d   <-- classified as
 6 3 1 0 | a = A
 4 4 2 0 | b = B
 0 2 7 1 | c = C
 1 0 0 9 | d = E
```

Fonte: Própria.

Na tentativa de melhorar o modelo, foi utilizado na aba seleção de atributos da ferramenta WEKA a avaliação de atributos por meio do filtro *ClassifierAttributeEval* que utiliza o método de busca *Ranker*. Com essa opção, é possível ordenar em um ranque, o valor de cada atributo de acordo com o classificador. Dessa maneira é possível eliminar atributos que contribuem menos para a classificação.

Do conjunto de 29 atributos, incluindo a classe, foram eliminados dois: QE_I18 e QE_I22. Foi verificado que com a eliminação desses atributos houve um pequeno aumento de instâncias corretamente classificadas.

A Figura 24, contém os novos parâmetros do modelo. O que mudou em relação ao modelo anterior, foi a questão da classificação de mais uma instância correta para as saídas B e E. Segundo a estatística *Kappa*, o modelo continua sendo de concordância moderada.

Assim sendo, o modelo tem certas limitações e pode não revelar padrões sofisticados.

Figura 24: Segundo modelo gerado para a classe QE_I8.

```
Correctly Classified Instances          28               70      %
Incorrectly Classified Instances        12               30      %
Kappa statistic                          0.6
Mean absolute error                      0.1909
Root mean squared error                  0.3789
Relative absolute error                 50.9055 %
Root relative squared error             87.4921 %
Total Number of Instances               40

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  6  3  1  0 |  a = A
  3  5  2  0 |  b = B
  0  2  7  1 |  c = C
  0  0  0 10 |  d = E
```

Fonte: própria.

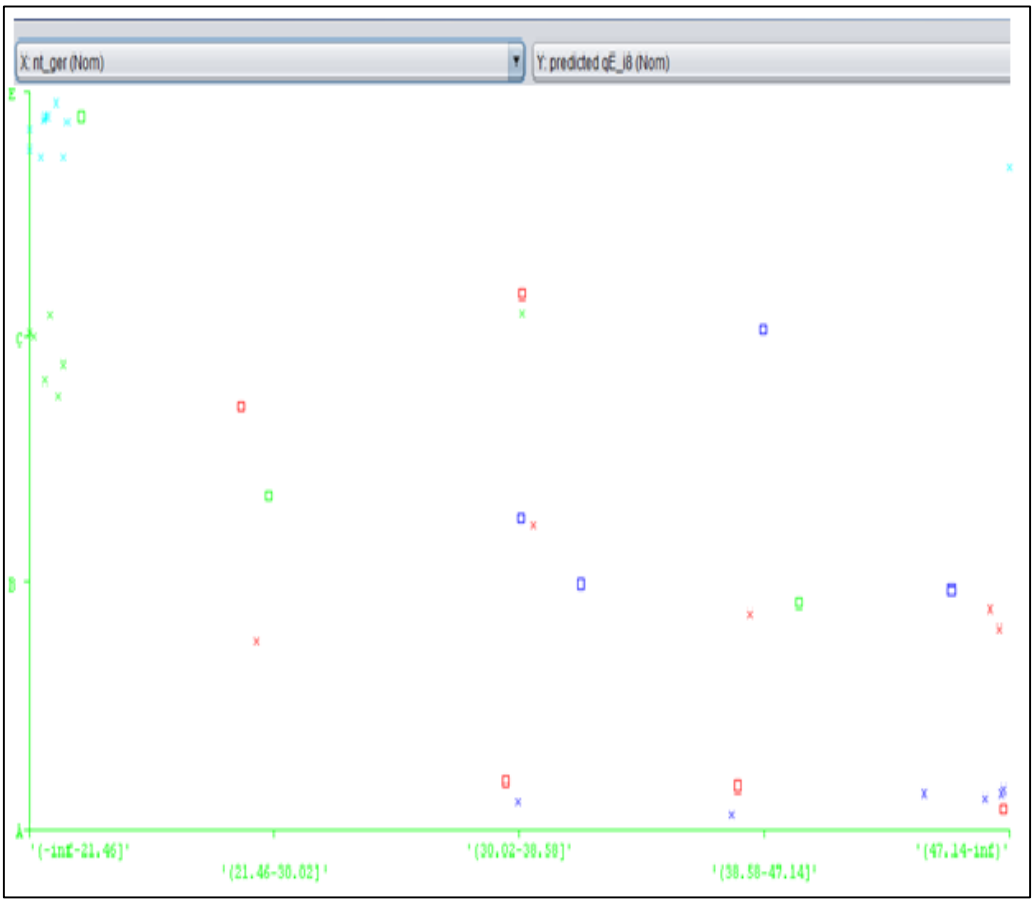
Uma das razões que pode explicar o fato da classificação não ter tido um desempenho melhor, refere-se a faixa salarial.

Como a faixa de salários nas opções para essa questão estão em intervalos grandes, o classificador erra em predizer principalmente as saídas A e B, porque não tem como saber por exemplo, se quem marcou a letra B, ganhava apenas 4 salários mínimos ou 10 salários mínimos, podendo esses grupos ter características similares em outros atributos.

Na tentativa de buscar algum padrão nesse modelo gerado para a questão QE_I8, com intuito de responder à pergunta sobre a renda anteriormente levantada, o modo de visualização da predição foi analisado. Não se pode afirmar que os estudantes de renda baixa têm as piores notas.

Contudo dá para perceber na Figura 25, que há uma concentração mais acentuada da opção dos estudantes que ganham até 3 salários mínimos no intervalo da maior nota. Dessa maneira, mesmo sem um padrão mais refinado, pode-se dizer que a renda familiar não é um dos fatores determinantes para um desempenho melhor na nota final.

Figura 25: Visualização da predição para a questão QE_I8.



Fonte: Própria.

Continuando a falar sobre renda, a questão QE_I9, pergunta sobre a situação financeira individual do estudante. Dessa vez foi utilizado o classificador NaiveBayes. Aplicando a classificação obteve-se um modelo com 62% das instâncias corretamente classificadas, e segundo a estatística Kappa, do mesmo modo que em QE_I8, o modelo tem um nível de concordância moderada com treinamento do classificador.

A Figura 26, apresenta o modelo gerado por esse classificador.

Figura 26: Primeiro modelo gerado para a classe QE_I9.

```
Correctly Classified Instances          31               62      %  === Confusion Matrix ===
Incorrectly Classified Instances        19               38      %
Kappa statistic                          0.525                        a  b  c  d  e   <-- classified as
Mean absolute error                      0.1752                       7  1  0  2  0 |  a = B
Root mean squared error                  0.3822                       0  5  4  1  0 |  b = E
Relative absolute error                 54.7515 %                     0  2  3  5  0 |  c = D
Root relative squared error             95.555  %                     2  0  2  6  0 |  d = C
Total Number of Instances               50                            0  0  0  0 10 |  e = F
```

Fonte: Própria.

Fazendo a análise da matriz de confusão, pode-se identificar que os extremos parecem ter um perfil mais harmônico, visto que a letra B, que representa que o estudante não tem renda e seus gastos não financiados por seus familiares, e a letra E, que representa que o estudante é o principal responsável pelo sustento da família, tiveram uma melhor predição.

Contudo, na tentativa de melhorar o modelo, a seleção de atributos foi aplicada novamente. Os atributos removidos foram: QP_I1, QP_I2, QP_I3, QP_I5, QP_I6, QP_I9, QE_I15, QE_I18, E QE_I22.

Entretanto, não houve uma melhorar significativa para o modelo no geral, mas o classificador conseguiu, predizer mais duas instâncias para a letra B. A Figura 27, apresenta o novo modelo.

Figura 27: Segundo modelo gerado para a classe QE_I9.

