UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO SOBRE OS DADOS SOCIOECONÔMICOS NA EDUCAÇÃO NA BASE DE DADOS DO INEP

AUREA MILENE TEIXEIRA BARBOSA DOS SANTOS

DM 13/2019

UFPA/ITEC/PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2019

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AUREA MILENE TEIXEIRA BARBOSA DOS SANTOS

MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO SOBRE OS DADOS SOCIOECONÔMICOS NA EDUCAÇÃO NA BASE DE DADOS DO INEP

DM 13/2019

UFPA/ITEC/PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2019

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AUREA MILENE TEIXEIRA BARBOSA DOS SANTOS

MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO SOBRE OS DADOS SOCIOECONÔMICOS NA EDUCAÇÃO NA BASE DE DADOS DO INEP

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica para obtenção do Grau de Mestre em Engenharia Elétrica com ênfase em Computação Aplicada sob a orientação do Prof. Dr. Marcelino Silva da Silva.

UFPA/ITEC/PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2019

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S237m Santos, Aurea Milene Teixeira Barbosa dos
MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO SOBRE OS DADOS SOCIOECONÔMICOS NA EDUCAÇÃO NA BASE DE DADOS DO INEP / Aurea Milene Teixeira Barbosa dos Santos. — 2019.
80 f. : il. color.

Orientador(a): Prof. Dr. Marcelino Silva da Silva
Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Elétrica, Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2019.

1. Mineração de dados educacionais. 2. Redes Bayesianas. 3. Enem. 4. Censo Escolar. I. Título.

CDD 006.312

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"MINERAÇÃO DE DADOS EDUCACIONAIS: UM ESTUDO SOBRE INDICADORES SOCIOECONÔMICOS EM EDUCAÇÃO NO BANCO DE DADOS DO INEP"

AUTORA: **AUREA MILENE TEIXEIRA BARBOSA DOS SANTOS**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRA EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 09/04/2019

BANCA EXAMINADORA:

Prof. Dr. Marcelino Silva da Silva
(Orientador – PPGEE/UFPA)

Prof. Dr. Claudomiro de Souza de Sales Junior
(Avaliador Interno – PPGEE/UFPA)

Prof. Dr. Edvar da Luz Oliveira
(Avaliador Externo – UFRA)

VISTO:

Prof.ª Dr.ª Maria Emília de Lima Tostes
(Coordenadora do PPGEE/ITEC/UFPA)

Sou grato a Deus, que me ajudou em cada etapa desse trabalho e não me deixou fraquejar, guiando os meus passos!

Dedico o presente trabalho a minha mãe e ao meu pai, que fizeram de tudo para que eu trilhasse os caminhos acadêmicos!

Aos amigos e colegas, que me incentivaram todos os dias e ofereceram apoio nos momentos críticos.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela vida, por iluminar a minha mente, guiar meus passos, guardar-me do mal e por me abençoar em todos os momentos até aqui.

Sou grata aos meus pais, pelo amor, esforço e dedicação. O maior tesouro que me deram foi a educação, essa caminhada árdua não me impediu de sonhar e conquistar muitas mudanças em nossas vidas. Agradeço, em especial, à minha mãe Aureni de Nazaré Teixeira que incontáveis vezes demonstrou na prática como é grande o seu amor, mostrando o verdadeiro valor da amizade e lealdade. Sempre me incentivando e acreditando em cada passo que decidi tomar nessa vida.

Ao meu irmão Aldri Mateus dos Santos, pessoa pela qual eu tenho muito orgulho e admiração.

Dedico meus agradecimentos especialmente, ao meu orientador Prof. Dr. Marcelino Silva da Silva, por toda paciência inigualável, orientação, aprendizado e pela grande oportunidade em realizar este trabalho, ao ter me acolhido e acreditado em mim como orientanda. Por todos os conselhos, puxões de orelha quando necessário e por toda amizade estabelecida nesta caminhada. Um grande exemplo de pesquisador e ser humano a ser seguido, muito obrigada por tudo professor.

Às pessoas que participaram e participam da minha vida e com o tempo me fizeram ser quem eu sou, auxiliando a evoluir as minhas ideias, as minhas aspirações, os meus objetivos e o meu futuro. Aos meus colegas do Laboratório de Pesquisa Operacional - LPO e Laboratório de Planejamento de Redes de Alto Desempenho – LPRAD da Universidade Federal do Pará, que de alguma forma contribuíram para este trabalho, em especial aos meus amigos: Jonatas Paulino, Priscila Aranha, Mariane Gonçalves, Anderson Souto, Rodrigo Alfaia, Maria da Penha e Eulália da Mata. Pessoas que me deram apoio em todos os desafios que enfrentei nessa jornada.

Ao CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico e CAPES - Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pelo suporte financeiro.

E por fim, mas não menos importante agradeço à Universidade Federal do Pará pela oportunidade dada a mim para a realização desta dissertação. Aos professores e funcionários do Programa de Pós-Graduação em Engenharia Elétrica pela oportunidade, aprendizado e todo apoio.

"Eu andarei vestido e armado com as armas de Jorge para que meus inimigos, tendo pés não me alcancem, tendo mãos não me peguem, tendo olhos não me vejam, e nem em pensamentos eles possam me fazer mal"

(Oração de São Jorge)

RESUMO

Este trabalho investiga os perfis dos alunos do terceiro ano do ensino médio brasileiro da rede escolar pública e privada, a fim de identificar quais fatores extraescolares e interescolares influenciam para que o estudante possa vim a ter um desempenho consideravelmente bom no Exame Nacional do Ensino Médio (ENEM). Dessa forma, foram realizados dois estudos de caso, um com os registros socioeconômicos contendo dezenas de milhares de amostras oriundas de alunos que realizaram o exame, dividido por cada região brasileira, possibilitando uma análise dos fatores socioeconômicos (extraescolares) influentes em cada região. E o outro estudo analisou os atributos relacionados às condições de infraestrutura escolar ofertada pelas escolas públicas (estaduais) de ensino médio no estado do Pará, para esse estudo foi relacionada cada nota que o estudante conseguiu no exame do Enem com a base do censo escolar, ou seja, esta base detalha as condições das escolas secundárias correspondente a cada aluno que participou da prova do Enem ambos de 2016. Para alcançar o objetivo proposto, os dois estudos de casos foram submetidos ao processo de Descoberta de Conhecimento em Base de Dados, a mineração de dados educacionais (MDE). No processo da MDE foi empregada a técnica de análise de componentes principais (PCA) na etapa do pré-processamento, com o intuito de diminuir a quantidade de variáveis sem perder as informações fornecidas pelo conjunto total, utilizando essa técnica foi possível diminuir de 43 para 22 o número de variáveis analisadas no estudo de caso um, e de 39 para 9 no segundo estudo de caso, com o percentual de 0.8226 % e 0.9099 % respectivamente. Tal técnica foi utilizada para propiciar a execução de mais outra técnica aplicada na pesquisa, as Redes Bayesianas, fazendo uso do *software Bayesware Discoverer*, sendo utilizada na etapa de mineração de dados, a escolha por essa técnica se deu por ela possibilitar o raciocínio sobre incertezas, especialmente em diagnósticos de causas e efeitos tendo como pressuposto o relacionamento das variáveis e suas probabilidades de ocorrências. Outro aspecto inerente é a sua estrutura que diz respeito à compreensibilidade da representação e dos resultados, os quais geram subsídios voltados para que especialistas e usuários inseridos no domínio realizem análises mais aprofundadas sobre o assunto tratado pelos dados. Os resultados atingidos apontaram o sucesso dessa metodologia e as técnicas empregadas, a pesquisa nos possibilitou ter uma análise a nível nacional dos estudantes do terceiro ano do ensino médio brasileiro, onde nenhum estudo realiza essa análise a nível Brasil se tratando dos dados do Enem. Foram apontadas fortes influências de variáveis socioeconômicas destacando como fatores influentes diretos no desempenho dos estudantes a diferença se ele estudou em escolas pública, privada ou federal. Aliada a essa variável encontra-se a questão da renda familiar, se o estudante abandonou ou reprovou no ensino fundamental, se tem acesso a computador e internet na sua residência e o turno em que estudou no ensino médio, a partir dessas variáveis foi possível realizar inferências e analisar o comportamento probabilístico das notas obtidas pelo aluno com cada uma dessas variáveis. Se tratando da análise da influência da estrutura escolar no desempenho do estudante paraense da escola pública, destacou-se as variáveis biblioteca e laboratório de ciências. Ao analisar só o estado do Pará se verificou que mais de 80% dos alunos da rede pública tiveram um desempenho ruim tirando notas iguais ou menores que 450 no Enem, mesmo que em sua escola tenha as duas variáveis aponta como influentes.

Palavras chaves: Mineração de dados educacionais; Redes Bayesianas; Enem; Censo Escolar.

ABSTRACT

This work investigates the profiles of the third year students of the Brazilian high school of the public and private school network, in order to identify which extracurricular and interscholastic factors influence the student to have a good performance in the National High School Examination (ENEM). In this way, two case studies were carried out, one with the socioeconomic records containing tens of thousands of samples from the students who took the exam, divided by each Brazilian region, making possible an analysis of the influential socioeconomic (extra-school) factors in each region. And the other study analyzed the attributes related to the conditions of school infrastructure offered by public (state) high schools in the state of Pará, for this study was related each note that the student obtained in the examination of enem with the base of the school census, that is, this database details the conditions of the secondary schools corresponding to each student who participated in the test of the enem both 2016. In order to reach the proposed objective, the two case studies were submitted to the process of Knowledge Discovery in Database, the educational data mining (EDM). In the MDE process, the main component analysis (PCA) technique was used in the preprocessing stage, in order to reduce the number of variables without losing the information provided by the total set, using this technique it was possible to decrease from 43 to 22 the number of variables analyzed in case study one, and 39 to 9 in the second case study, with a percentage of 0.8226% and 0.9099% respectively. This technique was used to propitiate the execution of another technique applied in the research, the Bayesian Networks, being used in the data mining stage, the choice for this technique was made possible by it to reason about uncertainties, especially in causes and effects having as presupposition the relationship of the variables and their probabilities of occurrences. Another inherent aspect is its structure, which concerns the comprehensibility of representation and results, which generate subsidies aimed at allowing specialists and users in the field to carry out more in-depth analysis on the subject treated by the data. The results showed the success of this methodology and the techniques used, the research allowed us to have a national analysis of the students of the third year of high school in Brazil, where no study performs this analysis at the Brazilian level when dealing with enem data. Strong influences of socioeconomic variables were pointed out highlighting as direct influential factors in student performance the difference if he studied in public, private or federal schools. Allied to this variable is the question of family income, if the student left or failed in elementary school, if he has access to the computer and internet in his residence and the shift in which he studied in high school, from these variables it was possible to perform inferences and analyze the probabilistic behavior of the grades obtained by the student with each one of these variables. When analyzing the influence of the school structure on the performance of the Paraense student of the public school, the variables library and science laboratory were highlighted. When analyzing only the state of Pará, it was verified that more than 80% of the students in the public network performed poorly, taking notes equal to or less than 450 in the enem, even though in their school the two variables were considered as influential.

Keywords: *Educational data mining; Bayesian Networks; Enem; School Census.*

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ENEM - Exame Nacional do Ensino Médio

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

PCA - Principal Component Analysis

PISA - Programme for International Student Assessment

OECD - Organisation for Economic Co-operation and Development

EDM - Educational Data Mining

MDE - Mineração de Dados Educacionais

W3C - World Wide Web Consortium

ISO - International Organization for Standardization

MD - Mineração de Dados

RB - Redes Bayesianas

TPC - Tabela de Probabilidade Condicional

PROUNI - Programa Universidade para Todos

FIES - Financiamento Estudantil

EJA - Educação de Jovens e Adultos

FUNDEB - Fundo de Manutenção e Desenvolvimento da Educação Básica e de valorização dos Profissionais da Educação

CSV - Comma-separated values

SEEC - Serviço de Estatística de Educação e Cultura

AVA - Ambiente Virtual de Aprendizagem

EAD - Educação a Distância

KDD - Knowledge Discovery in Databases

MEC - Ministério da Educação

MB - Muito Baixa

BA - Baixa

REG - Regular

EXC - Excelente

CN - Ciências da Natureza

CH - Ciências Humanas

LC - Linguagens e Codigos

MT - Matemática

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

1. INTRODUÇÃO

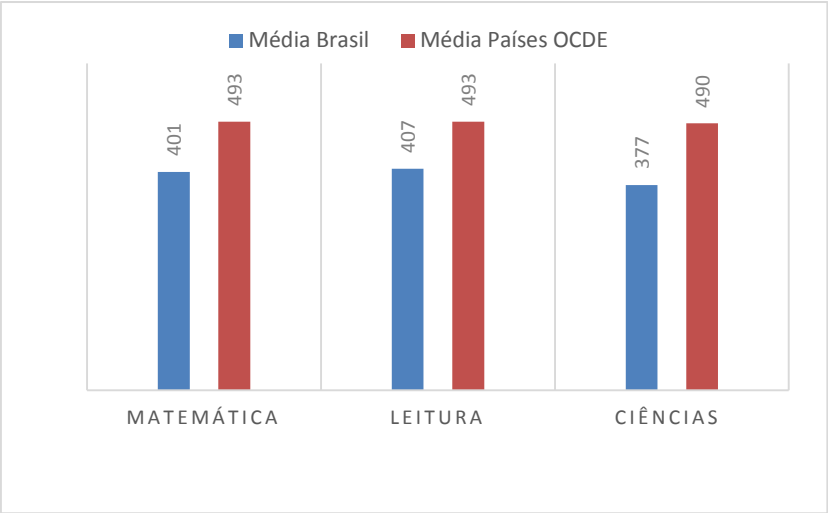
1.1. Motivação e Caracterização do Problema

O ensino médio se constitui em uma etapa da educação básica brasileira que tem mobilizado controversos debates em torno de questões como os persistentes problemas de acesso, a ausência de qualidade, a falta de sentidos e objetivos claros, o aparente descompasso entre suas práticas e os interesses de seu público, formado principalmente por jovens (Krawczyk, 2009 e 2011).

A educação possui um papel extremamente importante fornecendo às pessoas o conhecimento, as habilidades e as competências necessárias para uma participação efetiva na sociedade e na economia. Além disso, a educação pode melhorar a vida das pessoas em áreas como saúde, engajamento cívico, interesse político e felicidade. Estudos demonstram que pessoas instruídas vivem mais, participam mais ativamente da política e da comunidade onde vivem, cometem menos crimes e necessitam menos de assistência social (OCDE, 2016).

Ainda segundo a OCDE (2016), no Brasil, 49% dos adultos entre 25 e 64 anos concluíram o ensino médio, muito abaixo da média da OCDE de 74%. Em 2015, o PISA testou estudantes de 72 países, incluindo países da OCDE, Brasil e Federação Russa. Os alunos foram testados em sua capacidade de leitura, suas habilidades em matemática e nível em ciências. O aluno médio brasileiro teve um desempenho abaixo da média dos alunos em países avaliados, com médias de notas em matemática (401 pontos), leitura (407 pontos) ciências (377 pontos), quanto à média dos alunos de outros países da OCDE são respectivamente 493, 493 e 490 pontos, como ilustra o gráfico 1.

Gráfico 1: Gráfico comparando o desempenho médio brasileiro em relação aos países da OCDE.



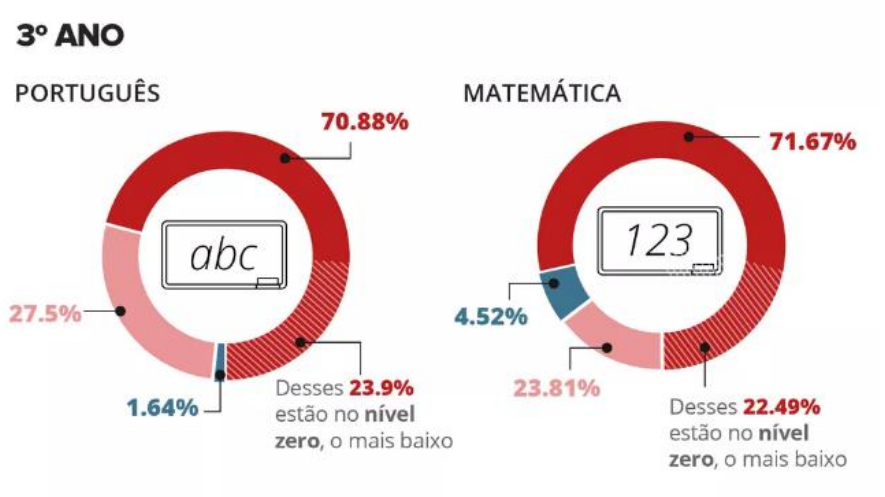
Fonte: Adaptação dos dados da OCDE, 2016.

Segundo o Inep em 2016, essas médias não apresentaram melhora em relação aos últimos anos. A média brasileira de ciências tem se mantido estável desde 2006, e a de leitura, desde 2000. A média de matemática apresentou crescimento significativo de 21 pontos desde 2003, porém diminuiu 11 pontos entre 2012 e 2015.

Analisando o ensino médio brasileiro com base nos dados da nota do Sistema de Avaliação da Educação Básica (Saeb), sistema que é utilizado pelo governo federal, a cada dois anos, para medir a aprendizagem dos alunos ao fim de cada etapa de ensino: ao 5º e 9º anos do ensino fundamental e 3º ano do ensino médio. O sistema é composto pelas médias de proficiências em português e matemática. Em 2017, cerca de 70% dos estudantes que concluíram o ensino médio no país apresentaram resultados considerados insuficientes em matemática. A mesma porcentagem não aprendeu nem mesmo o considerado básico em português, segundo dados do (Saeb, 2017).

O Ministério da Educação (MEC) classificou os níveis de proficiência que estão organizados em uma escala de 0 a 9 - quanto menor o número, pior o resultado. Níveis de 0 a 3 são considerados insuficientes; entre 4 e 6 os alunos têm nível de conhecimento básico; e a partir de 7 até 9, adequado. Etapa mais problemática da educação básica, o ensino médio foi classificado no nível 2 de proficiência, a figura 1 ilustra esses percentuais e considera os índices totais somando escolas da rede públicas e privada.