```
Correctly Classified Instances          32               64      %  === Confusion Matrix ===
Incorrectly Classified Instances        18               36      %
Kappa statistic                          0.55                         a  b  c  d  e   <-- classified as
Mean absolute error                      0.1539                       9  0  1  0  0 |  a = B
Root mean squared error                  0.3343                       0  5  4  1  0 |  b = E
Relative absolute error                 48.1064 %                     1  4  3  2  0 |  c = D
Root relative squared error             83.5837 %                     2  1  2  5  0 |  d = C
Total Number of Instances               50                            0  0  0  0 10 |  e = F
```

Fonte: Própria.

Como o NaiveBayes possibilita o cálculo das probabilidades para cada saída de uma forma mais intuitiva, o que difere do MPL, a análise para os modelos gerados pelo classificador, serão feitas de acordo com a tabela de frequência gerada na classificação.

A tabela a ser analisada será referente a nota final, dado que o objetivo é verificar como está o comportamento da saída em relação a nota. A Figura 28, traz os valores que o classificador gerou.

Figura 28: Tabela gerada pelo Naivebayes para a variável nt_ger (QE_I9).

nt_ger					
'(-inf-21.46]'	3.0	2.0	2.0	1.0	1.0
'(21.46-30.02]'	1.0	3.0	2.0	5.0	1.0
'(30.02-38.58]'	4.0	6.0	3.0	4.0	10.0
'(38.58-47.14]'	2.0	3.0	4.0	2.0	1.0
'(47.14-inf)'	5.0	1.0	4.0	3.0	2.0
[total]	15.0	15.0	15.0	15.0	15.0

Fonte: Própria.

Com os valores foi aplicado o teorema de Bayes, para cada saída da classe. Os valores obtidos com a aplicação do teorema estão presentes na Figura 29.

Figura 29: Valores obtidos pelo teorema de Bayes para QE_I9 (nt_ger).

	B	E	D	C	F	TOTAL
intevalo de nota 1	33,3333	22,22222	22,22222	11,11111	11,11111	100,0000
intervalo de nota 2	8,33333	25	16,66667	41,66667	8,333333	100,0000
intervalo de nota 3	14,8148	22,22222	11,11111	14,81481	37,03704	100,0000
intervalo de nota 4	16,6667	25	33,33333	16,66667	8,333333	100,0000
intervalo de nota 5	33,3333	6,666667	26,66667	20	13,33333	100,0000

Fonte: própria.

Diante dos valores da Figura 29, nota-se que não traz nenhuma evidência de um padrão um consistente. Os estudantes que declararam que não tinham renda e dependiam de suas famílias têm a maior probabilidade de tirar as melhores e piores notas, com 33,33% para ambos os casos.

As demais saídas também não apontam um perfil com maior probabilidade de estarem entre as piores notas, a não ser os estudantes que marcaram a letra C, que representa o grupo dos que tem renda, mas recebem ajuda da família, com 41,7% de

chances de estarem no segundo intervalo de nota, que corresponde ao intervalo (21.46-30.02], entretanto existe a probabilidade de 20% de estarem no intervalo com as maiores notas.

Apesar dos modelos criados serem de satisfação mediana, o parâmetro renda presente nas questões QE_I8, e QE_I9 não podem ser considerados como um dos fatores do desempenho negativo dos estudantes na prova do ENADE, por apresentar predições que deixa claro que, apesar das diferenças por parte da renda, os estudantes contém características semelhantes em outros atributos, que fazem com que os melhores classificadores confundam os perfis com base no parâmetro renda.

Seguindo com as análises da MD, a próxima classe a ser aplicada a classificação é a QE_I10, que retrata a situação de trabalho dos estudantes. O classificador escolhido para essa classe foi o MPL.

Com 82% das instâncias classificadas corretamente, o modelo criado para a questão QE_I10, obteve até agora o melhor desempenho. Estando segundo a estatística kappa, em nível de concordância substancial, sendo considerado como um bom modelo. A Figura 30, exibe as métricas para esse modelo.

Figura 30: Modelo gerado para a classe QE_I10.

```
Correctly Classified Instances          41               82      %
Incorrectly Classified Instances         9               18      %
Kappa statistic                          0.775
Mean absolute error                      0.0854
Root mean squared error                  0.2445
Relative absolute error                 26.7018 %
Root relative squared error             61.1323 %
Total Number of Instances               50

=== Confusion Matrix ===

 a b c d e   <-- classified as
 8 1 0 1 0 | a = A
 1 7 0 2 0 | b = E
 1 0 9 0 0 | c = C
 0 1 1 8 0 | d = D
 0 0 1 0 9 | e = B
```

Fonte: Própria.

Na tentativa de melhorar ainda mais o modelo foi aplicado o filtro de seleção de atributos, mas não retornou uma melhora para o modelo. Dessa maneira as análises partiram do modelo da Figura 30.

A pergunta que estava em volta da questão QE_I10, era se os estudantes que trabalhavam tiveram um rendimento pior do que os que apenas estudavam. A conclusão que se teve, foi que sim.

A concentração das melhores notas está entre os alunos que não trabalhavam. Os estudantes que responderam trabalhavam eventualmente, os que trabalhavam até

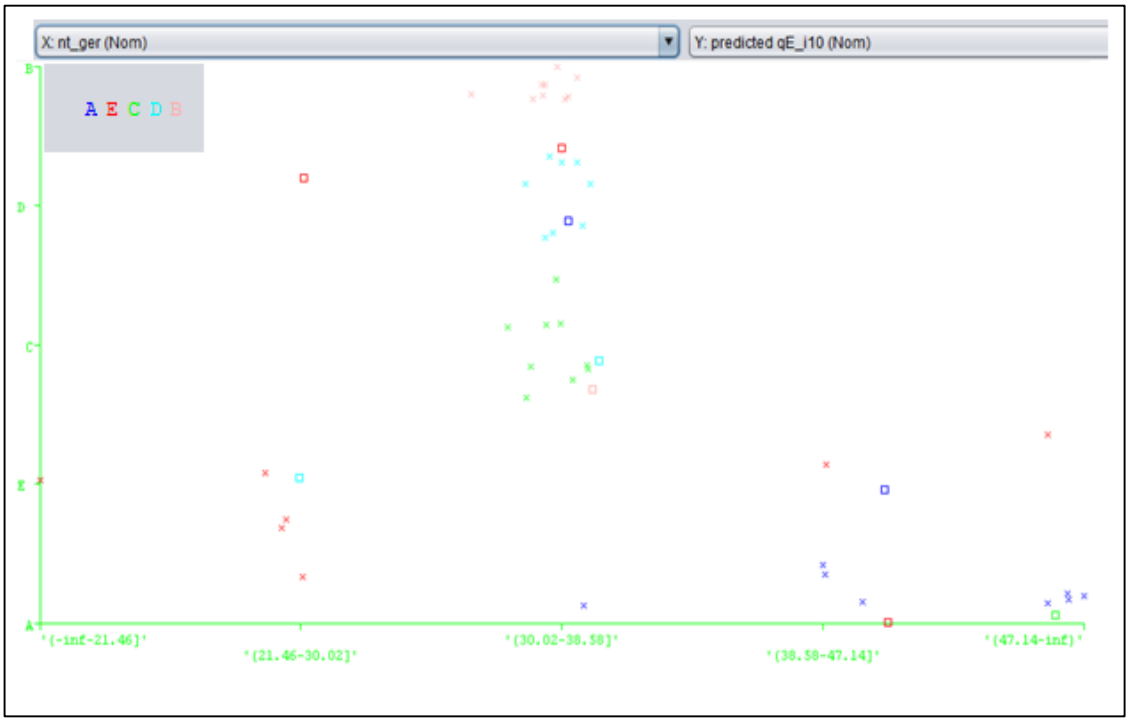
20 horas semanais, e os que trabalhavam de 21 a 39 horas semanais, ficaram concentrados no intervalo de nota 3, o nível intermediário entre as melhores e piores notas.

Ficando os estudantes que trabalhavam 40 horas semanais ou mais, nos intervalos das piores notas. Sendo que existe pontos fora da curva, que obtiveram as melhores notas, mas foge do padrão apresentado pelos demais.

Sendo assim, o parâmetro situação de trabalho, um fortíssimo candidato que entra na lista de fatores que podem influenciar no desempenho da nota final do ENADE.

A Figura 31, apresenta a distribuição da nota final de acordo com a situação de trabalho.

Figura 31: Predição para a classe QE_I10.



Fonte: Própria.

A próxima classificação é sobre a questão QE_I22 referente a quantidade de livros lidos durante o ano, excluindo os livros de bibliografia do curso. Para essa questão, o que procurasse saber, é se essa quantidade de livros lidos pode aumentar o desempenho dos estudantes no resultado final.

O classificador utilizado para a classe QE_I22, foi o MPL. O modelo criado, presente na Figura 32, possui uma concordância regular, segundo a estatística kappa, com 60% de instancias classificadas corretamente.

O filtro de seleção de atributos foi aplicado na tentativa de melhorar o modelo. Foram eliminadas as seguintes variáveis: nt_dis_ce, QP_I2, QP_I6, e QE_I9. O novo modelo está presente na Figura 33.

Com acerto de 65%, o novo modelo ainda não consegue uma boa classificação. E isso indica que o classificador encontrou dificuldade em detectar padrões mais acentuados, nos quais se possam inferir com mais precisão certos comportamentos.

Figura 32: Primeiro modelo gerado para a classe QE_I22.