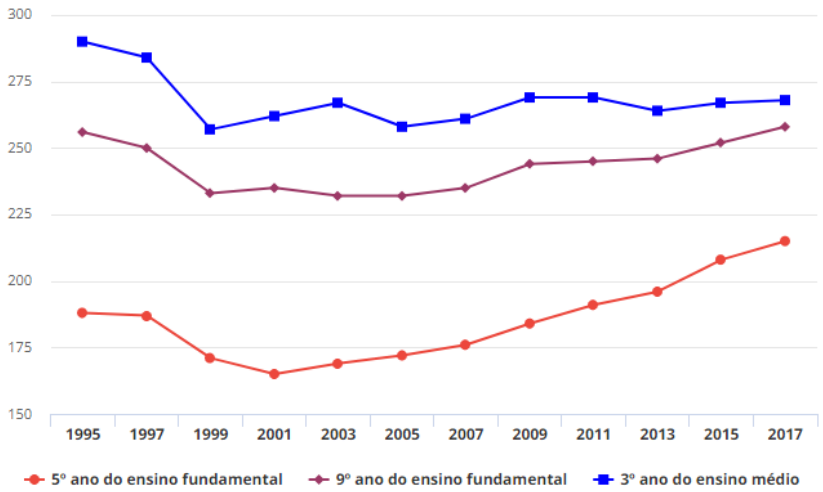
Figura 1: Gráfico dos percentuais em nível de proficiência e português e matemática, dos alunos do 3º ano do ensino médio brasileiro.



Fonte: Jornal G1, adaptado dos dados do Inep/Mec, 2018.

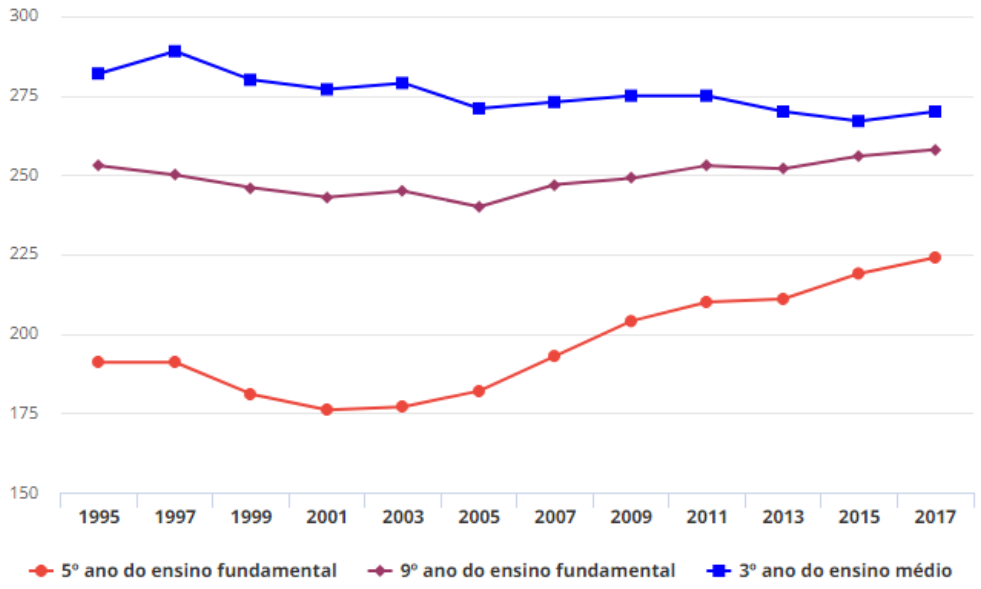
Podemos observar que em matemática, 71,67% dos alunos têm nível insuficiente de aprendizado. Desses, 23% estão no nível 0, o mais baixo da escala de proficiência organizada pelo MEC. Na matéria de língua portuguesa, 70,88% dos alunos têm nível insuficiente de aprendizado, sendo que 23,9% estão no nível zero, o nível mais baixo. Analisando a série histórica do sistema de avaliação, é possível observar o comportamento das notas dos alunos do ensino médio e sua estagnação, tanto em língua portuguesa quanto em matemática, o gráfico 2 e 3 demonstram esse comportamento, exibindo a série histórica entre os anos de 1995 e 2017 (Saeb, 2017).

Gráfico 2: Gráfico com as séries históricas das médias das notas em língua Portuguesa dos alunos do 3ª ano do ensino médio brasileiro.



Fonte: Inep/Mec, 2018.

Gráfico 3: Gráfico com as séries históricas das médias das notas em matemática dos alunos do 3ª ano do ensino médio brasileiro.



Fonte: Inep/Mec, 2018.

Esses dados mostram os muitos dos desafios contemporâneos encontrados na educação básica, principalmente no ensino médio. Dentre esses desafios, a qualidade, os baixos índices de aprendizagem, a possível promoção de uma educação integral, bem como a adequação do extenso currículo à realidade, ao mercado de trabalho, aos interesses dos jovens e da sociedade brasileira contemporânea, representam grandes desafios às políticas públicas educacionais.

Muito do que tem sido discutido visa tornar o ensino médio uma etapa mais atraente, comprometida com os desafios da sociedade do conhecimento, em rede e tecnológica (Castells, 2002), tal como expresso pelo Plano Nacional de Educação (Brasil, 2014). O ensino médio acumula as carências das etapas anteriores, que vão desde falhas no acesso à escola, na alfabetização e no aprendizado durante o ensino fundamental, até problemas para a conclusão das demais séries (Castro, 2011). Tornou-se urgente vencer as barreiras de rendimento e desempenho para a construção de gerações mais incluídas, produtivas e com acesso menos desigual às oportunidades.

Como já enfatizado por Krawczyk (2011), o ensino médio brasileiro é o nível de ensino que provoca os debates mais polêmicos, seja pelos persistentes problemas de acesso e permanência, seja pela qualidade da educação oferecida, ou pela discussão sobre a sua identidade.

Nos últimos anos foram produzidos poucos trabalhos acerca dos fatores socioeconômicos que possuem um real fator influente no desempenho do estudante do ensino médio brasileiro. É importante que se procure entender como e o quanto cada fator exerce influência sobre o estudo para que se possa pensar em políticas públicas mais adequadas a fim de melhorar a qualidade da educação oferecida e até mesmo combater a uma possível evasão escolar.

Existe uma preocupação crescente com políticas afirmativas de incentivo ao estudo por parte do governo e sociedade, ao mesmo tempo em que faltam estudos sobre que fatores influenciam no desempenho educacional do estudante e como cada fator contribui positiva ou negativamente para que o aprendizado do estudante seja garantido de maneira efetiva e consequentemente possa ter acesso ao ensino superior. A maior parte dos estudos sobre Mineração de Dados Educacionais prioriza a descrição das técnicas e poucos estudos procuram detectar as causas da evasão, repetência escolar ou até mesmo notas para o acesso ao ensino superior, enquanto esses estudos são muito importantes por apresentarem grandes efeitos financeiros e de impacto em toda a cadeia produtiva do país (RIGO et al 2014).

Nesse cenário preocupante o avanço na utilização de técnicas de mineração de dados vem tornando uma verdadeira aliada para capturar informações sobre os alunos, assim como suas interações com os ambientes de aprendizagem, conteúdos educacionais, avaliações, e assim favorecer para que haja melhorias no processo de aprendizagem, auxiliando gestores, educadores e formuladores de políticas educacionais. (SCAICO; QUEIROZ; SCAICO, 2014)

As instituições de ensino superior atualmente vêm fazendo uso da Mineração de Dados a fim de otimizar o padrão educacional, maximizando o sistema de ensino. As instituições do ensino básico não fazem o mesmo uso, os algoritmos de Data Mining são poucos utilizados, para se descobrir informações importantes que ajudem a instituição a melhorar a taxa de estudantes bem-sucedidos, a questão da proporção de abandono escolar, entre outros. Srivastava e Srivastava (2013) destacam que o papel das tarefas de Data Mining em educação é principalmente adaptar o ensino ao aluno através de recomendações para que o mesmo melhore sua aprendizagem, e até mesmo na criação de um ambiente de aprendizagem individual.

No Brasil, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) realiza o Exame Nacional do Ensino Médio (ENEM), o qual se trata de uma

avaliação criada e aplicada pelo Ministério da Educação para avaliar o desempenho dos estudantes ao fim da educação básica (INEP, 2017). O ENEM passou a ser utilizado como fonte de informações sobre a educação secundária do país. Além do conhecimento teórico do aluno, o ENEM captura informações sócio-econômico-culturais sobre o seu perfil (INEP, 2017). Porém, pouco conhecimento sistemático é extraído dessa base.

Apesar das aplicações de Mineração de Dados serem implementadas nos mais diversos setores, como saúde, varejo, indústria, serviços financeiros, entre outros, o foco deste trabalho é o setor de educação, onde a Mineração de Dados atua no sentido de desenvolver e adaptar algoritmos já existentes para compreender melhor os dados oriundos de contextos educacionais, produzidos por alunos, escolas e professores.

No Brasil nota-se ainda poucas pesquisas no campo emergente de Mineração de Dados Educacionais, desta forma busca-se novos resultados propícios de revelar um panorama da problemática, servindo como fonte de informações para gestores públicos, Diretores, professores e até estudantes interessados no tema, para que possam discutir políticas públicas mais efetivas.

A meta principal deste trabalho é a avaliação de fatores socioeconômicos, ou seja, fatores extraescolares que possuem alguma ligação de influência no rendimento individual dos alunos do ensino médio, tendo como parâmetro as notas individuais dos alunos que fizeram o ENEM. Aplica-se neste trabalho a metodologia *Educational Data Mining* (EDM), este fato impulsiona a realização de novas pesquisas e concede à base de dados governamentais mais que as funcionalidades primárias de armazenamento, recuperação de dados e síntese de relatórios, todavia, está se torna uma fonte de extração de conhecimento até então pouco explorada.

O trabalho proposto compartilha algumas características com os trabalhos presentes na literatura sobre mineração de dados educacionais. Contudo, possui alguns pontos que merecem relevância dentro da pesquisa:

a) Fortalecimento do campo de EDM, uma vez que esta área é nova e não há muitos estudos nacionais;

b) Aplicação em uma vasta base de dados, pois a maioria dos trabalhos usam datasets com poucas centenas de amostras;

c) Diagnóstico voltado também para características estruturais da escola que o aluno frequenta, enriquecendo mais a descoberta de fatores influentes. Não se baseado apenas nos dados socioeconômicos do estudante.

1.2. Objetivo Geral

O objetivo geral desta dissertação é identificar e analisar quais fatores extraescolares e intraescolares influenciam no desempenho dos estudantes do ensino médio, explorando as bases de micro dados do ENEM (Exame Nacional do Ensino Médio) e a do Censo Escolar ambas do ano de 2016, utilizando o algoritmo PCA (*Principal Component Analysis*) e as Redes Bayesianas ao domínio escolhido.

1.2.1. Objetivos Específicos

A realização do objetivo geral está associada com os objetivos específicos enumerados a seguir:

- Realizar um amplo levantamento do estado da arte das áreas contempladas nesta dissertação;
- Entender a modelagem da base de dados para identificar quais os dados estavam disponíveis e são mais relevantes à pesquisa;
- Projetar um banco de dados relacional de dados auxiliar que contemple os atributos selecionados à pesquisa;
- Aplicar à base de dados ferramentas de Mineração de Dados para extração de padrões, buscando obter conhecimento útil acerca da problemática abordada e consequentemente apoiar decisões no âmbito acadêmico;

1.3. Organização do Documento

Este trabalho está organizado da seguinte forma:

- Capítulo 2 apresenta a fundamentação teórica relativa à Mineração de Dados Educacionais, ENEM, Censo Escolar e conceitos de probabilidades e redes bayesianas.
- O Capítulo 3 são apresentados e analisados os trabalhos relacionados ao estudo realizado.

- O capitulo 4 trata por sua vez, de apresentar a modelagem da aplicação proposta.
- Capítulo 5 evidenciam-se os resultados de dois estudos de casos propostos nesta pesquisa.
- Capítulo 6 são debatidas as conclusões, contribuições, dificuldades e propostas de trabalhos futuros.

2. Referencial Teórico

2.1. Considerações Iniciais

Este capítulo tem por finalidade conceituar teoricamente todo o conhecimento referente ao que foi aplicado neste trabalho. São abordados os conceitos de mineração de dados educacionais (MDE), conhecimento a respeito da importância que as bases de dados utilizadas ENEM e Censo Escolar tem para a educação, analise das componentes principais PCA e o Teoria de Probabilidades tal como as características das Redes Bayesianas.

2.2. Mineração de Dados Educacionais

A mineração de dados consiste em extrair ou "minerar" conhecimento a partir de uma grande quantidade de dados. Em parte da literatura relacionada, a mineração de dados é também tratada como sinônimo para outros termos, a descoberta de conhecimento em base de dados (KDD, do ingês *Knowledge Discovery in Database*). Outros autores consideram a mineração uma etapa do processo de KDD, o qual é composto pelas etapas de seleção, pré-processamento e limpeza, transformação, mineração de dados e interpretação, conforme apresentado por Fayyad *et.al.*(1996).

A Mineração de dados (MD) tem sido diversamente aplicada em várias áreas do conhecimento, como por exemplo, vendas, bioinformática, e ações contraterrorismo. Com a expansão dos cursos à distância e também daqueles com suporte computacional, muitos pesquisadores da área de Informática na Educação (em particular, Inteligência Artificial Aplicada à Educação) têm mostrado interesse em utilizar mineração de dados para investigar perguntas científicas voltadas a área da educação (Quais são os fatores que afetam a aprendizagem? Ou como desenvolver sistemas educacionais mais eficazes?) (BAKER *et. al.*, 2011).

Dentro deste contexto, surgiu uma nova área de pesquisa conhecida como "Mineração de Dados Educacionais" (do inglês, "*Educational Data Mining*", ou EDM). A EDM é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais (INTERNATIONAL EDUCATIONAL DATA MINING SOCIETY, 2011).

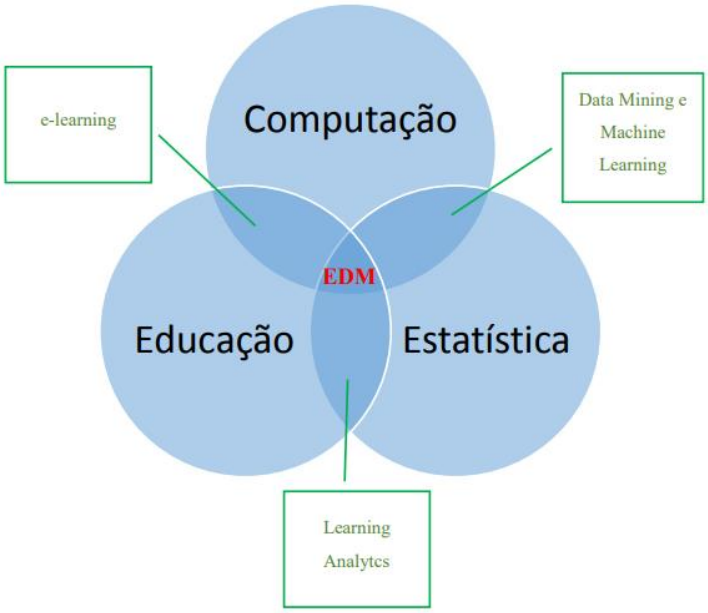
O objetivo geral do processo do KDD é extrais conhecimento de um conjunto dados existentes e transforma-lo em uma estrutura mais compreensível de ser analisada pelo ser humano (SOUMEN et. al., 2006). Por exemplo é possível minerar dados de educandos para verificar a relação entre uma abordagem pedagógica e o aprendizado do estudante. Por meio desta informação o professor poderia compreender se sua abordagem realmente está ajudando e desenvolver novos métodos de ensino mais eficazes (BAKER et. al., 2011).

A área de Mineração de Dados Educacionais diz respeito ao uso de técnicas de MD em domínios que abrangem dados educacionais. Conforme Baker, Carvalho e Isotani (2011), a EDM é definida como a área de pesquisa que tem como objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais.

Como apontado em (García et al. 2011) e (Baker & Yacef 2010), a EDM emergiu nos últimos anos como uma área de pesquisa para pesquisadores das mais diversas áreas (e.g., Computação, Educação, Estatística, Sistemas Tutores Inteligentes, *E-learning*, etc.), quando da análise de grandes volumes de dados com o objetivo de resolver questões voltadas às áreas educacionais associadas às respectivas áreas de pesquisa individuais.

A partir do seu relacionamento com as diversas áreas de conhecimento, Romero e Ventura (2013) afirmaram que a EDM é a combinação de três principais áreas de conhecimento (Figura 2): Computação, Educação e Estatística. A interseção dessas áreas fornece as três subáreas – *E-learning, Data Mining e Machine Learning e a Learning Analytcs* – que estão mais relacionadas com a EDM.

Figura 2: Principais áreas relacionadas com a EDM.



Fonte: Adaptado Adaptado de Romero e Ventura, 2013.

Com o aumento no número de repositórios com dados acadêmicos de alunos e, também, a rápida obtenção (embora restrita) de tais dados por meio de consultas a banco de dados, tem sido viabilizado, de certa forma, o seu uso em um contexto acadêmico, para extração de informações que podem reverter em replanejamento de disciplinas, redirecionamento de práticas pedagógicas, revisão de ementas, e outros. Processos utilizados em MDE podem ser vistos como conversores daqueles dados acadêmicos brutos que foram obtidos por sistemas educacionais tradicionais (sejam eles automatizados ou não), em informação útil ao sistema educacional como um todo, que pode ser utilizada por desenvolvedores de software, professores, pesquisadores educacionais, etc.

Segundo Queiroga (2017) muitos trabalhos buscam modelar comportamentos de alunos com o objetivo de realizar previsões sobre os mesmos, recorrendo a diversas técnicas e bases de dados, na sua maioria com resultados satisfatórios. Segundo Detoni, Cechinel e Araújo (2015), a identificação com antecedência de estudantes que sofrem algum tipo de risco de evasão pode ajudar de maneira decisiva o trabalho de professores e tutores.

Fazendo uso da EDM é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um tipo de abordagem instrucional (e.g. aprendizagem individual ou colaborativa) proporciona melhores benefícios educacionais ao aluno. Também é possível verificar se o aluno está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem.

Baseado no que já foi dito, percebe-se que é viável minerar dados de instituições de ensino, alunos e diversos outros dados ligados a educação, com objetivos normalmente de colher informações que possam levar a melhoria do ensino nas instituições ou aumento do desempenho do aluno. As informações sobre a relação entre dados e, posteriormente a descoberta de novos conhecimentos, podem ser muito úteis para realizar atividades de tomada de decisão.

2.2.1. **Principais Aplicações de EDM**

A MDE vem sendo utilizada para a obtenção de diversos objetivos pedagógicos nos últimos anos, a escolha da tarefa ou subárea de pesquisa de EDM depende dos objetivos definidos para a mineração, do que pretende buscar com os dados, tipos de regularidades ou categorias de padrões que tem o interesse de encontrar. Nesse contexto, Romero (2011) propõe a formação de categorias, as principais são:

- Comunicação com os *Stakeholders*: provê auxilio para os educadores para avaliar as atividades e participação dos estudantes. Os métodos geralmente utilizados são: mineração de processor, geração de relatórios, visualização de dados, e a analise estatística de dados (FERREIRA, 2012);
- Realizar melhorias e manutenção em cursos: dispõem aos educadores estratégias que auxiliem para a melhoria do curso. Os métodos de mineração aplicados geralmente são: Associação, cluster e classificação;
- Gera recomendações: provê recomendações de conteúdo no momento adequado vivenciado pelo estudante. Os métodos de mineração geralmente utilizados são: associação, cluster e classificação (ABEL,2010);

- Prever os resultados de atividades/ provas ou de avaliações de aprendizado: prevê os resultados de testes e de outras avaliações educacionais, com base na análise nas atividades realizadas pelo estudante. Para essa aplicação geralmente é utilizada como método de mineração a associação, cluster e classificação (FERREIRA, 2012);
- Criar modelo de estudantes: tem por objetivo estudar determinadas características dos alunos. Para isso são utilizados os métodos de análise estatística, rede bayesianas, modelos psicométricos e aprendizado por reforço;
- Análise da estrutura do domínio: busca avaliar a estrutura do domínio, através da análise da performance (como o domínio realiza uma tarefa). As técnicas mais comuns para esse tipo de aplicação incluem, regras de associação, métodos de clusterização e algoritmos de busca. Ex_1: A universidade ao oferecer um programa de monitoramento poderá comparar o desempenho de um domínio de alunos que usam a monitoria com os que não usam. Ex_2: Verificar quais fatores podem influenciar no conceito final do vestibular, como sexo, bairro, instituição do ensino médio e etc.
- Aprendizagem individualizada: foca nas necessidades individuais de cada aluno. O objetivo, em geral, é manter as características fortes e melhorar aquelas abaixo de uma média, após identificá-las. Algumas metas como sugestão de direcionamento usadas na EDM são a Previsão do resultado final, Dificuldade em uma matéria ou assunto e Sugestão de atividades complementares etc.

2.3. ENEM

O ENEM é uma prova de abrangência nacional, aplicada anualmente pelo INEP órgão vinculado ao MEC (Ministério da Educação). Apesar do exame ter sido criado, em 1998, com a finalidade avaliar o perfil dos concluintes do ensino médio no país, ele gradualmente incorporou uma série de outros usos como: critério parcial ou exclusivo para ingresso à ampla maioria das universidades federais, como principal quesito de acesso a bolsas de estudo (PROUNI) e crédito estudantil financiado pelo governo (FIES), (INEP, 2017).

O ENEM ganhou notoriedade principalmente quando passou a “democratizar as oportunidades de concorrência às vagas federais de ensino superior”, onde o

participante tem a oportunidade de concorrer a vagas em qualquer localidade do território nacional, de instituições de educação superior públicas, sem o ônus do deslocamento para a realização das provas.

Segundo o INEP (2017), o Ministério da Educação apresentou uma proposta de reformulação do ENEM, no ano de 2010, onde ficou decretado que o mesmo passaria a ter 180 questões ao invés de 63 como era até 2008, além disso o ENEM passou a ser aplicado em dois dias. Além de que começou a explorar quatro áreas do conhecimento humano, dividindo as questões igualmente (45 questões) para cada uma das seguintes áreas do conhecimento:

- Linguagens, códigos e suas tecnologias: Literatura, Educação Física, Língua Estrangeira (Inglês ou Espanhol), Língua Portuguesa, Artes e Tecnologias da Informação e Comunicação;
- Matemática e suas tecnologias;
- Ciências da Natureza e suas tecnologias: Biologia, Física e Química;
- Ciências Humanas e suas tecnologias: Filosofia, História, Geografia e Sociologia.

Atualmente o principal objetivo do ENEM é a análise do desempenho escolar e acadêmico ao final do Ensino Médio, segundo o INEP os resultados podem:

- Proporcionar o estabelecimento de parâmetros para a auto avaliação do participante, buscando a continuidade de sua formação e a sua introdução no mercado de trabalho;
- Proporcionar a criação de referência nacional para a melhoria dos currículos do Ensino Médio;
- Ser utilizados como mecanismo único, alternativo ou complementar para o ingresso na Educação Superior, especialmente, a disponibilizada pelas instituições federais de educação superior;
- Proporcionar a entrada do participante em programas governamentais de financiamento ou apoio ao estudante da Educação Superior;
- Ser usado como instrumento de seleção para ingresso nos diferentes setores do mundo do trabalho;
- Possibilitar o desenvolvimento de estudos e indicadores sobre a educação brasileira.

2.4. Censo Escolar

O Censo Escolar é um levantamento de dados estatístico-educacionais de âmbito nacional realizado todos os anos e coordenado pelo INEP Ele é feito com a colaboração das secretarias estaduais e municipais de Educação e com a participação de todas as escolas públicas e privadas do país. Trata-se do principal instrumento de coleta de informações da educação básica, que abrange as suas diferentes etapas e modalidades: ensino regular (educação Infantil e ensinos fundamental e médio), educação especial e educação de jovens e adultos (EJA), (INEP,2017).

Segundo o INEP (2017) o Censo Escolar coleta dados sobre estabelecimentos, matrículas, funções docentes, movimento e rendimento escolar. Essas informações são utilizadas para traçar um panorama nacional da educação básica e servem de referência para a formulação de políticas públicas e execução de programas na área da educação, incluindo os de transferência de recursos públicos como merenda e transporte escolar, distribuição de livros e uniformes, implantação de bibliotecas, instalação de energia elétrica, Dinheiro Direto na Escola e Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb).

É com base nas informações captadas pelo Censo Escolar que o Inep atualiza anualmente o Cadastro Nacional de Escolas e obtém dados referentes à matrícula, movimento e rendimento dos alunos, incluindo informações sobre o sexo, turnos, turmas, séries e períodos, condições físicas dos prédios escolares e equipamentos existentes, além de informações sobre o pessoal técnico, administrativo e docente, por nível de atuação e grau de formação. As informações captadas permitem, portanto, traçar o perfil dos diferentes segmentos da comunidade escolar – alunos e professores, sobre a situação de funcionamento e infraestrutura dos estabelecimentos de ensino – caracterização física, instalações, equipamentos e insumos pedagógicos; sobre as formas de organização do ensino – seriado, em ciclo ou por disciplina; e sobre o movimento e rendimento escolar afastamentos e transferências, aprovações e reprovações.

Os Micro dados passaram a ser estruturados em formato CSV (*Comma-Separated Values*), e seus dados estão delimitados por Pipe (|), de modo a garantir que praticamente qualquer software estatístico, inclusive *open source*, consiga importar e carregar as bases de dados. Devido à amplitude das bases, os arquivos são divididos por

região geográfica (Norte, Nordeste, Sudeste, Sul e Centro-Oeste), tanto para as variáveis de Matrículas, quanto para as de Docentes.

Todas as escolas das redes pública e privada respondem ao Censo Escolar por meio do preenchimento de questionário padronizado. O Censo Escolar abrange um universo de cerca de 52 milhões de alunos e 266 mil escolas, distribuídas em mais de 5.500 municípios. A coleta dos dados e o processamento das informações são operacionalizados pelas Secretarias Estaduais de Educação, sob a coordenação da Diretoria de Informações e Estatísticas Educacionais (Seec), do Inep.

2.5. PCA (Analise de Componentes Principais)

A análise de componentes principais, conhecida como *Principal Component Analysis* (PCA) é uma técnica utilizada para reduzir a dimensionalidade de um conjunto de dados onde há um grande número de variáveis inter-relacionadas. É feito de forma que o máximo de variância presente nos dados seja mantido. Essa redução se dá pela obtenção de um novo conjunto de variáveis não correlacionadas, denominadas de componentes principais, tais componentes estão ordenados de forma que os primeiros possuem a maior parte da variância presente nas variáveis originais.

Para Dunteman (1999), o PCA elimina informações redundantes, destacando os recursos ocultos, provenientes das informações contidas nas bases, e visualiza as principais relações existentes entre as observações vistas. Uma das ferramentas mais clássicas e populares para a análise de dados e redução de dimensionalidade, com ampla gama de aplicações bem-sucedidas em toda a ciência e engenharia (JOLLIFFE, 2002).

O método PCA é um método de aprendizagem não supervisionada, que visa encontrar a combinação de condições que explicam a maior variação nos dados (YANG et al., 2008), utilizado em muitos tipos de análises incluindo neurociência e computação gráfica (SHLENS, 2005), além de análises de dados de micro arranjos (HOLMES et al., 2011; YANG et al., 2008).

As definições a seguir abrangendo análise de componentes principais surgiram a partir dos estudos de Song et al. (2013). PCA é uma decomposição de valores próprios da matriz de covariância dos dados, utilizado para aproximação de baixo rank, que

compara os dados através de uma função linear de variáveis (MARKOS; VOZALIS; MARGARITIS, 2010).

Matematicamente, os componentes principais são obtidos calculando os autovalores da matriz de covariância C, como apresentada na equação (2.1):

$$\mathbf{Cvi = \lambda ivi} \qquad (2.7)$$

A matriz de covariância dos vetores dos dados originais X é representada por C, λi refere-se aos autovalores da matriz C e vi corresponde aos autovetores correspondentes. Consecutivamente, a fim de reduzir a dimensionalidade dos dados, os autovetores k, que correspondem aos maiores autovalores k, precisam ser computadorizados (XU et al., 2005).

Considerando $\boldsymbol{E_k}$ = [$v1$, $v2$, $v3$, … … . , vk] e Λ = [$\lambda 1$, $\lambda 2$, $\lambda 3$ … . . λk], logo tem-se $CEk = EK\Lambda$. Portanto, obtemos a seguinte equação (2.2):

$$\mathbf{X}^{\mathbf{PCA}} = \mathbf{E_k}^{\mathbf{T}}\,\mathbf{X} \qquad (2.2)$$

Em relação à Equação (2.2), o número das características da matriz de dados original X é reduzido pela multiplicação com a matriz d x k E_k que tem auto vetores k correspondentes aos maiores autovalores k. O resultado da matriz é X^{PCA} (BINGHAM; MANNILA, 2001).

2.6. Redes Bayesianas

A rede bayesiana é estruturada através do Teorema de Bayes, que foi formulado matematicamente para cálculo de probabilidades proposta por Thomas Bayes, em 1763.

Considerando os eventos arbitrários A e B, tal que P(A) ≠ 0 e P(B) ≠ 0, tem-se a Equação 2.3:

$$\boldsymbol{P(A|B) = \frac{P(B)P(B)}{P(A)}} \quad (2.3)$$

O teorema de Bayes representa uma conjugação do teorema de probabilidade condicional e da fórmula de probabilidades totais conforme está disposto na equação 2.4, permitindo-se calcular a probabilidade a *Posteriori P(B|A)* em termos da informação a *priori P(B) e P(A|B).*

$$\boldsymbol{P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}[P(A|B_j)P(B_j)]}} \quad (2.4)$$

A redes bayesianas (RB) é uma estrutura de dados proposta por Judea Pearl em 1986 (PEARL,1986), chamada Redes Bayesianas (RB); conhecida também por Redes de crença, Rede probabilística, Rede Causal e mapa de conhecimento (RUSSEL; NORVING, 2010) (HAN; KAMBER; PEI, 2012). Essa estrutura refere-se à representação gráfica de variáveis e suas relações causais sob um determinado cenário de aplicação.

Ko e Kim (2014) defendem a importância da representação do conhecimento em Inteligência Computacional, pois isso determina o quão fácil as características presentes na base de dados podem ser compreendidas. A argumentação dos autores, dentro do contexto estudado, sobreleva a relevância da Rede Probabilística, uma vez que Hlel, Jamoussi e Hamadou (2016) julgam a RB como um dos melhores modelos de interpretação do conhecimento baseado em formalismos teóricos.

As RBs têm sido empregadas na resolução de problemas em diversos campos de estudo. Borunda et al. (2016) destacam aplicações em ramos como diagnósticos médico, busca heurística, questões ambientais, gestão de bacias hidrográficas, sensores virtuais, entre outras. Essa gama de trabalhos confirma a utilidade e flexibilidade das redes de crenças na modelagem de cenários complexos em uma notação simples, não obstante objetiva. Isto posto, na pesquisa apresentada nesta dissertação explora-se esse potencial para o desenvolvimento de modelos voltados a descoberta de padrões alusivos à desistência ou retenção em cursos de graduação.

2.6.1. Estrutura das Redes Bayesianas

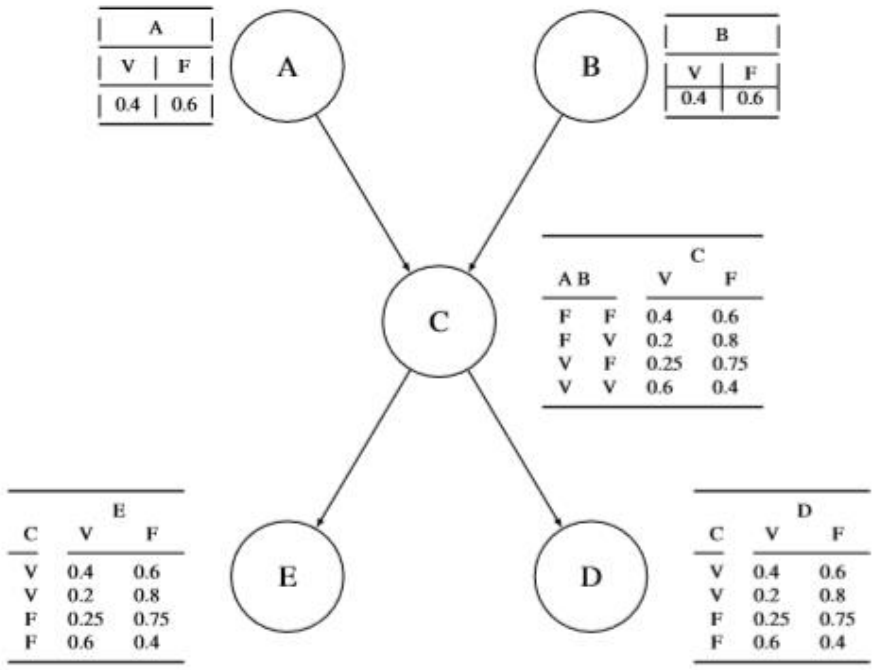
A rede de crenças é um grafo direcionado, conectado e acíclico, conhecido na literatura por *Directed Acyclic Graph* (DAG), onde cada nó é uma variável aleatória. Dessa forma os conceitos enumerados anteriormente acerca de teoria dos grafos significam, respectivamente: estruturas onde há a direção dos arcos; todos os nós estão conectados na rede e, finalmente, partindo-se de um nó arbitrário e seguindo por todos os arcos conectados não há retorno.

Os arcos referem-se às dependências ou influências, se houver um vínculo de um nó X até outro Y, denomina-se que X é pai de Y. A cada nó Xi há uma distribuição de probabilidade condicional P(Xi|Pais(Xi)) que quantifica o efeito dos pais sobre o nó (RUSSEL; NORVING, 2010). Segundo Russel e Norving (2010), a topologia da rede ou o conjunto de nós e vínculos (arcos) específica os relacionamentos de independência condicional e a sua semântica gráfica sugere que as causas são pais dos efeitos.

Outro elemento importante às redes probabilísticas é a tabela de probabilidade condicional (TPC), que denota uma matriz que contém a probabilidade condicional de cada valor de nó à combinação de possíveis valores dos nós pai. A soma de cada linha deve ser igual a 1, uma vez que as entradas representam combinações exaustivas dos possíveis eventos em cada variável.

A RB de um domínio hipotético mostrada na Figura 3, esquematicamente exemplifica os conceitos debatidos nesta seção. Observa-se a presença de cinco nós booleanos, sendo que A e B são pais de C, enquanto D e E, filhos de C, são nós folhas. As tabelas de probabilidade condicional são apresentadas nos nós C, E e D, exceto aos nós A e B que são as variáveis pais, as TPCs apresentam a probabilidade combinada exaustivamente para todas as possibilidades dos nós pais. Diante disso, por exemplo, nota-se que para P(C|A,B) onde a probabilidade de C = “v” (verdadeiro) é de 0,25, dado que A = “v” e B= “F” (falso).

Figura 3: Redes Bayesianas com tabelas de probabilidade condicional dos nós A, B, C, D e E.



Fonte: Couto, 2017.

As topologias das redes e as tabelas de probabilidades são elementos cruciais para o raciocínio sobre incertezas, isto é, na descoberta de padrões dos quais se conhece pouco acerca do domínio de aplicação, no que tange às relações de causas e efeitos. Assim pode-se responder muitas questões acerca do domínio somente por meio de inferências.

2.6.2. Processo de Inferência

Segundo (KORB; NICHOLSON, 2010) o termo inferência ou atualização de crença (*belief updating*), contextualizado em sistemas aplicados a redes de bayesianas, refere-se ao cálculo da distribuição de probabilidade posteriori, atualizada por toda a estrutura da rede, dado um conjunto de variáveis aleatórias de evidências.