```
Correctly Classified Instances          24               60      %
Incorrectly Classified Instances        16               40      %
Kappa statistic                          0.4667
Mean absolute error                      0.2205
Root mean squared error                  0.4114
Relative absolute error                 58.7984 %
Root relative squared error             95.0148 %
Total Number of Instances               40

=== Confusion Matrix ===

 a b c d   <-- classified as
 6 2 2 0 | a = A
 2 4 3 1 | b = B
 2 3 5 0 | c = C
 0 0 1 9 | d = E
```

Fonte: Própria.

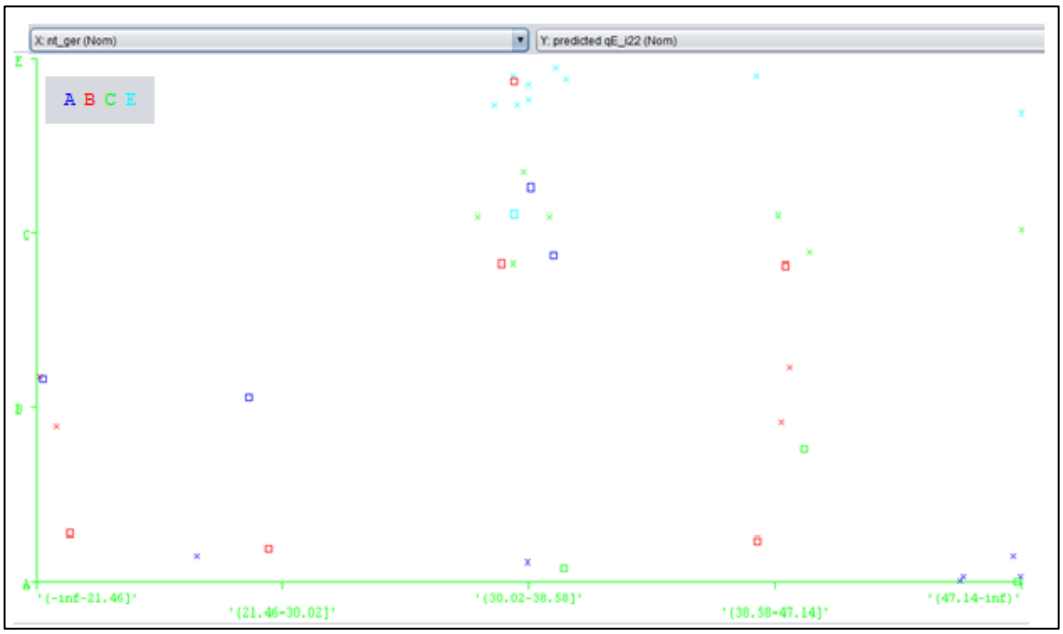
Figura 33: Segundo modelo gerado para a classe QE_I22.

```
Correctly Classified Instances          26               65      %
Incorrectly Classified Instances        14               35      %
Kappa statistic                          0.5333
Mean absolute error                      0.1957
Root mean squared error                  0.3824
Relative absolute error                 52.1987 %
Root relative squared error             88.321  %
Total Number of Instances               40

=== Confusion Matrix ===

 a b c d   <-- classified as
 6 2 2 0 | a = A
 3 4 2 1 | b = B
 2 1 7 0 | c = C
 0 0 1 9 | d = E
```

Fonte: Própria.

De acordo com a visualização da predição presente na Figura 34, não existe um padrão específico que evidencie que a quantidade de livros lido, tem algum impacto negativo na nota final. E por incrível que pareça, as maiores notas estão concentradas nos estudantes que não leram nenhum livro.

Figura 34: Predição para a classe QE_I22.



Fonte: Própria.

Sendo assim, a quantidade de livros lidos não é um fator que contribui para um melhor desempenho na prova.

Outra pergunta que segue o mesmo teor da questão anterior, é a questão QE_I23, sobre a quantidade de horas de estudo. Para essa questão o classificador utilizado foi o *NaiveBayes* foi selecionado para as predições.

O modelo gerado para a classe QE_I23 conseguiu classificar corretamente 64% das instâncias, tendo uma concordância moderada de acordo com o treinamento do classificador.

Contudo não obteve sucesso em classificar a saída B, referente aos alunos que apenas assistiam às aulas. Com a intenção de melhorar a classificação foi aplicado o filtro de seleção de atributos, que fez com que as seguintes variáveis fossem eliminadas: QP_I7, QE_I1, e QE_I9.

O segundo modelo ainda teve dificuldade para a classificação da saída B, porém houve um aumento na classificação das instâncias, passando para 70%, deixando o modelo em uma concordância substancial. Os modelos gerados estão presentes nas Figuras 35 e 36.

Figura 35: Primeiro modelo gerado para a classe QE_I23.

```
Correctly Classified Instances          32               64      %
Incorrectly Classified Instances        18               36      %
Kappa statistic                          0.55
Mean absolute error                      0.1466
Root mean squared error                  0.3677
Relative absolute error                 45.805  %
Root relative squared error             91.9198 %
Total Number of Instances               50

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  2  8  0  0 |  a = B
  0  9  0  1  0 |  b = D
  3  0  6  1  0 |  c = C
  3  0  0  7  0 |  d = E
  0  0  0  0 10 |  e = A
```

Fonte: Própria.

Figura 36: Segundo modelo gerado para a classe QE_I23.