O teorema de Bayes, apresentado na Equação 2.3, é a base de todos os sistemas modernos de inferência probabilística (RUSSEL; NORVING, 2010). A probabilidade posteriori é calculada a partir da generalização desse teorema (Equação 2.4), onde o

termo $\frac{1}{\sum_{j=1}^{n}[P(A|Bj)P(Bj)]}$ denota uma constante de normalização α, necessária para tornar a soma das entradas P(A|Bj) igual a 1. Dessa discussão resulta que a probabilidade a posteriori é conseguida pela Equação 2.5, tal que j representa os possíveis estados da variável B.

$$\boldsymbol{P(A) = \alpha P(Bj)P(Bj)} \tag{2.5}$$

Retomando à topologia disposta na Figura 1, é permitida a formalização do processo de inferência. Por conseguinte, considera-se saber a saída do sistema, dadas as consultas P(D,C,E,A,B). A resposta da inferência se obtém a partir do teorema de Bayes e o seu detalhamento matemático está mostrado na Equação 2.6.

$$\begin{aligned} \boldsymbol{P(D,\underline{C},E,A,B)} &= \boldsymbol{P(\underline{C})P(\underline{C},E,A,B)} \\ &= \boldsymbol{P(D,\underline{C})P(\underline{C})P(\underline{C},A,B)} \\ &= \boldsymbol{P(\underline{C})P(\underline{C})P(A,B)P(A)P(B)} \end{aligned} \tag{2.6}$$

As inferências aplicadas nas Redes Bayesianas referem-se ao cálculo da distribuição de probabilidade posteriori, atualizada por toda a estrutura da rede, dado um conjunto de variáveis aleatórias de evidências. (RUSSEL; NORVING, 2010) denota a forma geral da inferência, onde Pais (Xi) representam todos os valores em Pais(Xi) que aparece em X1,.., Xn, representada na equação 2.7:

$$\boldsymbol{P(X_1, \dots\dots, X_n) = \prod_{i=1}^{n} P\,(X_i|pais(X_i))} \tag{2.7}$$

2.6.3. Aprendizagem da Estrutura

O processo de aprendizagem da estrutura de redes de crenças pode ser concebido em duas modalidades distintas. A primeira forma, recorre ao emprego de algoritmos que aprendem a partir de uma base de dados o grafo que melhor representa a distribuição de probabilidade do conjunto de dados (SARDINHA; PAES; ZAVERUCHA, 2009); enquanto a outra possibilidade, vale-se de interações com usuários especialistas no domínio de aplicação. De acordo com Santana (2008), contudo, verifica-se grande consumo de tempo na execução desta última, visto que é mais factível a combinação das duas estratégias dirigida à validação de modelos.

Santana (2008) assegura a grande relevância do estudo de técnicas dedicadas ao aprendizado de estrutura, pelo fato de o tamanho do espaço de busca de possíveis estruturas aumentar exponencialmente junto com o número de variáveis do modelo. Em virtude disso, as topologias de certas redes tornam-se demasiadamente complexas, assim dificultam a interpretabilidade oriunda dos seus elementos estruturais.

Sardinha, Paes e Zaverucha (2009) destacam a existência de três categorias de algoritmos de aprendizado de estrutura, a saber: baseado em pontuação (score-based); baseados em restrições (constraint-based); e híbridos, agregadores das particularidades de ambos os anteriores. Os algoritmos score-based adotam uma função de avaliação para qualificar a rede bayesiana é uma heurística para contornar situações nas quais o problema de buscas das redes, em todo o espaço, torna-se intratável. A abordagem constraint-based, por sua vez, faz uso de diversos testes de independência condicional como por exemplo, algum teste de hipótese ou score oriundo da teoria da informação (SARDINHA; PAES; ZAVERUCHA, 2009)

Neste trabalho é empregado um método clássico de aprendizado de estrutura baseado em pontuação, denominado K2 (COOPER; HERSKOVITS,1992). Esta opção se justifica, segundo as asserções de Ko e Kim (2014), como sendo um dos melhores e mais eficientes métodos de aprendizado, além da facilidade de implementação (YANG; CHANG, 2002). O K2 percorre todo o espaço de busca valendo-se da métrica de pontuação dada por Equação 2.8.

$$\boldsymbol{P(B_s|X) = \prod_{i=1}^{n} \prod_{j=1}^{qi} \frac{(ri-1)!}{(Nij+ri-1)!} \prod_{k=1}^{ri} Nijk!} \quad (2.8)$$

Onde:

— *X* é a base de dados com *n* observações;
— *Bs* a dimensão de estrutura;
— *ri* é a quantidade de valores que a variável *Xi* pode assumir;
— $N_i j_k$ é o número de observações na base *X*, tal que X_i é configurado com o valor *k* e seus pais com o valor *j*.

Resumindo e destacando os pontos positivos para se utilizar as Redes Bayesianas, segundo Luna (2004) e Tan et. al. (2009) dentre as suas principais características destacam-se:

- Permitir expressar as assertivas de independência de forma visual facilitando a percepção;
- Tornam o processo de inferência eficiente computacionalmente;
- Permite analisar grande quantidade de dados;
- Pode ser aplicada em diversos domínios;
- São apropriadas para tratar situações com valores de atributos incompletos por meio de soma ou integração de probabilidades pelos valores possíveis do atributo;
- São robustas para tratar *overfitting* devido à combinação probabilística dos dados com conhecimento anterior sobre a situação.

3. TRABALHOS CORRELATOS

3.1. Considerações Iniciais

Nesta seção será apresentado um levantamento dos trabalhos correlatos pesquisados e selecionados que fazem o uso da descoberta do conhecimento no contexto educacional, mostrando as técnicas de mineração que são aplicadas e traz algumas abordagens que se relacionam com o objetivo desta proposta de trabalho.

3.2. Correlatos

No Brasil, o desafio de analisar e compreender o comportamento dos alunos é muito grande devido a diversidade da população. De acordo com Blanchard et al., existe uma correlação entre os dados socioculturais dos alunos e suas ações, atitudes e comportamentos apresentados durante a aprendizagem.

As correlações dos dados sociais e econômicos com o desempenho do estudante do ensino médio são demonstradas nos trabalhos de Silva, L. A., Morino, A. H.; Sato, T. M. C. (2014) e Adeodato, Paulo JL; Santos Filho, Maílson M.; Rodrigues, Rodrigo L. (2014), os trabalhos utilizam os dados dos questionários socioeconômicos preenchidos por alunos participantes da edição de 2010 do Enem. No trabalho de Silva, L. A., Morino, A. H.; Sato, T. M. C. (2014) eles analisam apenas os dados das capitais da região Sudeste do Brasil. Os autores adotam a técnica de mineração baseada em regras de associação, utilizando diferentes parametrizações do algoritmo *Apriori*, analisando somente as respostas de quatro perguntas, são relacionadas com os resultados de desempenho desses alunos. Em suas conclusões, são identificados fatores que diminuem o desempenho dos alunos, como a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas que moram com o estudante.

Já a pesquisa realizada por Adeodato, Paulo JL; Santos Filho, Maílson M.; Rodrigues, Rodrigo L. (2014), diferente do trabalho anterior usa a base do Enem junto com o Censo Escolar, mas analisa apenas as escolas da rede privada no Brasil, classificando as escolas em boas ou não. Para obter os resultados os autores utilizaram a Regressão logística e árvore de decisão como técnicas para identifica e quantifica os principais atributos que têm influência na qualidade da escola. Os resultados mostraram

que os principais fatores que influenciam a boa qualidade das escolas estão ligados a situação econômica e financeira, seja de maneira direta (renda familiar) ou indiretamente, ou em aspectos culturais (nível de educação da mãe ou do pai) da família.

Continuando a analisar os trabalhos que realizam o estudo com as variáveis socioeconômicas coletadas no Enem, o trabalho de Stearns, B., Rangel, F., Rangel, F., Firmino, F., and Oliveira, J. (2017) e Simon, Augusto; Cazella, Sílvio (2017) utilizam dados do Enem do ano de 2015. Stearns, B., Rangel, F., Rangel, F., Firmino, F., and Oliveira, J. (2017) utilizam apenas os resultados das notas de matemática devida a sua alta variância, com o objetivo de comparar a capacidade de generalização de dois métodos de agrupamento por árvore de decisão para a nota dos estudantes no exame. Os modelos utilizados foram os métodos de regressão baseados em boosting de árvore de decisão, o AdaBoost e o Gradient Boosting. Para avaliar a performance dos regressores as seguintes métricas foram utilizadas Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) e R-Squared (R^2). Os melhores resultados foram obtidos com o Gradient Boosting com alta relevância estatística. Os autores concluem que os indicadores socioeconômicos podem explicar um viés na pontuação dos alunos, mas não fazem nenhuma analise especifica com os fatores pois não era a finalidade da pesquisa.

Continuando a usar a nota de matemática como alvo de analise, Alves, Cechinel e Queiroga (2018) propõem um trabalho com o objetivo de encontrar padrões e gerar um modelo preditivo do indicador de desempenho das notas da prova de matemática e suas Tecnologias das escolas do ensino médio, por meio dos dados abertos referentes ao ENEM 2015. Os algoritmos utilizados foram *Naive bayes* e J48 por meio do pacote de software WEKA.

Analisando o trabalho de Simon, Augusto; Cazella, Sílvio (2017), diferente dos autores anteriores, eles realizaram um estudo que possui o objetivo de gerar um modelo preditivo do indicador de desempenho médio na área de ciências da natureza e suas tecnologias através dos valores médios de proficiência dos alunos, utilizando a a base por escolas disponibilizada nos microdados do Enem. Através do software WEKA os mesmos realizaram a análise por árvore de decisão usando o algoritmo j48, que foi feita a partir da opção de *cross-validation*, com o valor para *fold* igual a dez e com a variável

a ser predita Média Escola. A árvore de decisão gerada conseguiu acertar 77,02% instâncias das 15998 inseridas.

Esses trabalhos mesmo analisando a variáveis socioeconômicas a partir dos dados dos estudantes que realizaram o exame do Enem não importando o ano da sua realização, não se detiveram em realizar uma análise com todas as variáveis presentes na base, não fizeram analise das variáveis influentes a nível Brasil, não utilizam as cinco áreas de conhecimento avaliadas no exame e não utilizaram as três redes de ensino ao mesmo tempo, rede privada, estadual e federal. Além disso os trabalhos analisados preocupam-se mais com os resultados de desempenho dos algoritmos, sendo a principal parte discutida.

Continuando com pesquisas usando dados abertos do Inep, tem o trabalho de Ferreira (2015) com o foco na identificação de fatores relacionados à conclusão do Ensino Fundamental utilizando os microdados do Censo Escolar da educação básica do Inep de 2014. Foram aplicados os algoritmos de classificação J48 e o filtro CfsSubsetEval, ambos da ferramenta Weka. A autora concluiu que é interessante a aplicação de filtros, pois permite uma melhor visualização das regras obtidas, mesmo diante de uma diminuição da acurácia dos resultados.

Alves (2018), tendo em vista que o mesmo buscou gerar modelos para predição do desempenho da redação do ENEM, por meio dos microdados do ENEM 2016 e algoritmos de classificação. A base de dados utilizada por Alves (2018) em sua pesquisa, continha 8.627.367 de instâncias. Neste cenário, foram realizados testes onde os dados foram categorizados, para alcançar um melhor resultado no uso dos algoritmos. A classe predita foi a nota da redação que foi categorizada como: baixo, médio, alto e nulo. Os modelos finais foram treinados e testados por meio dos algoritmos: *Naive Bayes* e J48. O uso desses algoritmos foi realizado por meio do software WEKA. O modelo com a maior eficácia conseguiu prever 61.7464% das amostras presentes na base de dados do ENEM 2016.

No âmbito do desempenho escolar sem utilizar os dados do Enem ou Censo, Laisa e Nunes (2015) têm como objetivo analisar uma base de dados de alunos do ensino médio de uma escola particular do ano de 2011 a 2014. Utilizando o algoritmo J48, para assim realizar uma análise de padrões e características dos alunos que

estiverem em risco de reprovação. Os experimentos chegaram a taxas de 77,9% de acertos. Segundo as autoras, a técnica de classificação aplicada permitiu compreender as informações armazenadas dos estudantes de ensino médio.

No contexto da evasão escolar, o objetivo do trabalho de Calixto et al. (2017) consistiu na identificação de variáveis relacionadas a este indicador educacional, utilizando os dados do censo escolar no âmbito do Ceará e Sergipe. As análises se deram por meio de técnicas de indução de regras e regressão logística. A idade, a etapa e a modalidade de ensino, a existência de laboratórios e localização da escola se destacaram como variáveis influentes na evasão escolar.

No trabalho de Gottardo (2012) foi investigado como os dados armazenados por um AVA (Ambiente Virtual de Aprendizagem) poderiam ser transformados em informações potencialmente úteis para apoiar o acompanhamento de estudantes em cursos de ensino a distância (EAD). As informações geradas foram inferências envolvendo estimativas de desempenho acadêmico futuro de estudantes. Os resultados obtidos com a aplicação dos algoritmos de classificação *RandomForest* e *MultilayerPerceptron*, sobre o conjunto de atributos selecionados demonstram que é possível obter inferências relativas ao desempenho dos estudantes com taxas de acurácia global variando entre 72% e 80%.

Em Detoni et al. (2015), é investigado a reprovação no cenário acadêmico do ensino à distância (EaD). Utiliza-se como atributo as contagens de interações com ambiente virtual e demonstrou que as redes bayesianas se mostraram o modelo mais adequado de predição.

Todos os trabalhos apresentados possuem contribuições na área desta dissertação e, em alguma medida, colaboraram para a construção desta dissertação. O quadro 1 abaixo descreve as lacunas e síntese dos principais trabalhos abordados neste capítulo.

Quadro 1- Síntese dos trabalhos pesquisados.

Autores	Principais Lacunas Encontradas
Silva, l. A., Morino, a. H.; Sato, t. M. C. (2014)	O trabalho utiliza apenas a o algoritmo A Priori, e a sua base se limita apenas a estudantes das capitais do Sudeste.
Adeodato, Maílson, Rodrigues (2014)	Avaliam apenas a qualidade das escolas particulares. Não junção das bases ENEM e Censo escolar, deixam de lado as características socioeconômicas dos estudantes.
Laisa e Nunes (2015)	Fazem do algoritmo J48 para identificar padrões de alunos que tende a reprovação no ensino médio de uma escola particular.
Ferreira (2015)	Utiliza apenas os dados do Censo Escolar e aplica o algoritmo de classificação J48 para visualizar regras para identificar fatores relacionados à conclusão do Ensino Fundamental.
Detoni et al. (2015)	Investigam apenas se a rede Bayesiana é um modelo adequado para investigar as variáveis relacionadas a reprovação de estudantes do ensino a distância.
Stearns, B., Rangel, F., Rangel, F., Firmino, F., and Oliveira, J. (2017)	Faz apenas uma comparação entre dois métodos de agrupamento por árvore de decisão para avaliar se fatores socioeconômicos influenciam nas notas de matemática dos alunos que fizeram o ENEM 2015.

Simon, augusto; Cazella, Sílvio (2017)	Faz predição apenas dos indicadores de desempenho médio na área de ciências da natureza e suas tecnologias.
Calixto et al. (2017)	Utilizam apenas dados do Censo Escolar para identificar as variáveis relacionadas a evasão escolar no estado do Ceará e Sergipe.
Alves, Cechinel e Queiroga (2018)	Utiliza o Naive bayes e J48 com o objetivo de encontrar padrões e gerar um modelo preditivo do indicador de desempenho apenas nas notas da prova de matemática e suas Tecnologias das escolas de ensino médio.
Alves (2018)	Prever o desempenho dos estudantes utilizando apenas as notas da redação de ENEM 2016, utilizou o Naive Bayes e J48.

Fonte: Autor, 2019.

3.3. Considerações Finais

Pelo estudo dos trabalhos expostos neste capítulo, observa-se que há um conjunto de pesquisas sendo realizadas no sentido de aplicar técnicas de mineração de dados para estabelecer padrões ou classificar fatores que afetam o desempenho do estudante. Existem poucos trabalhos publicados relacionados à utilização das Redes Bayesianas no âmbito educacional.

Desta forma a pesquisa contribui auxiliando no diagnóstico e avaliação de quais condições socioeconômicas e estruturais das escolas que os alunos frequentam mais contribuem para que os estudantes possam ter um bom desempenho no ENEM. A consolidação da proposta deste trabalho baseia se em apresentar uma resposta inicial sobre a problemática analisada no contexto nacional, realizando um diagnóstico por regiões. Nessa pesquisa apresentada nessa dissertação, explora o potencial para o

desenvolvimento de modelos voltados a descoberta de padrões e fatores que influenciam no desempenho de estudantes que estão concluindo o ensino médio e pretende ingressar no ensino superior.

4. METODOLOGIA EMPREGADA

4.1. Considerações Iniciais

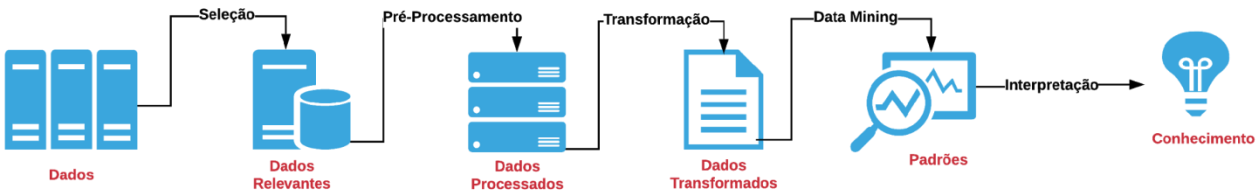
Neste capítulo é apresentada a metodologia utilizada no presente trabalho, desde a etapa da seleção até a forma de análise dos resultados.

4.2. Metodologia Empregada

A metodologia proposta engloba o emprego da mineração de dados, conhecida também como Descoberta de Conhecimentos em Bancos de Dados, ou KDD (do inglês, "*Knowledge Discovery in Databases*"). Com o objetivo de extrair conhecimento, que sejam importantes para tomadas de decisões, que podem ser encontradas nas informações que estão ocultas no grande volume de dados, como no caso dessa dissertação, que tem como objeto de estudo os dados fornecidos pelo INEP.

Para a obtenção deste modelo, foram realizadas as cinco etapas que compõem o processo do KDD em duas bases fornecidas pelo Inep, os microdados do ENEM e uma base originada da junção do ENEM e Censo Escolar ambas de 2016, bases essas que constam informações sobre as questões, as notas dos candidatos, informações socioeconômicas dos estudantes, características estruturais da escola onde estudam e formação do corpo docente, as etapas consistem em: Seleção, Pré-Processamento, Transformação, Mineração de Dados e Interpretação/Avaliação. A Figura 4 abaixo exemplifica o esquema proposto.

Figura 4: Diagrama da metodologia do KDD utilizada no trabalho.



Fontes: Adaptação de Fayyad et al. 1996.

4.3. Seleção Dados Selecionados

Nessa etapa realizou-se a seleção manual de quais variáveis se faziam necessárias ao domínio de aplicação escolhido, em cima da base do Inep, reduzindo os números de dados do ENEM e Censo Escolar de 2016 que foram analisados mantendo seus aspectos relevantes que são de suma importância para a análise. Sendo exigida logo de início uma boa compreensão do domínio da aplicação e auxílio de especialistas para a definição de que atributos eram mais relevantes para o domínio da pesquisa.

4.3.1. **Base de dados do ENEM 2016**

Os microdados do ENEM 2016, possui um tamanho de mais de 5 gigabytes e conta com um conjunto de 166 variáveis e conta com mais de 8 milhões de instâncias que representa o número de inscritos no exame em todo país.

Após análise inicial da base ENEM, observou-se a necessidade de remoção de algumas variáveis. Para tanto, descartou-se as variáveis julgadas como redundantes, como por exemplo, o código do município e o nome do município onde o candidato realizou a prova, as variáveis consideradas não relevantes, como o número da inscrição do candidato, e as variáveis que representam casos específicos, os quais não estão sob análise neste trabalho, como por exemplo, se o candidato necessitou realizar a prova em braile e as demais questões em relação à deficiência. Assim, o número total de variáveis a serem analisadas foi reduzido para 43, as quais são listadas na tabela 1.

Tabela 1: Tabela com as variáveis selecionadas da base de dados ENEM 2016 e sua descrição.

Variáveis	Descrição
TP_SEXO	Tipo de Gênero do participante.
SG_UF_ESC	Sigla da Unidade da Federação da escola.
TP_ESTADO_CIVIL	Estado Civil.
TP_COR_RACA	Cor/raça que o participante declarou.
TP_ESCOLA	Tipo de escola do Ensino Médio.
TP_ENSINO	Tipo de instituição que concluiu o Ensino Médio
IN_TREINEIRO	Inscrito fez a prova com intuito de apenas treinar.

TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola) Federal, Estadual ou Municipal.
TP_PRESENCA_CH	Presença na prova objetiva
TP_PRESENCA_CN	Presença na prova objetiva
TP_PRESENCA_LC	Presença na prova objetiva
TP_PRESENCA_MT	Presença na prova objetiva
NU_NOTA_CN	Nota da prova Ciências da Natureza
NU_NOTA_CH	Nota da prova de Ciências Humanas
NU_NOTA_LC	Nota da prova Linguagem e Seus Códigos
NU_NOTA_MT	Nota da prova Matemática
NU_NOTA_REDACAO	Nota da prova de Redação
Q001	Grau de estudo do Pai
Q002	Grau de estudo da Mãe
Q003	Ocupação do Pai
Q004	Ocupação da Mãe
Q005	Número de pessoas na residência
Q006	Renda mensal Familiar
Q010	Se a família Possui carro
Q011	Se a família Possui Moto
Q013	Na casa possui Freezer
Q016	Na casa possui micro-ondas
Q019	Na casa possui Televisão
Q022	Na casa possui celular
Q023	Na casa possui Telefone fixo
Q024	Na casa possui tem Computador
Q025	Na casa tem acesso a internet
Q026	Exerce ou já exerceu atividade remunerada
Q043	Modalidade de Ensino Fundamental que Frequentou.
Q044	Turno que frequentou o Ensino Fundamental

Q045	Abandonou ou reprovou no ens. Fundamental.
Q048	Modalidade de Ensino Médio.
Q049	Turno que frequentou o Ensino Médio.
Q050	Abandonou ou reprovou no Ensino Médio.

Fonte: Autor, 2019

4.3.2. **Base de dados do Censo Escolar**

A base do Censo Escolar 2016 é composta por três tabelas, uma com os dados referentes as matriculas dos alunos, outra é a tabela escola que possui os dados referentes a estrutura e os equipamentos presentes nas escolas e uma com os dados dos docentes nas escolas brasileiras. Para a realização do estudo foi utilizada apenas a tabela escola, porque nesse primeiro momento queremos analisar os fatores intraescolares influentes, ou seja, investigar quais componentes que fazem parte da estrutura escolar possuem alguma influência no desempenho do estudante no Enem.

A tabela escola possui mais de 160 variáveis com mais de 100 mil registros, que abrange a Educação Básica em seus diferentes níveis – Educação Infantil, Ensino Fundamental e Ensino Médio – e modalidades – Ensino Regular, Educação Especial e Educação de Jovens e Adultos, entre escolas públicas e privadas de todo o país.

A análise baseou-se apenas nos registros de escolas públicas do ensino médio, presentes no estado do Pará, contendo 8078 registros. Para a realização do trabalho realizou-se a remoção de variáveis consideradas redundantes, e todas que estavam relacionadas a necessidades especiais. No fim foram selecionadas 45 variáveis como mostra a Tabela 2.

Tabela 2: Tabela com as variáveis selecionadas da base de dados do Censo Escolar 2016 e sua descrição.

Variáveis	Descrição
IN_LOCAL_FUNC_PREDIO_ESCOLAR	Local de funcionamento da escola - Prédio Escolar
TP_OCUPACAO_PREDIO_ESCOLAR	Forma de ocupação do prédio escolar
IN_AGUA_FILTRADA	Água consumida pelos alunos na escola passa por um processo de filtragem

IN_AGUA_INEXISTENTE	Abastecimento de água - Inexistente
IN_ENERGIA_REDE_PUBLICA	Abastecimento de energia elétrica - Rede pública
IN_ENERGIA_INEXISTENTE	Abastecimento de energia elétrica – Inexistente
IN_ESGOTO_INEXISTENTE	Esgoto sanitário – Inexistente
IN_LIXO_COLETA_PERIODICA	Destinação do lixo - Coleta periódica
IN_SALA_DIRETORIA	Dependências existentes na escola - Sala de diretoria
IN_SALA_PROFESSOR	Dependências existentes na escola - Sala de professores
IN_LABORATORIO_INFORMATICA	Dependências existentes na escola - Laboratório de informática
IN_LABORATORIO_CIENCIAS	Dependências existentes na escola - Laboratório de ciências
IN_SALA_ATENDIMENTO_ESPECIAL	Dependências existentes na escola: Sala de recursos multifuncionais para Atendimento Educacional Especializado (AEE)
IN_QUADRA_ESPORTES	Dependências existentes na escola - Quadra de esportes coberta ou descoberta
IN_COZINHA	Dependências existentes na escola – Cozinha
IN_BIBLIOTECA_SALA_LEITURA	Dependências existentes na escola - Biblioteca e/ou Sala de leitura
IN_BANHEIRO_DENTRO_PREDIO	Dependências existentes na escola - Banheiro dentro do prédio
IN_DEPENDENCIAS_PNE	Dependências e vias adequadas a alunos com deficiência ou mobilidade reduzida
IN_SECRETARIA	Sala de secretaria
IN_REFEITORIO	Refeitório
IN_DESPENSA	Dependências existentes na escola – Despensa
IN_AUDITORIO	Auditório
IN_AREA_VERDE	Área Verde
IN_EQUIP_TV	Aparelho de televisão
IN_EQUIP_VIDEOCASSETE	Equipamentos existentes na escola – Videocassete

IN_EQUIP_DVD	Equipamentos existentes na escola – DVD
IN_EQUIP_PARABOLICA	Equipamentos existentes na escola - Antena parabólica
IN_EQUIP_RETROPROJETOR	Retroprojetor
IN_EQUIP_SOM	Se na escola tem Som
IN_EQUIP_MULTIMIDIA	Projetor multimídia (Datashow)
IN_EQUIP_FOTO	Máquina fotográfica/ Filmadora
IN_COMPUTADOR	Se a escola tem Computador
IN_INTERNET	Acesso à Internet
IN_BANDA_LARGA	Internet Banda Larga
IN_ALIMENTACAO	Alimentação escolar para os alunos
TP_ATIVIDADE_COMPLEMENTAR	Atividade Complementar
IN_MEDIACAO_PRESENCIAL	Mediação didático-pedagógica oferecida pela escola – Presencial
IN_MEDIACAO_SEMIPRESENCIAL	Mediação didático-pedagógica oferecida pela escola – Semipresencial
IN_MEDIACAO_EAD	Mediação didático-pedagógica oferecida pela escola – Educação a Distância – EaD
IN_REGULAR	Modo, maneira ou metodologia de ensino correspondente às turmas com etapas de escolarização consecutivas, Creche ao Ensino Médio.

Fonte: Autor,2019.

4.4. Pré-Processamento

Nessa etapa ocorreu a eliminação de dados que foram considerados redundantes e incompletos (campos vazios) que acabam influenciando no processamento do algoritmo fazendo com que não consigam executar, essa etapa foi realizada na base do ENEM com dados de todos os alunos do Brasil que foi dividida por região e depois na base originada da junção Enem e Censo Escolar. A base oriunda da junção de Enem e Censo Escolar foi realizada com o intuito de inferir quais as variáveis da estrutura escolar teriam alguma influência nas notas obtidas pelos alunos, analisando apenas as escolas públicas no estado do Pará.

Realizou-se também a discretização dos dados contínuos (numéricos) que a base apresentava reduzindo o espaço amostral de valores possíveis, agrupando-os em intervalos facilitando a sua aplicação e compreensão de comportamento.

A técnica do PCA também foi utilizada nessa etapa, fazendo uso do software WEKA, com o intuito de reduzir o número de variáveis que ainda se encontrava muito elevado influenciando no desempenho final das RBs obtidas.

4.4.1. Pré-Processamento ENEM

Na fase do pré-processamento foi realizada a limpeza da base, removendo os registros dos candidatos faltosos, os treineiros, daqueles que não completaram o questionário socioeconômico e os dados ausentes. Desta forma, a base de dados foi reduzida a 1.101.136 registros.

Buscando reduzir o número de atributos na composição da rede, ajudando a facilitar a interpretação dos resultados, tendo em vista que muitos nós, dificultam a interpretação, aplicou-se a técnica do PCA, utilizando o software WEKA, foi realizado um PCA em cada região. Essa técnica foi utilizada uma vez que a base apresentava um total das 43 variáveis, para que a interpretação dos resultados não fosse prejudicada pela dimensão dos fatores considerados.

A Tabela 3 mostra a quantidade de instâncias analisadas, a porcentagem da variância do grupo de variáveis que ficaram em primeiro no ranking de atributos e a lista detalhada dos atributos selecionados em cada região. Foi determinado no Weka para selecionar um conjunto de 15 variáveis, fazendo com que os percentuais de variância ficassem em um pouco mais de 83%.

Tabela 3: Tabela com o percentual do grupo de variáveis e as variáveis de cada região.

Regiões	**Norte**	**Nordeste**	**Centro-Oeste**	**Sudeste**	**Sul**
Instancias	113240	308121	110556	399277	169942
Percentual	0.8226 %	0.8077 %	0.8106 %	0.8204 %	0.8312 %
Variáveis	Q42	Q042	Q042	Q042	Q042
	Q047	Q047	Q047	Q047	Q047
	Q006	Q006	Q006	Q006	Q006
	TP_DEPENDENCIA_ADM_ESC	TP_DEPENDENCIA_ADM_ESC	TP_DEPENDENCIA_ADM_ESC	TP_DEPENDENCIA_ADM_ESC	TP_DEPENDENCIA_ADM_ESC

	Q010	Q010	Q010	Q010	Q010
	Q024	Q024	Q024	Q024	Q024
	Q025	Q025	Q025	Q025	Q025
	Q002	Q002	Q002	Q002	Q002
	Q004	Q004	Q004	Q004	Q004
	Q049	Q049	Q049	Q049	Q049
	Q013	Q013	Q013	Q013	Q013
	Q001	Q001	Q001	Q001	Q001
	Q045	Q045	Q045	Q045	Q045
	Q044	Q044	Q048	Q050	Q050
	Q048	**COR_RACA**	**Q050**	COR_RACA	Q044

Fonte: Autor,2019

O conjunto de variáveis iniciais na base do Enem foi reduzido a um subconjunto de atributos mutuamente relevantes. Sendo assim a quantidade de variáveis diminuiu para 22, esse número se dá pela soma das variáveis selecionadas no PCA, variáveis presentes na tabela 3, mais as variáveis que são referentes as notas obtidas pelos estudantes no exame.

4.4.2. Pré-Processamento Censo Escolar

Antes de iniciar essa etapa realizou-se a junção da base do ENEM com a do Censo Escolar, com o objetivo de obter mais informações a respeito do aluno, podendo avaliar quais componentes da estrutura escolar estão ligadas ao seu desempenho no exame no ENEM. Essa junção foi viabilizada devido as duas bases terem uma mesma variável em comum, que no caso é o CO_ENTIDADE (código da escola). Com essa junção a quantidade de registros diminuiu para 6700.

Logo após o mesmo pré-processamento foi realizado na base do Censo Escolar, removendo os registros de dados ausentes, diminuindo a quantidade de registros para 6.300. Nessa base também foi preciso diminuir o número de atributos fazendo uso novamente do PCA. A tabela 4 abaixo mostra o conjunto de atributos e o percentual de variância, o conjunto que obteve o melhor percentual de variância foi o que continha 9 atributos e mostrou que esse conjunto atingiu um percentual de variância de 0,9099%.

Essa diminuição resultou em economia de tempo e de recursos no processo de mineração que é aplicada nessa mesma base de dados, sem perda significativa de informação.

Tabela 4: Variáveis do Censo Escolar que foram selecionadas no PCA.

Percentual: 0.9099 %
VARIÁVEIS
BIBLIOTECA_SALA_LEITURA
EQUIP_RETROPROJETOR
EQUIPAMENTO_MULTIMIDIA
EQUIPAMENTO_TV
IN_COMPUTADOR
IN_INTERNET
LABORATORIO_CIENCIAS
LABORATORIO_INFORMATICA
QUADRA_ESPORTES

Fonte: Autor,2019.

4.5. Transformação

Com o objetivo de usar os dados, já presentes na base, como entrada para os algoritmos de mineração, é preciso realizar a etapa de transformação dos dados, para que os mesmos fiquem adequados para esses algoritmos e melhorar o desempenho. Dessa forma após todas as etapas de pré-processamento dos dados, o próximo passo realizado foi a transformações dos dados. Uma das principais tarefas realizadas nessa etapa de transformações dos dados foi à categorização, pois a partir da mesma se torna possível categorizar as variáveis fazendo com que os valores da mesma, fiquem divididos em categorias diminuindo então a amplitude desses valores, podendo dessa forma fazer com que os algoritmos de mineração de dados tenham resultados melhores.

As primeiras variáveis a passar pelo processo de categorização foram os atributos que contém as notas obtidas pelos alunos que fizeram o ENEM, os atributos são “NU_NOTA_REDACAO”, “NU_NOTA_CN”, “NU_NOTA_CH”, “NU_NOTA_LC”

e “NU_NOTA_MT”. A categorização dessas variáveis foi realizada de acordo com as faixas de classificação que são realizadas pelo próprio ENEM. As faixas ficaram como mostrada na Tabela 5 e 6 abaixo. Todas as notas em cada área do conhecimento possuem uma escala numérica que vai de 0 a 1000.

Tabela 5: Categorização das notas obtidas pelos estudantes nas provas de LC, MT, CH e CN no ENEM 2016.

Faixas	**Classificação**
N<450	Muito Baixa (MB)
451 a 549,99	Baixa (BA)
550 a 649,99	Regular (REG)
650 a 749,99	Bom
750 a 1000 ou N >= 750	Excelente (EXC)

Fonte: Autor, 2019.

Tabela 6: Categorização das notas obtidas pelos estudantes na prova de redação no ENEM 2016.

Faixas	Classificação
N<500	Muito Baixa (MB)
500 a 599,99	Baixa (BA)
600 a 699,99	Regular (REG)
700 a 799,99	Bom
800 a 1000 ou N >= 800	Excelente (EXC)

Fonte: Autor, 2019.

Para tratarmos a variável Renda familiar (Q006), cujos valores são faixas de renda (e.g. “De R$ 2.364,01 até R$ 3.152,00”), com um grande número de intervalos, e a variável Número de pessoas na residência (Q005), que também possui uma faixa extensa, escolhemos substituir o texto original e agrupar por número de salários mínimos e quantidade de pessoas. As Tabelas 7 e 8 mostram o agrupamento.

Tabela 7: Categorização da renda familiar do estudante que realizou a prova no ENEM 2016.

Faixas	Classificação
R$ 0,00	Nenhuma Renda
Até R$ 880,00	1 Salário
R$ 880,01 a R$ 2640,00	3 salários
R$ 2640,01 a R$ 5280,00	6 salários
R$ 5280,01 a R$ 10560,00	12 salários
R$ 10560,01 a R$ 17600,00	20 Salários

Fonte: Autor, 2019.

Tabela 8: Categorização do número de pessoas na residência do estudante que realizou a prova no ENEM 2016.

Faixas	**Classificação**
Mora sozinho	1
2 a 6 pessoas	2
7 a 11 pessoas	3
12 a 20 pessoas	4

Fonte: Autor, 2019.

Os dados apresentados na base do Censo Escolar são binários onde as respostas para cada variável se limita a 0 (zero) significa “não” e 1 (um) “sim”. Fez-se necessário realizar a categorização dos atributos “NU_NOTA_REDACAO”, “NU_NOTA_CN”, “NU_NOTA_CH”, “NU_NOTA_LC” e “NU_NOTA_MT” para apenas duas faixas.

Para a realização da RB no Censo Escolar, foi necessário tirar a média das notas de cada aluno, esse processo se dá pela soma de todas as notas (quatro objetivas + redação) e a divisão desse valor por 5, uma vez que todas as provas possuem o mesmo peso na distribuição de seus valores. As notas finais são classificadas em duas faixas BOM e RUIM, que é exibido na tabela 9. A faixa de nota foi estabelecida levando em consideração as notas de corte para o ingresso nos cursos de ensino superior da UFPA.

Tabela 9: Categorização da média das notas obtidas pelos estudantes no ENEM 2016 para o Censo.

Faixas	Classificação
N<=459,99	RUIM
551 a 1000	BOM

Fonte: Autor, 2019.

4.6. Mineração de dados

Essa etapa corresponde ao processo de mineração de dados e a busca de informação, nessa etapa foi aplicada a técnica de Redes Bayesianas, sendo utilizando o *software Bayesware Discoverer*, o mesmo foi escolhida por ter uma interface descomplicada e de fácil uso, que proporciona a criação e a geração de redes bayesianas. Este software construiu as redes a partir dos atributos do banco de dados criado, exibindo as tabelas de probabilidade condicional ou incondicional de cada nó (atributo).