```
Correctly Classified Instances          35               70      %
Incorrectly Classified Instances        15               30      %
Kappa statistic                          0.625
Mean absolute error                      0.1369
Root mean squared error                  0.3472
Relative absolute error                 42.7809 %
Root relative squared error             86.792  %
Total Number of Instances               50

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  2  2  6  0  0 |  a = B
  0  9  0  1  0 |  b = D
  2  0  7  1  0 |  c = C
  3  0  0  7  0 |  d = E
  0  0  0  0 10 |  e = A
```

Fonte: Própria.

As análises partiram por meio do teorema de Bayes aplicado a tabela de frequência gerada pelo classificador. E, por conseguinte pelos valores gerados pela aplicação do teorema. A Figura 37, contém os valores da tabela, e a Figura 38, os resultados.

Figura 37: Tabela gerada pelo Naivebayes para a variável nt_ger (QE_I23).

nt_ger					
'(-inf-21.46]'	1.0	1.0	1.0	1.0	11.0
'(21.46-30.02]'	4.0	1.0	1.0	2.0	1.0
'(30.02-38.58]'	5.0	9.0	6.0	8.0	1.0
'(38.58-47.14]'	3.0	2.0	3.0	1.0	1.0
'(47.14-inf)'	2.0	2.0	4.0	3.0	1.0
[total]	15.0	15.0	15.0	15.0	15.0

Fonte: Própria.

Figura 38: Valores obtidos pelo teorema de Bayes para QE_I23 (nt_ger).

	questão E_I23					
	B	D	C	E	A	TOTAL
intevalo de nota 1	6,666667	6,666667	6,666667	6,666667	73,33333	100
intervalo de nota 2	18,51852	33,33333	14,81481	29,62963	3,703704	100
intervalo de nota 3	22,22222	5,555556	22,22222	44,44444	5,555556	100
intervalo de nota 4	27,27273	18,18182	36,36364	9,090909	9,090909	100
intervalo de nota 5	20	20	20	30	10	100

Fonte: Própria.

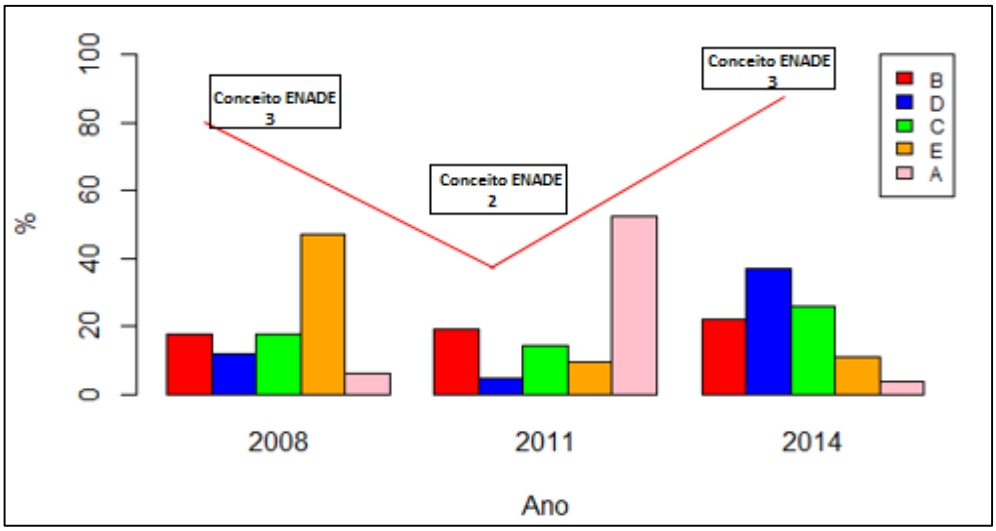
Como resultado da análise não se pode afirmar que a quantidade de horas é um fator forte para influenciar o desempenho do aluno. Não se pode afirmar que quem se dedicou mais de doze horas aos estudos (Letra E), teve um desempenho melhor do que quem estou de uma a três horas (Letra B).

As probabilidades estão distribuídas de modo que não se pode fazer tal afirmação, por esse motivo, a questão das horas de estudos não entra na lista de fatores que podem fazer com que haja um melhor desempenho por parte dos estudantes na nota final, contudo, quem apenas assistiu às aulas, tem uma probabilidade maior de 73,3% de terem os piores desempenhos na prova.

No Gráfico 16, está a aplicação do Teorema de *Bayes* para a variável ano de acordo com a classe QE_I22. O gráfico apresenta que a maior probabilidade de estudantes apenas terem assistidos às aulas está presente no ano de 2011, com quase 60%.

Em 2011, foi quando o conceito ENADE do curso era de 2. No Gráfico 16, ainda consta uma distribuição mais uniforme das respostas para a quantidade de horas de estudo semanalmente no ano de 2014.

Gráfico 16: Teorema de *Bayes* aplicado com relação ao ano (QE_I22).



Fonte: Própria.

Com essa questão, encerra-se o quadro de classes referentes ao questionário socioeconômico do estudante, as próximas pertencem ao questionário de percepção da prova. E o classificador paras as questões seguintes é o *NaiveBayes.*

A pergunta que está em volta das classes QP_I2, e QP_I5, diz respeito à parte específica da prova. Consiste em verificar se o grau de dificuldade, e a clareza e objetividade dos enunciados estão condizentes com suas notas.

Os modelos finais para as classes citadas, estão presentes respectivamente nas Figuras 39 e 40.

Figura 39: Modelo final gerado para a classe QP_I2.

```
Correctly Classified Instances          29               72.5    % === Confusion Matrix ===
Incorrectly Classified Instances        11               27.5    %
Kappa statistic                          0.6333                      a b c d   <-- classified as
Mean absolute error                      0.1385                     6 0 4 0 | a = C
Root mean squared error                  0.3415                     1 9 0 0 | b = E
Relative absolute error                 36.9369 %                   5 0 5 0 | c = D
Root relative squared error             78.8715 %                   0 0 1 9 | d = .
Total Number of Instances               40
```

Fonte: Própria.

Figura 40: Modelo final gerado para a classe QP_I5.