O conjunto depois de transformado foi submetido à técnica de Redes Bayesianas, que explorou e analisou os dados identificando padrões, sendo possível também visualizar tabelas de probabilidade condicional ou incondicional de cada nó (atributo). No primeiro momento dessa etapa foram geradas cinco RBs com os dados do ENEM, cada uma correspondendo a uma região do país, que são elas: Região Norte, Região Nordeste, Região Centro-Oeste, Região Sudeste e Região Sul. Podendo assim fazer uma análise de cada região isoladamente, verificando os padrões de influência socioeconômica no desempenho do estudante em cada região traçando um perfil regional e obtendo parâmetros nacionais. Em outro momento obteve-se uma RB para identificação dos padrões da estrutura escolar publica paraense, identificando quais dessas variáveis são influentes nas notas obtidas no ENEM.

4.7. Identificação e Interpretação dos Padrões

Com a realização da mineração de dados, foi iniciada a geração das Redes Bayesianas e os conhecimentos extraídos da própria topologia da rede fazendo uso do *software Bayesware*. Nessa seção são apresentadas as redes geradas dos dois estudos de

caso, detalhados nas seções 5.2 a 5.3. Vale destacar que, a partir das topologias apresentadas, interpretações são possíveis e válidas no contexto pesquisado.

Nessa fase os padrões extraídos foram sujeitos a interpretações e avaliações determinando a sua qualidade e utilidade, sendo informações úteis para gestores, professores e pesquisadores envolvidos com a temática da educação.

4.8. Considerações Finais

Este capítulo teve como objetivo demonstrar toda a metodologia necessária para realizar as etapas do conhecimento nas bases de dados objetos estudo e como se deu a aplicação da técnica de mineração aplicada ao domínio escolhido.

5. Resultados

5.1.Considerações Iniciais

Neste capítulo serão expostos os resultados obtidos a partir da aplicação da metodologia proposta nos dos dados educacionais, os microdados do ENEM e Censo Escolar, dados esses que são fornecidos pelo INEP.

Para a realização dessa pesquisa foram realizados dois estudos de casos: no primeiro momento foi utilizada a técnica das redes Bayesianas apenas na base de dados do ENEM 2016 com o intuito de identificar as variáveis socioeconômicas de cada aluno brasileiro influentes no seu rendimento na realização dessa prova, fazendo uma análise a nível nacional. O próximo estudo seguiu os mesmos parâmetros que o estudo anterior, no entanto objetivou investigar quais variáveis escolares possuem ligação com o seu rendimento no final do ensino médio, para tal foram analisadas as escolas do ensino público do estado do Pará, para tal realizou a junção da base ENEM e Censo Escolar.

5.2. Estudo de Caso I

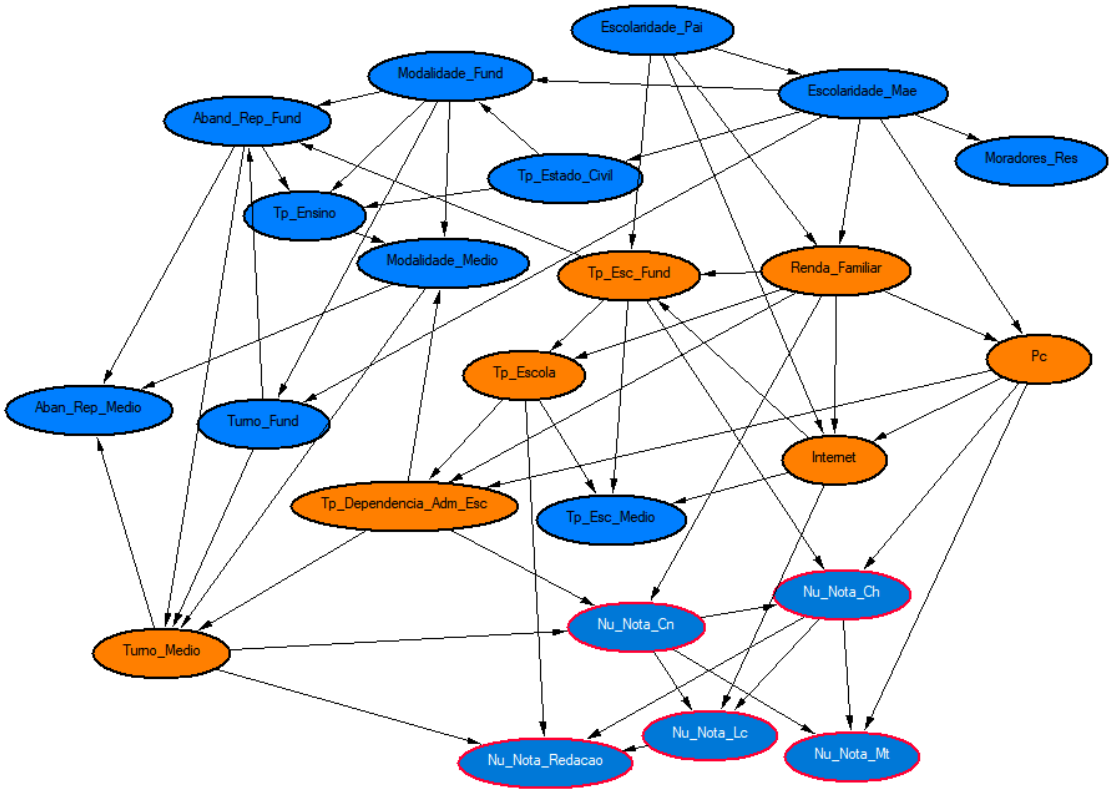
O primeiro estudo de caso evidencia a análise das Redes Bayesianas obtidas após sua aplicação na base do ENEM 2016. Foram geradas cinco redes cada uma correspondendo a uma região do país, que são elas: Região Norte, Região Nordeste, Região Centro-Oeste, Região Sudeste e Região Sul. Podendo assim fazer uma análise de cada região isoladamente e posteriormente comparar quais variáveis influenciam no desempenho do aluno por região e áreas de conhecimento.

As redes foram geradas com os 22 atributos mais relevantes, que foram definidos no pré-processamento (seção 4.3.1). As cores dos nós então diferentes com o propósito de facilitar a identificação e visualização das variáveis. Os elementos em laranja são as variáveis que estão ligadas diretamente as notas que por sua vez estão com as bordas na cor vermelha, como é ilustrado na Figura 3.

A RB resultante da região Norte, exibida na figura 5, evidencia que a região possui sete variáveis influentes em suas notas, são elas: Turno do Ensino Médio, Tipo de Dependência Administrativa escolar, Tipo de Escola, Tipo de Escola do Ensino Fundamental, Renda Familiar e se possui computador e Internet em sua residência.

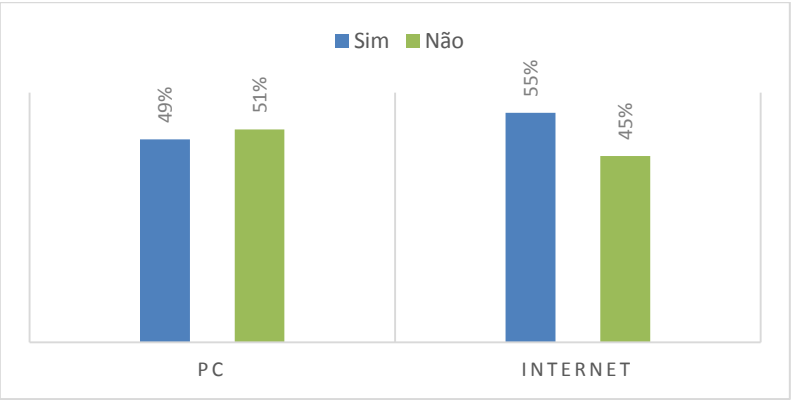
O comportamento das variáveis referente a acesso à internet e possuir computador em sua residência e demostrado no gráfico 4, onde 45% dos alunos da região Norte não tem acesso à internet e que 51% dos estudantes não tem computador na residência.

Figura 5: Rede Bayesiana da Região Norte.



Fonte: Autor, 2019.

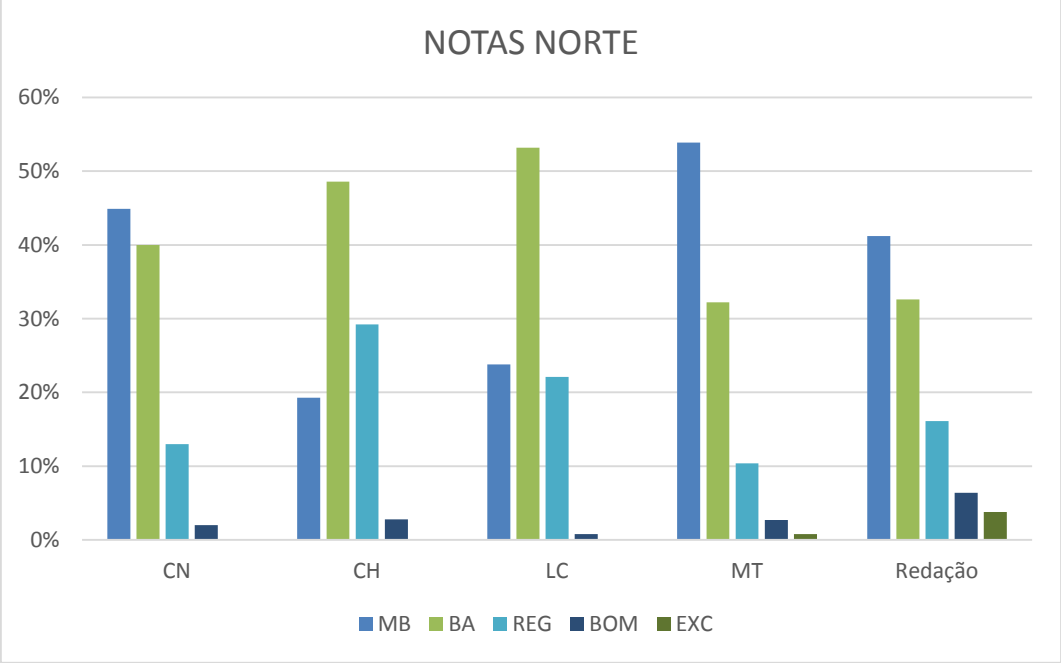
Gráfico 4 – Quantidade de alunos com e sem Internet e Computador.



Fonte: Autor, 2019.

O gráfico 5 mostra a concentração das notas dos alunos dessa região em cada área do conhecimento, tendo uma maior concentração de notas as Muito Baixas (MB) e as Baixas (BA) nas matérias de Linguagens e seus códigos e Matemática, em especial em matemática onde mais da metade dos estudantes, tiraram menos de 450 pontos.

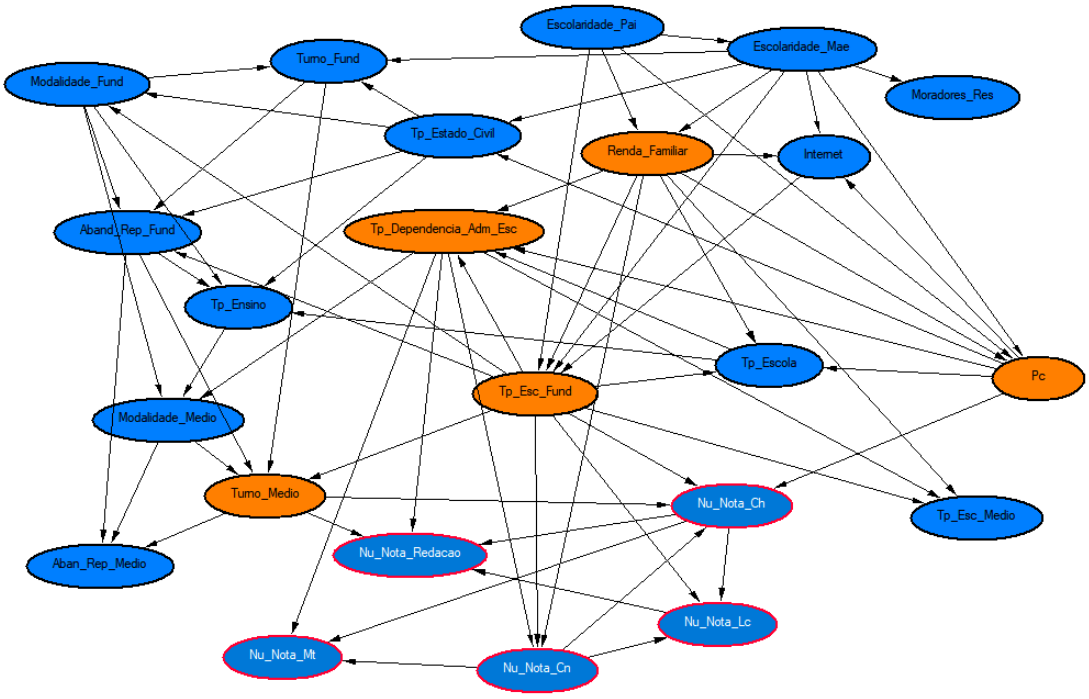
Gráfico 5 – Porcentagem das notas da região Norte em cada área do conhecimento.



Fonte: Autor, 2019.

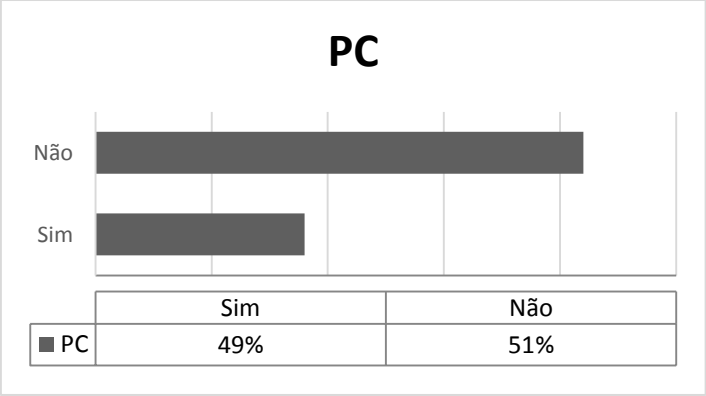
Na rede obtida da região Nordeste apresentada na figura 6 as variáveis influentes são o Tipo de Dependência Administrativa Escolar, Renda Familiar, Tipo de Escola do Ensino Fundamental, Turno do Ensino Médio e se o aluno possui computador. Diferente da região Norte o estudante ter acesso ou não a internet não influencia em nenhuma nota, o gráfico 6 expõem que um pouco mais da metade dos estudantes dessa região não tem computador.

Figura 6: Rede Bayesiana da Região Nordeste.



Fonte: Autor, 2019.

Gráfico 6 – Percentual de alunos com ou sem Computador em sua residência região Nordeste.



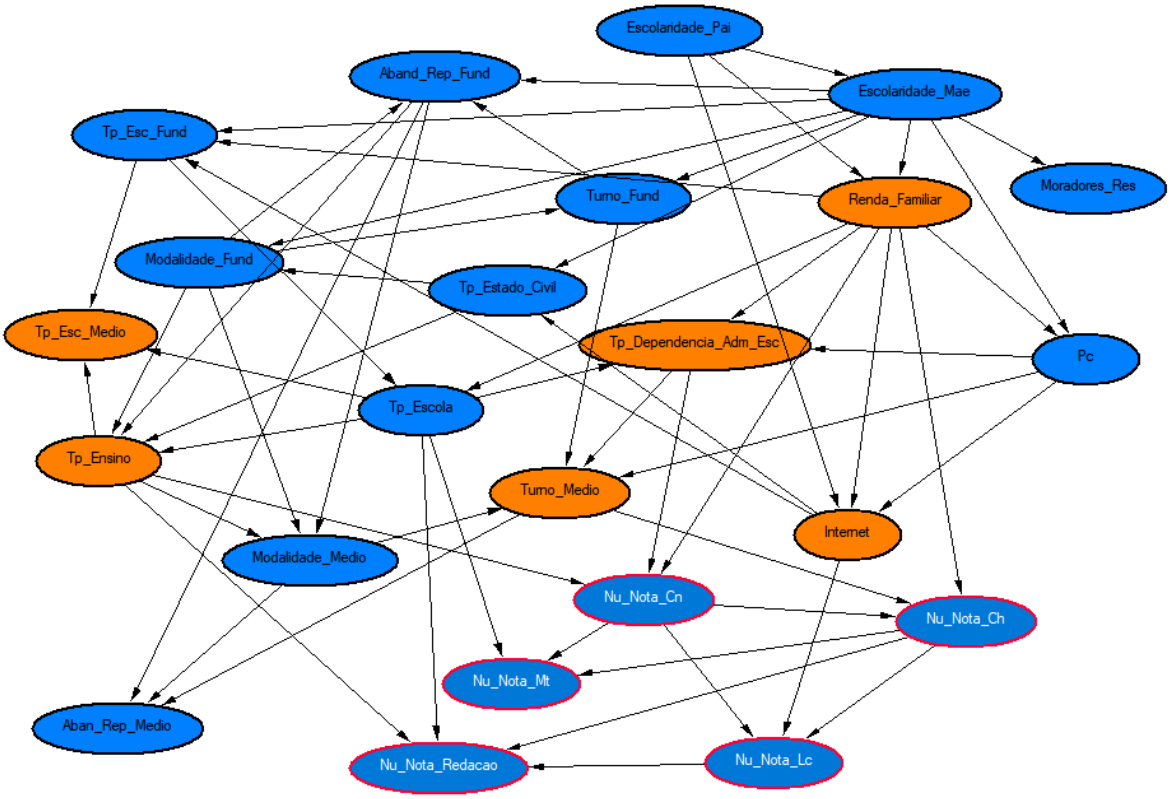
Fonte: Autor, 2019.

As notas no Nordeste concentram-se também entre MB e BA o mesmo comportamento da região Norte. O mesmo comportamento se repete para a região seguinte a Centro-Oeste. Já na região Sudeste e Sul observou-se uma diminuição da porcentagem das notas MB e BA e um aumento principalmente das notas regulares

acima dos 45% ou seja notas até 649,99. Mas não divergem muito das regiões anteriores consideradas menos desenvolvidas. Os gráficos com as concentrações das notas podem ser analisados nos apêndices.