```
Correctly Classified Instances          35               70      % === Confusion Matrix ===
Incorrectly Classified Instances        15               30      %
Kappa statistic                          0.625                        a  b  c  d  e   <-- classified as
Mean absolute error                      0.1238                       6  0  4  0  0 |  a = B
Root mean squared error                  0.299                        2  8  0  0  0 |  b = D
Relative absolute error                 38.6857 %                     4  3  3  0  0 |  c = C
Root relative squared error             74.7431 %                     0  0  0 10  0 |  d = .
Total Number of Instances               50                            1  0  1  0  8 |  e = A
```

Fonte: Própria.

Para a classe QP_I2, foi aplicado o filtro de seleção de atributos, e foi removido as seguintes variáveis: tp_sexo, QP_I2, QP_I9, QE_I1, QE_I9, QE_18, QE_I22, e QE_I23.

Para a classe QP_I5, não foi aplicado o filtro de seleção de atributos, por constatar que a remoção dos atributos prejudicava na predição.

Os modelos das duas classes são bem parecidos, estando na estatística kappa, como sendo de concordância substancial.

Sobre o grau de dificuldade da prova específica (QP_I2), pode-se dizer que sim, estão condizentes, uma vez que, quem marcou alternativas como muito difícil, e difícil, tem uma probabilidade um pouco maior de tirar as piores notas, assim como quem não respondeu.

Mas não existe padrão consistente, uma vez que no intervalo de nota 5 (42,82 -inf), onde estão concentradas as melhores notas, a probabilidade está igualmente distribuída em 25%.

Sobre a objetividade e clareza dos enunciados (QP_I5), há uma aparente contradição, quem marcou que “sim, em todos”, tem as menores chances de ter um desempenho entre as melhores notas, assim como os que não responderam.

Enquanto quem disse “sim, na maioria”, os resultados estão bem distribuídos. A probabilidade maior é 35,7% no intervalo 5 (42,82 -inf), para os que responderam cerca da metade.

Nessa questão o padrão encontrado foi que quem não responde a essa pergunta, ou quem marca a opção como “sim, todos” (letra A), tendem a estar menos concentrados nos intervalos com as melhores notas.

Nas Figuras 41 e 42, estão presentes as tabelas de frequências paras as questões, e nas Figura 43 e 44, os valores da aplicação do teorema de *Bayes*.

Figura 41: Tabela gerada pelo Naivebayes para a variável nt_ce (QP_I2).

nt_ce				
'(-inf-16.78]'	2.0	1.0	3.0	2.0
'(16.78-25.46]'	4.0	9.0	4.0	9.0
'(25.46-34.14]'	4.0	1.0	4.0	1.0
'(34.14-42.82]'	3.0	2.0	2.0	1.0
'(42.82-inf)'	2.0	2.0	2.0	2.0
[total]	15.0	15.0	15.0	15.0

Fonte: Própria.

Figura 42: Tabela gerada pelo Naivebayes para a variável nt_ce (QP_I5)

nt_ce				
'(-inf-16.78]'	2.0	1.0	3.0	2.0
'(16.78-25.46]'	4.0	9.0	4.0	9.0
'(25.46-34.14]'	4.0	1.0	4.0	1.0
'(34.14-42.82]'	3.0	2.0	2.0	1.0
'(42.82-inf)'	2.0	2.0	2.0	2.0
[total]	15.0	15.0	15.0	15.0

Fonte: Própria.

Figura 43: Valores obtidos pelo teorema de Bayes para QP_I2 (nt_ce).

	questão QP_I2				
	C	E	D	.	TOTAL
intevalo de nota 1	25	12,5	37,5	25	100
intervalo de nota 2	15,38462	34,61538	15,38462	34,61538	100
intervalo de nota 3	40	10	40	10	100
intervalo de nota 4	37,5	25	25	12,5	100
intervalo de nota 5	25	25	25	25	100

Fonte: Própria.

Figura 44: Valores obtidos pelo teorema de Bayes para QP_I5 (nt_ce).

	questão QP_I5					
	B	D	C	.	A	TOTAL
intevalo de nota 1	28,57143	14,28571	14,28571	28,57143	14,28571	100
intervalo de nota 2	10,34483	6,896552	13,7931	31,03448	37,93103	100
intervalo de nota 3	16,66667	58,33333	8,333333	8,333333	8,333333	100
intervalo de nota 4	38,46154	15,38462	30,76923	7,692308	7,692308	100
intervalo de nota 5	21,42857	21,42857	35,71429	14,28571	7,142857	100

Fonte: Própria.

A última questão que encerra as classificações é sobre o tempo de permanência para concluir a prova. E a pergunta norte associada a essa questão é: Estudantes que ficaram um tempo maior tentando fazer a resolução da prova, tiveram um aproveitamento melhor?

O modelo criado para a classe QP_I9 teve um fraco desempenho e precisou da aplicação do filtro de seleção de atributos, que removeu doze variáveis. Depois disso o classificador que estava em um nível de concordância regular, passou para o nível de concordância moderada.

Antes a quantidade de instâncias classificadas corretamente era de 48%, depois 66%. Mesmo com a melhora, a distribuição dos acertos da saída ainda não estava equilibrada, das dez saídas da letra C, apenas três tinham sido classificadas corretamente.

Na Figura 45 estão as métricas do modelo para a classe em questão.

Figura 45: Modelo gerado para a classe QP_I9.

```
Correctly Classified Instances          33               66      % === Confusion Matrix ===
Incorrectly Classified Instances        17               34      %
Kappa statistic                          0.575                     a b c d e   <-- classified as
Mean absolute error                      0.1502                    6 2 2 0 0 | a = D
Root mean squared error                  0.3346                    1 6 3 0 0 | b = B
Relative absolute error                 46.9531 %                  1 6 3 0 0 | c = C
Root relative squared error             83.6426 %                  1 0 0 9 0 | d = .
Total Number of Instances               50                         0 0 1 0 9 | e = E
```

Fonte: Própria.

Quando um classificador não consegue distribuir corretamente as classificações entre as saídas da classe, isso afeta diretamente o modelo, e quer dizer que não há padrões consistentes, como aconteceu anteriormente, mas pode haver alguns pontos que podem ser destacados.

Para a classe QP_I9, não foi diferente. Aplicando mais uma vez o teorema de *Bayes*, no que diz respeito a quantidade de horas que os estudantes gastaram para concluir a prova, mais uma vez quem não responde à pergunta tem uma maior probabilidade de tirar as notas mais baixas com 53,3%.

Já quem marca “entre duas e três” e “entre três e quatro” tem a maior probabilidade de ter um desempenho melhor com 35,7%, e 28,6% respectivamente. E isso vale para o intervalo de nota 5 (47,14-inf), já quando o intervalo de nota é o 4

(38,58-47,17], quem assume a maior probabilidade é quem marcou “entre uma e duas horas”.

Então dessa maneira, mais uma vez não tem um padrão que indique que quem fica mais tempo tentando resolver as questões, irá conseguir alcançar as melhores notas, uma vez que os intervalos das “melhores” notas são próximos, e há esse revezamento de quem as consegue alcançar.

Figura 46: Tabela gerada pelo Naivebayes para a variável nt_ger (QP_I9).

nt_ger					
'(-inf-21.46]'	1.0	3.0	2.0	8.0	1.0
'(21.46-30.02]'	2.0	2.0	2.0	3.0	1.0
'(30.02-38.58]'	4.0	4.0	3.0	1.0	10.0
'(38.58-47.14]'	4.0	5.0	3.0	1.0	1.0
'(47.14-inf)'	4.0	1.0	5.0	2.0	2.0
[total]	15.0	15.0	15.0	15.0	15.0

Fonte: Própria.

A tabela de frequência gerada pelo *naivebayes* está presente na Figura 46. Já os valores obtidos com aplicação do teorema de *Bayes* estão na Figura 47.

Figura 47: Valores obtidos pelo teorema de Bayes para QP_I9 (nt_ger).

	questão QP_I9					
	D	B	C	.	E	TOTAL
intevalo de nota 1	6,666667	20	13,33333	53,33333	6,666667	100
intervalo de nota 2	20	20	20	30	10	100
intervalo de nota 3	18,18182	18,18182	13,63636	4,545455	45,45455	100
intervalo de nota 4	28,57143	35,71429	21,42857	7,142857	7,142857	100
intervalo de nota 5	28,57143	7,142857	35,71429	14,28571	14,28571	100

Fonte: Própria.

As análises da etapa da classificação terminam, e com isso o padrão mais acentuado que foi obtido é que estudantes que trabalhavam 40 horas ou mais tiveram um rendimento inferior as demais saídas da classe, tirando as melhores notas os estudantes que não trabalhavam.

Espera-se com essa etapa ter conseguido responder as perguntas em volta dos fatores socioeconômicos que poderiam determinar o desempenho final da prova.

5 CONCLUSÃO

A base de dados retirada do ENADE, possibilitou a extração de informações referentes ao perfil do estudante do curso de Ciência da Computação do Campus Avançado de Natal da UERN, ao desempenho dos estudantes ao longo dos anos analisados, as questões específicas em que os estudantes tiveram um melhor desempenho.

Pertinente ao processo KDD, “a extração do conhecimento”, aconteceu por meio da mineração de dados, diante da aplicação do aprendizado não supervisionado e supervisionado.

Quanto ao perfil socioeconômico do estudante, merece destaque que, a grande maioria dos estudantes são solteiros (82,3%). Contava com 88,2% dos estudantes provenientes de um ensino médio tradicional, mesmo levando em conta a difusão dos institutos federais, que oferecem um ensino técnico.

Com relação à questão étnica racial, apenas 3,9% se consideram negros, e o perfil que traça o estudante ficou divido entre brancos (52,9%) e pardos (41,2%).

Quanto a renda familiar e a situação de trabalho, 58,8% dos estudantes tinham uma renda familiar numa faixa maior que 3 salários mínimos até 10 salários mínimos, e 56,8% dos estudantes trabalhavam 40 horas semanais ou mais.

Sobre a política de cotas, 68,6% dos estudantes disseram não ter utilizado como recurso de ingresso curso, 23,5% responderam que utilizaram por recorte social. Apenas um estudante respondeu que utilizou por critério étnico-racial. 51% dos estudantes cursaram o ensino médio em escola pública, e 43,14% em escola privada (particular).

No quesito desempenho na prova, tudo indica que, mesmo com notas baixas, os estudantes do campus de Natal, estão dentro do cenário nacional. Em 2014, a média dos estudantes, foi de 43,2, na nota final da prova, um pouco abaixo da média nacional de 45,3, no entanto, estando dentro do intervalo modal de (40; 50].

Em 2011, a média da nota final dos estudantes da UERN foi de 29,8, também abaixo da média nacional de 31,8, mas dentro do intervalo modal (20;30]. Já no ano de 2008, os estudantes superaram a média nacional de 34,8, conseguindo a média final de 36,6.

Com relação as demais notas, a realidade dos estudantes do campus de Natal vem acompanhado de perto o cenário nacional. O quadro nacional para nota

discursiva específica, chama atenção pelas notas baixíssimas do curso de ciência da computação, estando no intervalo modal de [0; 10] para os anos analisados, sendo as notas dos estudantes de Natal 16,5, 4,7, 16,8, respectivamente para os anos de 2014, 2011, e 2008.

No que se trata das questões objetivas específicas em que os estudantes tiveram um rendimento maior, as áreas de atuação são: Engenharia de software/interação humano-computador, Banco de dados, Sistemas operacionais e arquitetura de computadores, e Lógica e matemática discreta, com aproveitamento de, 57,8%, 50%, 45,4%, e 35% respectivamente.

Sendo os piores rendimentos em teoria dos grafos, Fundamentos e técnicas de programação, Inteligência artificial e computacional, e Probabilidade e estatística, com 12,8%, 16,8%, 16,8%, e 19,7%, respectivamente.

Contudo os as áreas com os piores rendimentos apresentam apenas questões de complexidades de nível difícil e muito difícil, em contrapartida a área de Engenharia de software/interação humano-computador, contém na sua maioria, questões de nível fácil e médio.

Banco de dados, possui apenas duas apenas duas questões, uma de nível fácil, e uma muito difícil, por isso o resultado final, ficou equilibrado. Já em sistemas operacionais e arquitetura de computadores, também consegue manter o equilíbrio no rendimento por possuir três questões de nível médio, e uma fácil, assim como em Lógica e matemática discreta.

No que tange as correlações entre as vinte e nove variáveis selecionadas para as tarefas descritiva e preditiva, os resultados mostram que a nota do componente específico tem uma forte correlação com a nota geral da prova, com valor de 0,94, evidenciando assim a maior importância para a nota final. Já com relação a nota de formação geral, o valor é de 0,73.

Sem considerar as demais correlações entre as notas, as demais variáveis não tiveram correlações expressivas.

No que concerne ao aprendizado não supervisionado, o *apriori*, algoritmo utilizado, gerou algumas regras que merecem ser destacadas. A primeira é a regra {nt_ce=C} → {net_ger=C}, que reforça o que foi encontrado na correlação das variáveis da nota específica e nota geral, para estudante alcançar uma nota final igual a C, que compreende o intervalo de [41,60], será necessário obter uma nota equivalente na nota específica.

A segunda regra em destaque é a {qe_i15=C} → {qe_I17=A}, onde apresenta que alunos que estudaram o ensino médio todo em escola pública, utilizam cota do recorte social. Ou seja 21,6% dos estudantes, são ambos de escola pública e utilizaram o recorte social, como cota.

Em contrapartida a regra {qE_I17=B} → {qE_I15=A} coloca em evidência que estudantes que estudaram o ensino médio todo em escola privada (particular), estão altamente associados com o fato de não usarem diretamente as cotas para o ingresso na universidade. Ou seja, 35,4% dos estudantes, são ambos de escola privada, e não utilizaram nenhum tipo de cota.