Na figura 7 a RB da região Centro-Oeste evidencia que a Renda Familiar, Turno do Ensino Médio, Tipo de Ensino, Tipo de Dependência Administrativa, Tipo de Escola e se o aluno tem acesso à Internet.

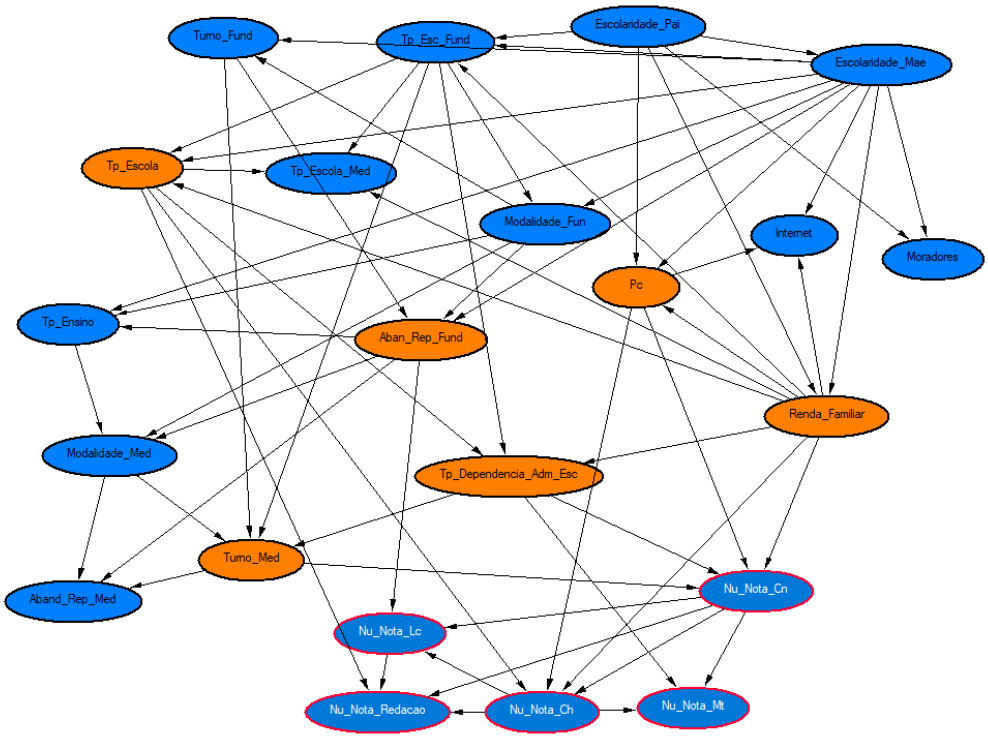
Figura 7: Rede Bayesiana da Região Centro-Oeste.



Fonte: Autor, 2019.

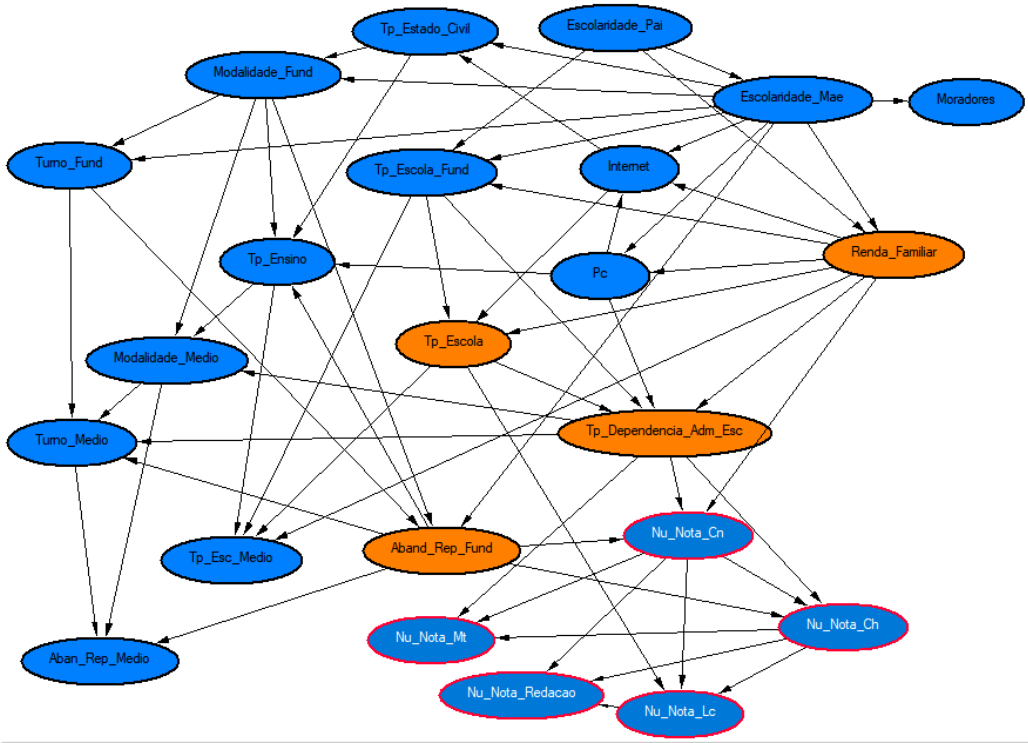
Na região Sudeste o Tipo Escola, se o Aluno Abandonou ou Não o Ensino Fundamental, o Tipo de Dependência Administrativa, Turno do ensino Médio, Renda Familiar e se tem acesso ou não a Internet em casa são as variáveis que a rede Bayesiana gerada apontou como influentes, na figura 8.

Figura 8: Rede Bayesiana da Região Sudeste.



Fonte: Autor, 2019.

Figura 9: Rede Bayesiana da Região Sul.



Fonte: Autor, 2019.

A figura 9 exibe a rede Bayesiana da última região a Sul, que demonstra ter apenas quatro variáveis influentes, entre elas está a Renda Familiar, o Tipo de Dependência Administrativa Escolar, se o aluno Abandonou ou Não o ensino Fundamental e o Tipo de Escola que ele frequentou.

A partir das redes apresentadas percebe-se que cada região possui um conjunto de variáveis influentes distinto, contendo ou não algumas variáveis em comum ligadas ao desempenho dos estudantes de cada região. A tabela 10, lista as variáveis de cada região, das variáveis influentes, que são as que apresentaram ligação direta aos nós das notas em cada região, destacam-se as de Tipo de Dependência Administrativa Escolar, variável que classifica se a escola possui administração municipal, estadual, federal e privada e a variável Renda Familiar, que aparecem em todas as regiões. Outro ponto a ser destacado é a não influencia dos atributos de acesso à internet e se o aluno possui ou não computador em sua residência, não influenciarem na região Sul, diferente das 4 regiões anteriores.

Tabela 10 – Tabela que contem a listagem de todas as variáveis influentes em cada região.

REGIÃO / VARIAVEIS	Norte	Nordeste	Centro-Oeste	Sudeste	Sul
Abandonou ou Reprovou Ens. Fundamental	X			X	X
Turno Ens. Médio	X	X	X	X	
Tipo de Administração Escolar	**X**	**X**	**X**	**X**	**X**
Tipo escola Ens. Fundamental	X	X			
Acesso à Internet	X		X		
Renda Familiar	**X**	**X**	**X**	**X**	**X**
Possui Computador	X	X		X	
Tipo de Escola Ens. Médio	X		X	X	X
Tipo de Ensino			X		

Fonte: Autor, 2019.

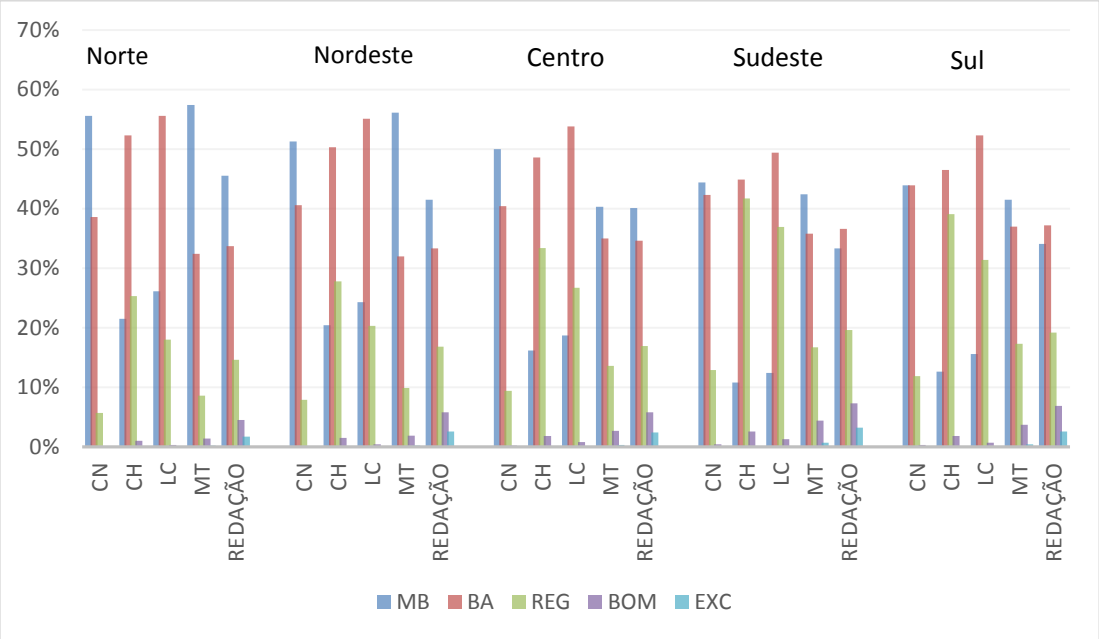
5.2.1.1.Inferências ENEM 2016

Nessa seção foi evidenciada e analisada o comportamento das notas quando realizado o processor de inferência em três variáveis, o tipo de Dependência Administrativa Escolar, acesso à Internet e se o estudante possui computador em sua residência.

— **Tipo de Dependência Administrativa Escolar**

Realizando as inferências em cima da variável TP_ADM_ESC podemos novamente analisar o comportamento das notas, em cada região. Os gráficos 7, 8 e 9 apresentam o comportamento de cada área do conhecimento quando assumimos que 100% dos alunos são de escola da rede estadual, privada e federal, respectivamente.

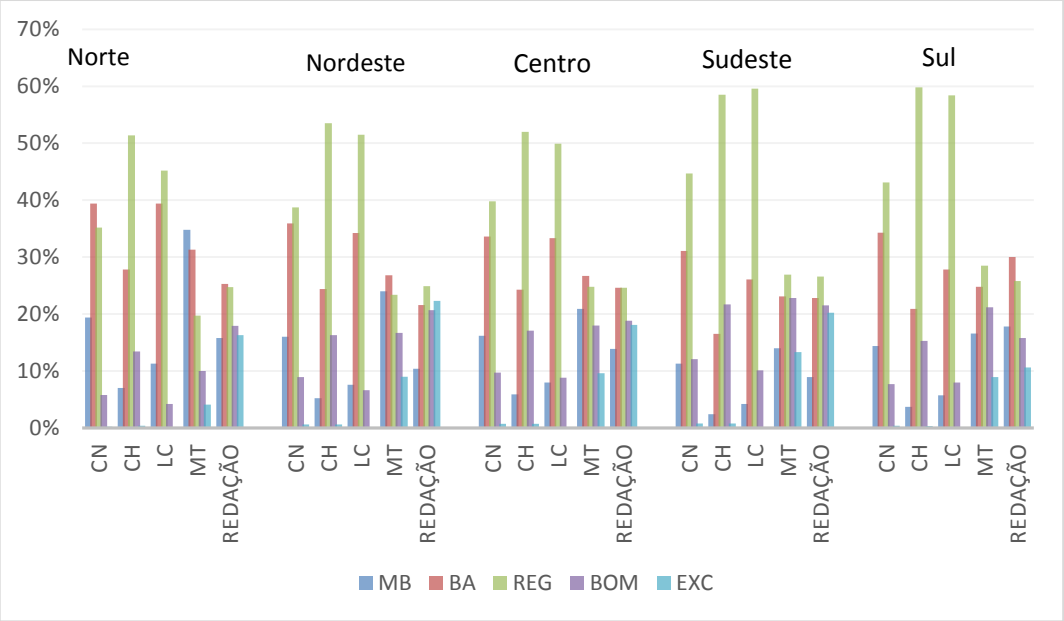
Gráfico 7 – Comportamento das notas com inferência dos alunos da rede Estadual.



Fonte: Autor, 2019.

Assumindo como todos os alunos da rede estadual as notas continuam se concentrando em Muito Baixa e Baixa em todas as regiões, tendo maior probabilidades os alunos da região Norte e Nordeste com percentuais acima de 50% em todas as áreas do conhecimento.

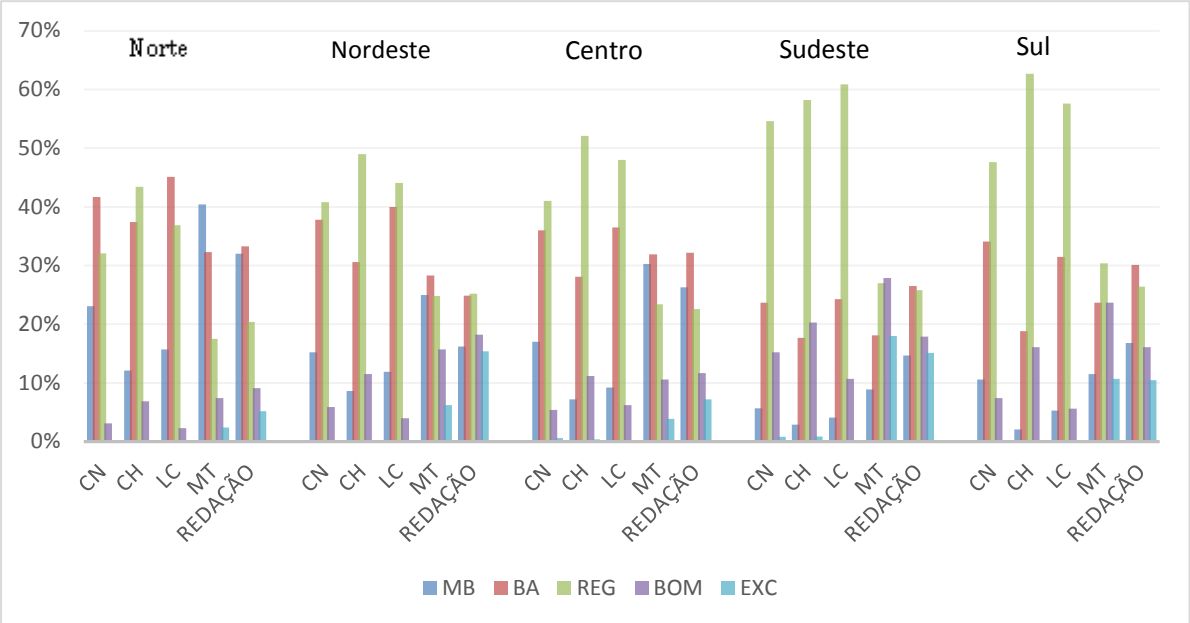
Gráfico 8 – Comportamento das notas com inferência dos alunos da rede Privada.



Fonte: Autor, 2019.

Quando evidenciamos alunos da rede privada o cenário das notas sofre uma alteração e passam a ter uma concentração de probabilidades para REG, BOM e EXC mostrando que tende a ter um melhor desempenho que os da rede estadual. Sendo que nas regiões Sudeste e Sul mais de 60% dos estudantes tiraram notas regulares.

Gráfico 9 – Comportamento das notas com inferência dos alunos da rede Federal.



Fonte: Autor, 2019.

Evidenciando a rede Federal, o cenário de concentração das notas tem um panorama ainda melhor, com uma concentração de notas consideradas boas superando os 60%. Essa evidencia é melhor visualizada quando são analisadas as regiões Centro-Oeste, Sudeste e Sul, em contra partida na região Norte e Nordeste esse percentual não é tão expressivo a diferença deve-se pela diminuição de tendências a notas MB e BA.

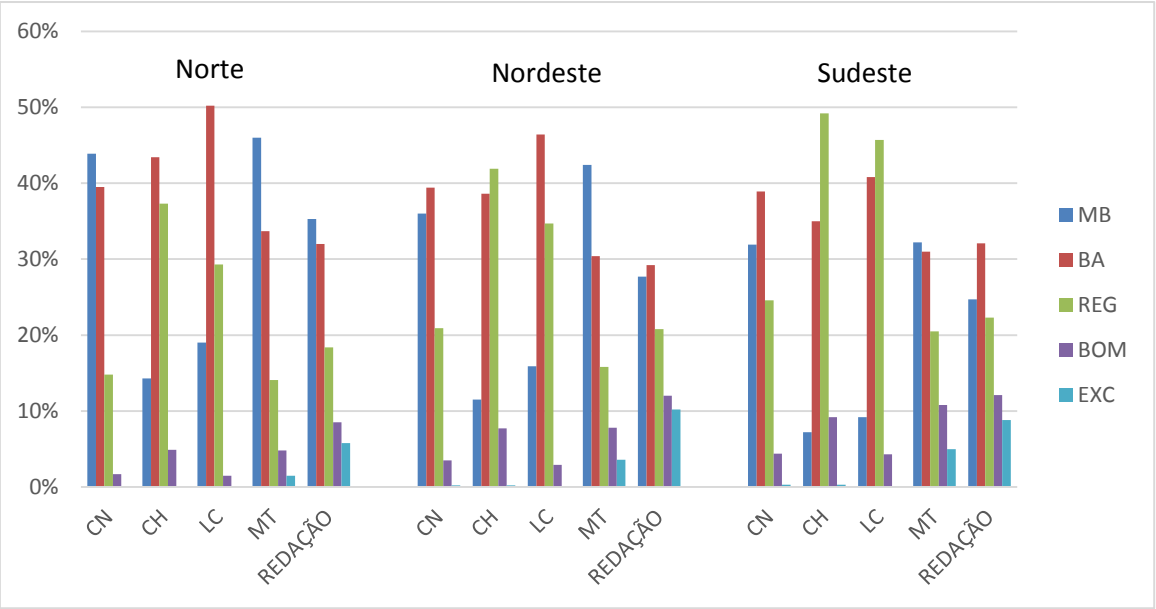
Realizando a inferência no Tipo de dependência administrativa escolar é possível fazer uma análise um pouco mais ampla em relação a variável Tipo de escola que se limita a classificar apenas se o estudante estudou em escola pública e particular. Nessa analise pode-se identificar que os estudantes das escolas federais possuem um desempenho superior em relação aos estudantes que frequentaram escolas da rede privada e principalmente em relação aos alunos da rede estadual que apresentaram os piores índices das notas.