No que se refere à etapa da classificação. Percebe-se que as classes em que foram aplicadas o filtro SMOTE, tiveram um desempenho melhor na maioria dos casos. O balanceamento dos dados foi primordial para melhorar o desempenho das classificações, ainda que não se tenha alcançada predições ditas quase perfeitas.

A aplicação do filtro de seleção de atributos *ClassifierAttributeEval,* teve uma considerável contribuição para a eliminação de variáveis que pouco contribuíam para a classificação, de acordo com o classificador.

O melhor desempenho foi do classificador MPL para a classe QE_10, com 82%, onde visualiza-se que, quem trabalha 40 horas semanais ou mais, tem o rendimento na prova reduzido.

As demais classificações criaram modelos medianos, onde não se obtém padrões consistentes, mas que em alguns casos são retiradas algumas informações relevantes, como no caso da classe QE_I23, onde quem apenas assistiu às aulas, tem uma probabilidade maior de 73,3% de terem os piores desempenhos na prova.

Por conta dos intervalos da renda familiar escolhidos, talvez a classificação para essa classe tinha sido prejudicada. Contudo não se pode “forçar” os dados a revelarem padrões que não existem. Uma vez que todos os procedimentos foram feitos, com intuito de melhorar a classificação.

Em síntese, a classificação foi um processo importante para esse trabalho, por responder às perguntas que estavam sendo levantadas, e por possibilitar conseguir outras informações complementares.

5.1 TRABALHOS FUTUROS

Trabalhos que utilizam a mineração de dados educacionais são muito importantes porque nos dão a possibilidade de tentar responder perguntas, para corrigir e aperfeiçoar determinado sistema educacional.

Com relação a trabalhos futuros, no que diz respeito aos dados da UERN de Ciência da Computação Campus Natal, ainda há muito a ser explorado, tanto no perfil socioeconômico, quanto da na parte didática pedagógica do curso, e com o acréscimo de novos conjuntos de dados provenientes do ENADE de 2017, e dos próximos anos que virão, a consistência para a criação de modelos por parte da técnica de classificação, vai se aperfeiçoando, e com isso padrões vão se acentuando, dando a possibilidade de entender melhor determinados comportamentos.

Quanto à base de dados, o INEP conta com milhares de dados que podem ser explorados. Bases como as do ENADE, do Exame Nacional do Ensino Médio (ENEM), e do Censo Escolar, podem ser utilizadas para entender cenários no contexto educacional como, evasão escolar, e desempenho em áreas de conhecimento do ensino médio.

Como trabalhos futuros, pretende-se utilizar a mesma metodologia deste trabalho, aplicada aos estudantes do mesmo curso, no entanto, sendo os da sede da UERN, que fica na cidade de Mossoró, tendo como objetivo fazer uma comparação de desempenho entre os estudantes de Natal e Mossoró. Outro trabalho em mente, segue a mesma linha de pensamento, no entanto, para maiores quantidades de dados, que no caso seria, utilizando novamente os dados do ENADE para comparar o desempenho dos estudantes do curso de ciência da computação das universidades federais das regiões do Brasil, sofrendo a metodologia modificações para atender às necessidades de se trabalhar com uma maior quantidade de dados.

REFERÊNCIAS

ALMEIDA, Leandro et al. DEMOCRATIZAÇÃO DO ACESSO E DO SUCESSO NO ENSINO SUPERIOR:UMA REFLEXÃO A PARTIR DAS REALIDADES DE PORTUGAL E DO BRASIL. **Avaliação**: Revista da Avaliação da Educação Superior, Campinas, v. 17, n. 3, p.899-920, nov. 2012. Disponível em: <http://submission.scielo.br/index.php/aval/article/view/78445/7579>. Acesso em: 20 jun. 2018.

ALI, A. B. M. Shawkat. Performance Analysis of Statistical Classifier SMO with other Data Mining Classifiers. In: BENITEZ, Jose et al. **Advances in Soft Computing: Engineering Design and Manufacturing.** Londres: Springe, 2003. p. 205-2012.

ALPAYDM, Ethem. **Introduction to Machine Learning.** 2. ed. Londres: The Mit Press, 2010.

AMO, Sandra de. **Técnicas de Mineração de Dados.** Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 15 jun. 2018.

AYODELE, Taiwo Oladipupo. Types of Machine Learning Algorithms. In: GUAN, Yizhang. **New Advances in Machine Learning.** Londres: Intechopen, 2010. Cap. 3. p. 19-48.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, [s.l.], v. 19, n. 02, p.3-13, 31 ago. 2011. Comissão Especial de Informática na Educação.

BAKER, Ryan S.j.d.; YACEF, Kalina. The State of Educational Data Mining in 2009: A Review and Future Visions. **Journal Of Educational Data Mining.** [s. L.], p. 3-16. Fall. 2009.

BAN, Jung-chao; CHANG, Chih-hung. The learning problem of multi-layer neural networks. **Neural Networks**, [s.l.], v. 46, p.116-123, out. 2013.

BARBOSA, Glauber de Castro; FREIRE, Fátima de Souza; CRISÓSTOMO, Vicente Lima. ANÁLISE DOS INDICADORES DE GESTÃO DAS IFES E O DESEMPENHO DISCENTE NO ENADE. **Avaliação**: Revista da avaliação da educação superior, Campinas, v. 16, n. 2, p.317-344, jul. 2011. Disponível em: <http://www.scielo.br/pdf/aval/v16n2/a05v16n2.pdf>. Acesso em: 2 jun. 2018.

BEZERRA, Miguel Eugênio Ramalho. **MÉTODOS BASEADOS NA REGRA DO VIZINHO MAIS PRÓXIMO PARA RECONHECIMENTO DE IMAGENS.** 2006. 56 f. TCC (Graduação) - Curso de Engenharia da Computação, Escola Politécnica de Pernambuco – Universidade de Pernambuco, Recife, 2006. Disponível em: <https://tcc.ecomp.poli.br/20052/MiguelEugenio.pdf>. Acesso em: 10 jun. 2018.

BRAZILIAN SYMPOSIUM ON COMPUTER GAMES AND DIGITAL ENTERTAINMENT, 7., 2008, Belo Horizonte. **GeoplanoPEC:** Um Jogo Inteligente Para o Ensino de Geometria Plana. Belo Horizonte: Sociedade Brasileira de

Computação – Sbc, 2008. 175 p. Disponível em: <http://www.sbgames.org/papers/sbgames08/Proceedings-SBGames-Computing-2008-Final-CD.pdf>. Acesso em: 16 jun. 2018.

CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 3., 2014, Dourados – Ms. **Prática de Mineração de Dados no Exame Nacional do Ensino Médio.** Dourados – Ms: Sociedade Brasileira de Computação – Sbc, 2014. 660 p. Disponível em: <http://www.br-ie.org/pub/index.php/wcbie/article/view/3289/2827>. Acesso em: 22 maio 2018.

CRETTON, Nícollas Nogueira; GOMES, Georgia Rodrigues. APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO ENADE COM ENFOQUE NOS CURSOS DE MEDICINA. **Acta Biomédica Brasiliensia**, [s.l.], v. 7, n. 1, p.74-88, 20 jun. 2016. Universidade Iguacu - Campus V.

DOMINGOS, Pedro. A few useful things to know about machine learning. **Communications Of The Acm**, [s.l.], v. 55, n. 10, p.78-87, 1 Fall. 2012. Association for Computing Machinery (ACM).

EAVES, David. **The Three Laws of Open Government Data.** 2009. Disponível em: <https://eaves.ca/2009/09/30/three-law-of-open-government-data/>. Acesso em: 25 maio 2018.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **Ai Magazine.** Palo Alto, California, p. 37-54. Outono. 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>. Acesso em: 11 jun. 2018.

FERNANDES, Susana; PINTO, Mónica. **Afinal o que são e como se calculam os quartis?** Disponível em: <https://sapientia.ualg.pt/bitstream/10400.1/2963/1/SFernandes_MMPinto_quartis_no_ensino.pdf>. Acesso em: 22 jun. 2018.

FGV. **Índice de Dados Abertos Para o Brasil.** 2017. Disponível em: <http://dapp.fgv.br/dapp-e-open-knowledge-lancam-indice-de-dados-abertos-para-o-brasil/>. Acesso em: 22 jun. 2018.

FRANK, Eibe. **Machine Learning Techniques for Data Mining.** 2000. Disponível em: <http://informatique.umons.ac.be/ssi/teaching/dwdm/ML_part_I.pdf>. Acesso em: 11 jun. 2018.

FRANK, Eibe; HALL, Mark A.; WITTEN, Ian H.. **The WEKA Workbench.** 2016. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf>. Acesso em: 2 jun. 2018.

FREIRE, Fátima de Souza; CRISÓSTOMO, Vicente Lima; CASTRO, Juscelino Emanoel Gomes de. Análise do desempenho acadêmico e indicadores de gestão

das IFES. **Revista Produção Online**, [s.l.], v. 7, n. 4, p.1-25, 5 jul. 2008. Associacao Brasileira de Engenharia de Producao – ABEPRO

FUNAHASHI, Ken-ichi. On the Approximate Realization of Continuous Mappings by Neural Networks. **Neural Networks.** [s. L.], p. 183-192. set. 1988.
GAMA, João et al. **Extração de conhecimento de dados: data mining.** 2. ed. Lisboa: Sílabo, 2015. 428 p.

GARDNER, M. W.; DORLING, S. R.. Artificial Neural Networks (The Multilayer Perceptron): A Review of Applications in The Atmospheric Sciences. **Atmopheric Enviromment.** [s. L.], p. 2627-2636. jun. 1998.

GARNER, Stephen R. **WEKA: The Waikato Environment for Knowledge Analysis.** Disponível em: <https://www.cs.waikato.ac.nz/~ml/publications/1995/Garner95-WEKA.pdf>. Acesso em: 2 jun. 2018.

GIRARDELLO, Adriano Douglas. **Um Estudo sobre o Uso de Máquinas de Vetores de Suporte em Problemas de Classificação.** 2010. 58 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Estadual do Oeste do Paraná, Cascavel, 2010. Disponível em: <http://www.inf.unioeste.br/~tcc/2010/TCC-Adriano Douglas Girardello.pdf>. Acesso em: 10 jun. 2018.

GOYAL, Anshul; MEHTA, Rajni. Performance Comparison of Naïve Bayes and J48 Classification Algorithms. **International Journal Of Applied Engineering Research.** [s. L.], p. 1389-1393. 2012.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques. 3. ed. Estados Unidos: Elsevier, 2012.

HSSINA, Badr et al. A comparative study of decision tree ID3 and C4.5. **International Journal Of Advanced Computer Science And Applications**, [s.l.], v. 4, n. 2, p.13-19, 2014. The Science and Information Organization.

HUANG, Fenghua; YAN, Luming. Combined Kernel-Based BDT-SMO Classification of Hyperspectral Fused Images. **The Scientific World Journal**, [s.l.], v. 2014, n. 2014, p.1-13, 27 ago. 2014. Hindawi Limited. http://dx.doi.org/10.1155/2014/738250. Disponível em: <https://www.hindawi.com/journals/tswj/2014/738250/>. Acesso em: 15 jun. 2018.

JUNG, Carlos Fernando. **Metodologia Científica e Tecnológica:** Módulo 3 – Variáveis e Constantes. 2009. Disponível em: <http://www.dsce.fee.unicamp.br/~antenor/mod3.pdf>. Acesso em: 15 jun. 2018.
KAUR, Gaganjot; CHHABRA. Improved J48 Classification Algorithm for the Prediction of Diabetes. **International Journal Of Computer Applications.** [s. L.], p. 13-17. jul. 2014.

LAROSE, Daniel T.. **Discovering Knowledge in Data:** An Introduction to Data Mining. New Jersey: Wiley Interscience, 2005.

LEUNG, K. Ming. **Naive Bayesian Classifier.** 2007. Disponível em: <http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf>. Acesso em: 1 jun. 2018.

LIEW, Anthony. DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. **Business Management Dynamics: Double Blind Peer Reviewed - Open Access Journal.** Inglaterra, p. 49-62. abr. 2013. Disponível em: <http://bmdynamics.com/issue_pdf/bmd110349- 49-62.pdf>. Acesso em: 15 jun. 2018.

LIEW, Anthony. **Understanding Data, Information, Knowledge And Their Inter-Relationships.** 2007. Disponível em: <http://www.tlainc.com/articl134.htm>. Acesso em: 11 jun. 2018.

MALAMUD, Carl. **Open Government Working Group Meeting in Sebastopol.** [mensagem pessoal] Mensagem recebida por: <Attendees>. em: 22 out. 2007

MITCHELL, Tom M.. **Machine Learning.** Portland, Oregon: McGraw Hill, 1997. 432 p.

INEP. **ENADE 2008:** Relatório Síntese - Computação. [s.l.]: Mec, 2008. Disponível em: < http://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2008/2008_rel_sint_computacao_informatica.pdf>. Acesso em: 21 jun. 2018.

INEP. **ENADE 2011:** Relatório Síntese - Computação. [s.l.]: Mec, 2011. Disponível em: < http://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2011/2011_rel_computacao.pdf >. Acesso em: 21 jun. 2018.

INEP. **ENADE 2014:** Relatório de Área - Ciência da Computação. [s.l.]: Mec, 2016. Disponível em: <http://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2014/2014_rel_ciencia_da_computacao.pdf>. Acesso em: 21 jun. 2018.

INEP. **Resumo Técnico:** Censo da Educação Superior 2014. Brasília: Mec, 2016. 55 p. Disponível em: <http://portal.inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/636024>. Acesso em: 21 jun. 2018.

NOGUEIRA, Eduardo Dimas Andrino; TSUNODA, Denise Fukumi. Mineração de dados para análise de relação entre as características socioeconômicas de concluintes do ensino superior e o desempenho desses estudantes no Enade 2012. **Percurso**, Curitiba, v. 16, n. 1, p.245-268. 2015. Quadrimestral.

NOVAES, Ivan Luiz; SALES, Kathia Marise Borges. **Boletim ENADE 2015.** 2015. Disponível em: <http://www.uneb.br/files/2016/01/Boletim-ENADE-2015.pdf>. Acesso em: 20 jun. 2018.

PLATT, John C.. **Sequential Minimal Optimization:** A Fast Algorithm for Training Support Vector Machines. [s. L.]: Microsoft Research, 1998. 21 p. Disponível em: <https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/>. Acesso em: 10 jun. 2018.

PROLO, Carlos Augusto; HESSEL, Fabiano Passuelo; SAYÃO, Miriam. **ENADE Comentado 2008:** Computação. Porto Alegre: Edipucrs, 2009. 184 p. Disponível em: <http://www.pucrs.br/edipucrs/enade/computacao2008.pdf>. Acesso em: 21 jun. 2018

RISTOFF, Dilvo. O novo perfil do campus brasileiro: uma análise do perfil socioeconômico do estudante de graduação.. **Avaliação**: Revista da Avaliação da Educação Superior, Campinas, v. 19, n. 3, p.723-747, nov. 2014

RISTOFF, Dilvo; LIMANA, Amir. **O Enade como parte da avaliação da educação superior.** Disponível em: <http://3em.ubi.pt/o_enade.pdf>. Acesso em: 25 maio 2018.

ROMÃO, Wesley et al. **EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO EM C&T:** O ALGORITMO APRIORI. 1999. Disponível em: <http://www.abepro.org.br/biblioteca/ENEGEP1999_A0901.PDF>. Acesso em: 25 maio 2018.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews**: Data Mining and Knowledge Discovery, [s.l.], v. 3, n. 1, p.12-27, 14 dez. 2012. Wiley.

SAKHARE, Nitin Nandkumar; JOSHI, Swati Atul. Classification of Criminal Data Using J48-Decision Tree Algorithm. **International Journal Of Datawarehousing And Mining.** [s. L.], p. 167-171. ago. 2014. Disponível em: <https://www.researchgate.net/publication/265908026_Classification_of_Criminal_data_using_J48_Algorithm?enrichId=rgreq-8f007a691cd2ba23b86dfea73638aa47-XXX&enrichSource=Y292ZXJQYWdlOzI2NTkwODAyNjtBUzoxOTA0Njk3OTk2Nzc5NTJAMTQyMjQyMjk0NzU0Nw==&el=1_x_2&_esc=publicationCoverPdf>. Acesso em: 12 jun. 2018.

SHAFIQUE, Umair; QAISER, Haseeb. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). **International Journal Of Innovation And Scientific Research.** [s.l.], p. 217-222. nov. 2014. Disponível em: <http://www.ijisr.issr-journals.org/abstract.php?article=IJISR-14-281-04>. Acesso em: 19 jun. 2018.

SILVA, Luís M.; SÁ, J Marques de; ALEXANDRE, Luís A.. Data classification with multilayer perceptrons using a generalized error function. **Neural Networks.** [s. L.], p. 1302-1310. abr. 2008.

SUMATHI, S.; SIVANANDAM, S. N.. **Introduction to Data Mining ad Its Applications.** Berlin: Springer, 2006.

SUTTON, Oliver. **Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction.** 2012. Disponível em: <http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf>. Acesso em: 9 jun. 2018.

TOURETZKY, David S.; POMERLEAU, Dean A.. What's Hidden in the Hidden Layers? **Neural Networks.** [s. L.], p. 227-233. set. 1989. Disponível em: <https://www.cs.cmu.edu/~dst/pubs/byte-hiddenlayer-1989.pdf>. Acesso em: 2 jun. 2018.

VASCONCELOS, Lívia Maria Rocha de; CARVALHO, Cedric Luiz de. **Aplicação de Regras de Associação para Mineração de Dados na Web.** Goiás: Instituto de Informática Universidade Federal de Goiás, 2004. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf>. Acesso em: 25 maio 2018.

VIERA, Anthony J.; GARRETT, Joanne M.. Understanding Interobserver Agreement: The Kappa Statistic. **Family Medicine.** [s. L.], p. 360-363. maio 2005. Disponível em: <http://www.stfm.org/Portals/49/Documents/FMPDF/FamilyMedicineVol37Issue5Viera360.pdf>. Acesso em: 3 jun. 2018.

VIJAYARANI, Ms S.; MUTHULAKSHMI, Ms M. Comparative Analysis of Bayes and Lazy Classification Algorithms. **International Journal Of Advanced Research In Computer And Communication Engineering.** [s. L.], p. 3118-3124. ago. 2013
WRITTEN, Ian H.; FRANK, Eibe. **Data Mining:** Practical Machine Learning Tools and Techniques. 2. ed. São Francisco, California: Elsevier, 2005.

WITTEN, Ian H. et al. **Weka:** Practical machine learning tools and techniques with Java implementations. 1999. Disponível em: <https://researchcommons.waikato.ac.nz/handle/10289/1040>. Acesso em: 2 jun. 2018.

W3C (Brasil). **Dados Abertos Governamentais.** Disponível em: <http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf>. Acesso em: 22 jun. 2018.

W3C (Brasil). **Manual dos Dados Abertos:** Governo. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 22 jun. 2018.

YAMAGUTI, Marcelo Hideki; BACELO, Ana Paula Terra. **ENADE Comentado:** Computação 2011. Porto Alegre: Edipucrs, 2014. 80 p. Disponível em: <http://ebooks.pucrs.br/edipucrs/Ebooks/Pdf/978-85-397-0521-4.pdf>. Acesso em: 21 jun. 2018.

YU-WEI; CHIU, David. **R for Data Science Cookbook.** Birmingham: Packt Publishing, 2016. 452 p.

.