— **Possui ou não Computador na residência**

O processo de inferência referente a computador foi realizado apenas nas regiões que tiveram essa variável como influente nas notas dos estudantes do ensino médio. Considerando que todos os estudantes possuem computador em sua residência as notas da região Norte, Nordeste e Sudeste tendem a percentuais menores nas notas MB e BA.

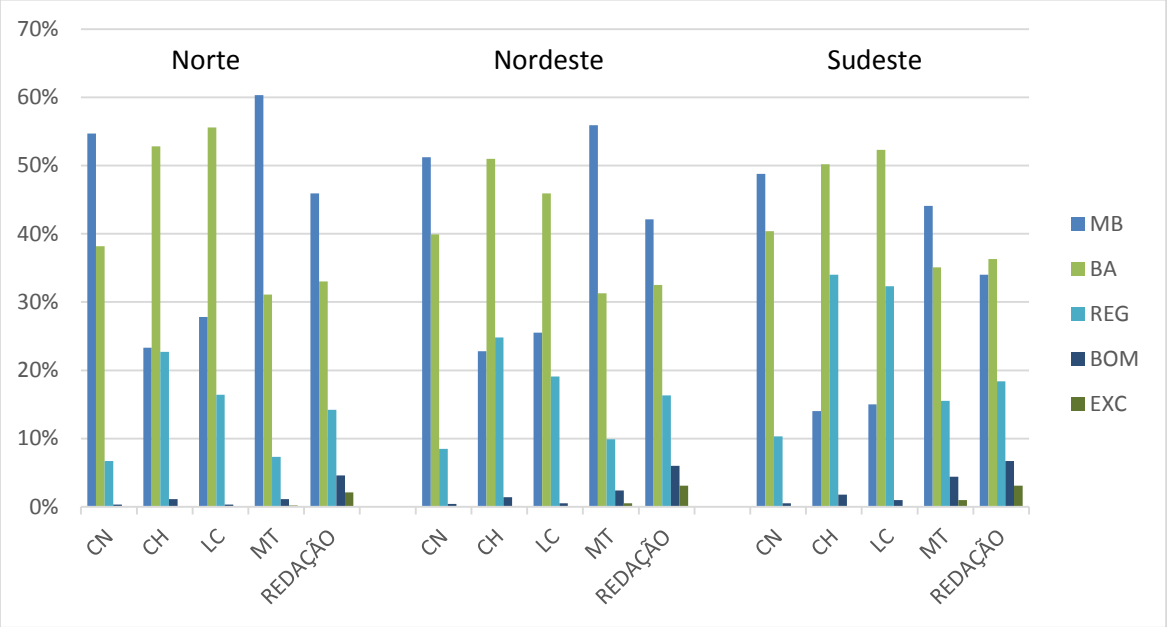
Comportamento diferente é exibido quando evidenciamos os que não possuem computador chegando a ter 60% de probabilidade de ter nota MB na região Norte e percentuais acima de 50% na região Nordeste. Mesmo a diferença não sendo tão expressiva é possível mostrar que há uma diferença, como mostra o gráfico 10 e 11, além de demonstrar o comportamento de cada área de conhecimento avaliado na prova.

Gráfico 10 – Comportamento das notas com inferência dos alunos que possuem computador.



Fonte: Autor, 2019.

Gráfico 11 – Comportamento das notas com inferência dos alunos que não possuem computador.



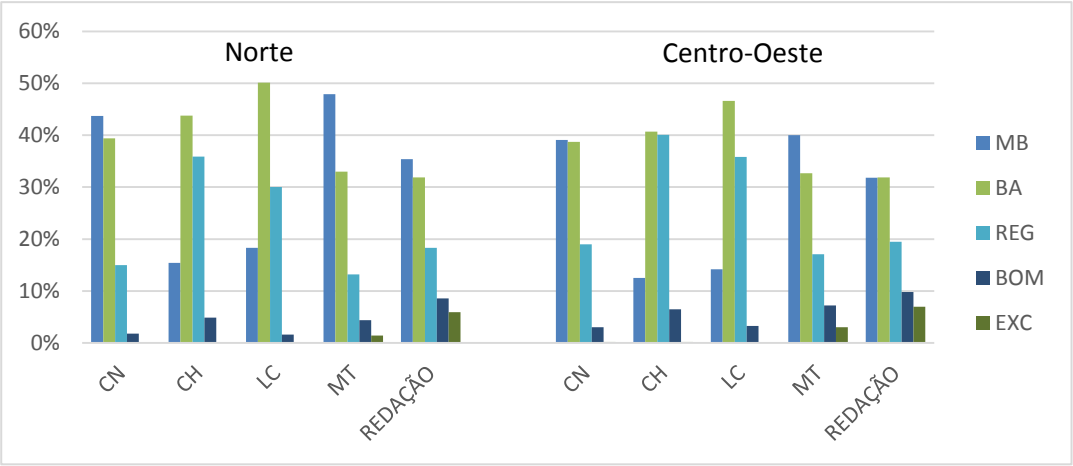
Fonte: Autor, 2019.

— Acesso à Internet

Por meio do questionário do ENEM, não ficou esclarecido se o acesso à internet deveria ser exclusivo através do computador. Assim, pode se supor que ao cesso através de tablets e smartphones podem ser considerados como dispositivos de acesso à internet. Os gráficos 12 e 13 demonstram o comportamento das notas dos alunos das

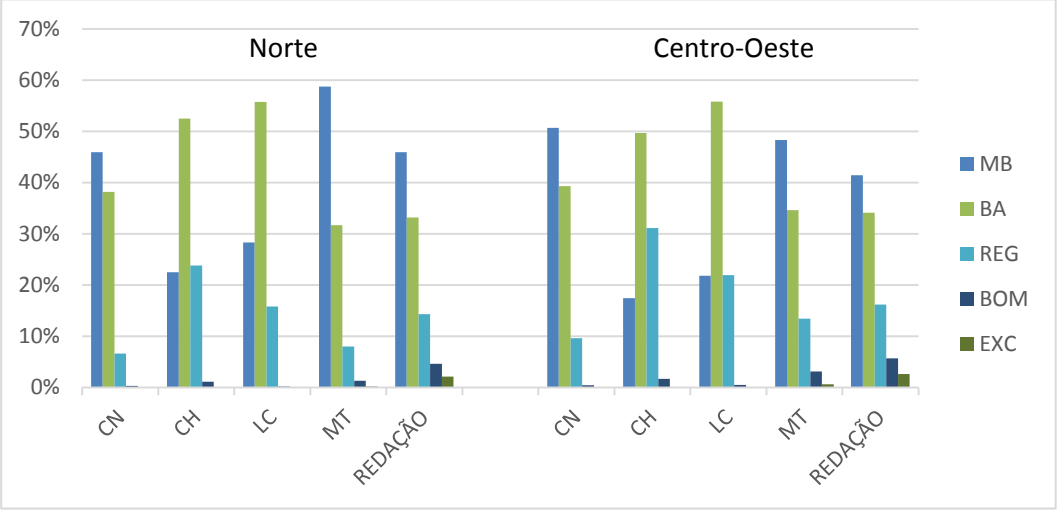
regiões Norte e Centro-Oeste com e sem acesso à internet. As inferências são realizadas apenas com as duas regiões que tiveram essa variável como influência direta nas notas.

Gráfico 12 – Comportamento das notas com inferência dos alunos que tem acesso à internet.



Fonte: Autor, 2019.

Gráfico 13 – Comportamento das notas com inferência dos alunos que não tem acesso à internet.



Fonte: Autor, 2019.

Podemos verificar que não se tem uma mudança tão expressiva na concentração das notas, mas percentuais de diferença são observados. Assumindo que todos os alunos não têm acesso a internet as notas das mais baixas até as mais altas teve o seguinte comportamento: 38%, 42%, 16%, 2% e 1%. Em comparação aos que tem acesso 30%, 39%, 24%, 5% e 2%, mostrando uma significativa diminuição das notas baixas e um aumento nas notas mais altas.

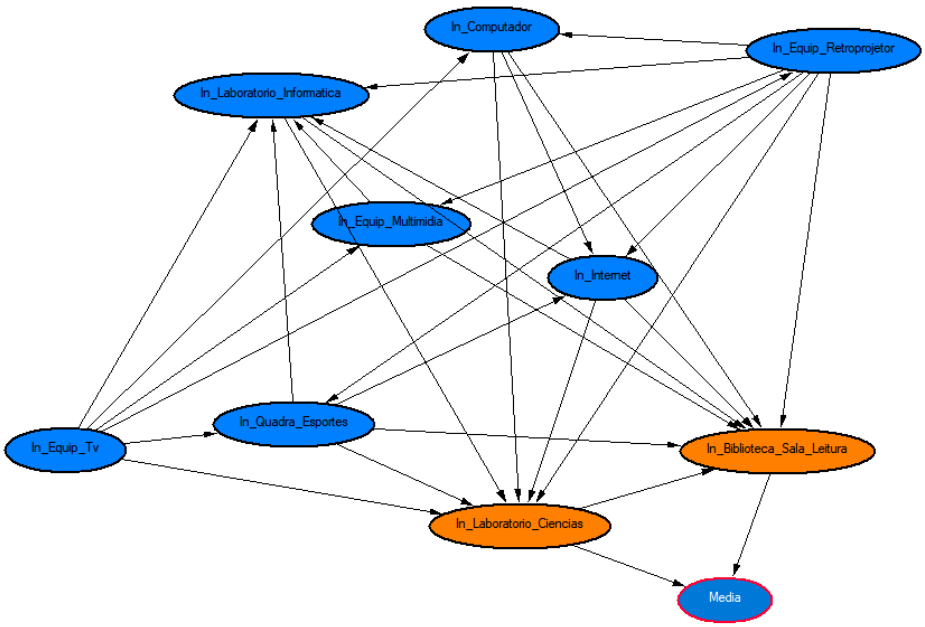
5.2.2. Estudo de Caso II

Esse estudo analisou os atributos relacionados as condições de infraestrutura escolar ofertada pelas escolas públicas (estaduais) de ensino médio no estado do Pará. Esta base detalha as condições das escolas secundárias correspondente a cada aluno que participou da prova do ENEM 2016. Para tal realizou-se a junção das bases de dados do ENEM com a do Censo Escolar, foram integradas pelo código único de identificação da que se faz presente nas duas bases.

A rede executada com o auxílio do algoritmo K2 é mostrada na figura 10. Na rede obtida contém os 10 atributos selecionados através da técnica do PCA. Infere-se da rede gerada a influência dos atributos Laboratório de Ciências e se a escola possui biblioteca e sala de leitura.

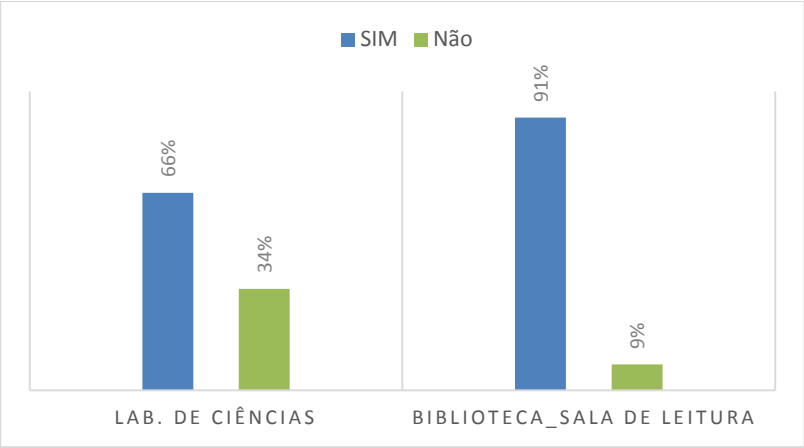
O gráfico 14 e 15 exibem os percentuais de escolas que possuem as variáveis identificadas como influentes nas escolas públicas do Pará, nos revelando que a maioria das escolas públicas possuem laboratório de Ciências e biblioteca com sala de leitura, e mesmo assim a maioria dos alunos obtiveram um desempenho classificado como ruim, ou seja, a maioria tirou menos de 550 pontos na média das notas obtidas no ENEM 2016. Essa primeira analise nos faz questionar, se o aluno possui uma estrutura por que ainda continua tendo um desempenho ruim?

Figura 10: Rede Bayesiana gerada com da base Censo ENEM.



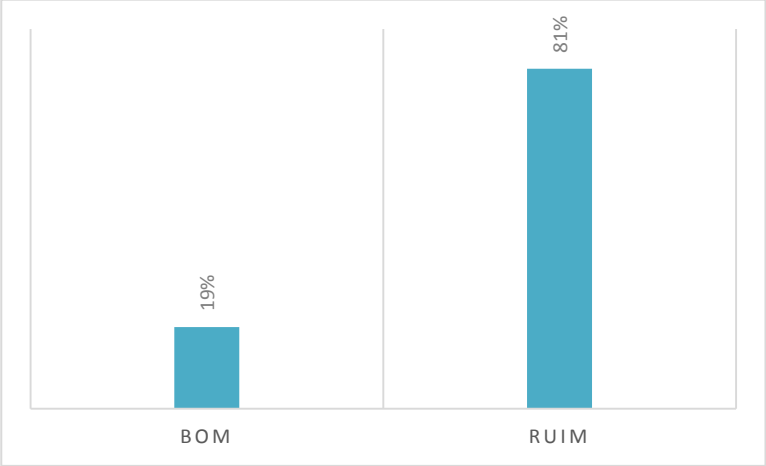
Fonte: Autor, 2019.

Gráfico 14: Gráfico com informações a respeito da porcentagem de alunos com escolas que tem Lab. de Ciências e Bibliotecas com salas de leitura.



Fonte: Autor, 2019.

Gráfico 15: Gráfico referente a classificação das notas da escola pública do Pará.



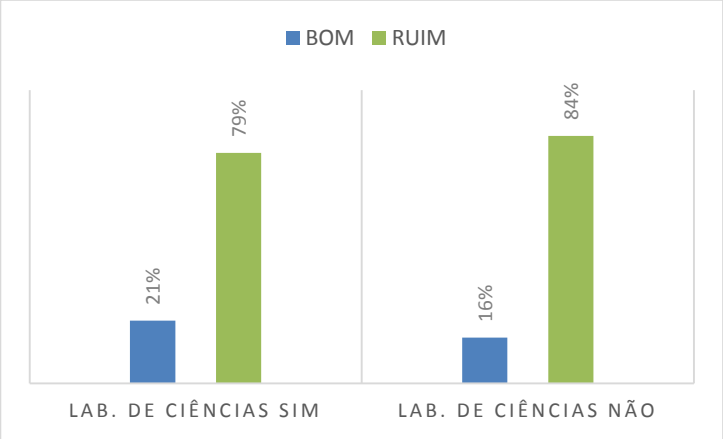
Fonte: Autor, 2019.

5.7.2.1 Inferências Censo Escolar e ENEM 2016

O processo de inferência nessa rede se deu nas duas variáveis que estão diretamente ligadas à média de notas obtidas pelos alunos que são de escola pública do estado do Pará. O primeiro processo de inferência relaciona a questão de o aluno ter acesso a laboratório de ciências ou não, quando assumimos que todos dos alunos têm acesso, o percentual de notas segue o seguinte padrão 21% e 79%. Quando evidenciamos que não

tem o comportamento muda minimamente diminuindo em 5% a porcentagem de notas boas e aumento em notas ruins, como mostra o gráfico 16.

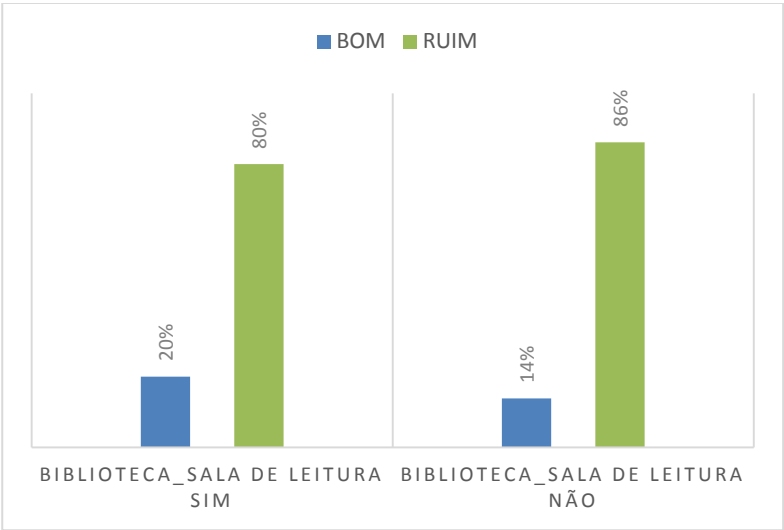
Gráfico 16: Gráfico das inferências de ter ou não acesso a Laboratório de Ciências.



Fonte: Autor, 2019.

O segundo processo de inferência relacionasse ao aluno ter na escola que frequentou biblioteca e sala de leitura, como foi mostrado no gráfico 15, 91% dos estudantes da escola pública do Pará teve durante seu estudo biblioteca e sala de leitura na escola em que frequentou. O gráfico 17 expõe as inferências de ter ou não em sua escola biblioteca e sala de leitura, apontando uma diferença de 6% para mais na probabilidade de notas boas e uma diminuição de notas ruins quando assumimos que sim.

Gráfico 17: Gráfico das porcentagens das inferências de ter ou não biblioteca e sala de leitura.



Fonte: Autor,2019.

Uma das dimensões teoricamente relevantes para uma possível explicação desse desempenho, diz respeito às dinâmicas que ocorrem cotidianamente nas salas de aula, na gestão da classe e do conteúdo e cobertura das disciplinas, assim como nas diferenças das características dos alunos dentro das turmas e do ambiente da classe.

5.3. Considerações Finais

Este capítulo foi essencial à exposição dos resultados obtidos com aplicação da metodologia proposta, expondo aplicação de um algoritmo que auxiliou no processo de pré-seleção dos dados e evidenciou dois estudos de caso nos quais são abordados os padrões extraídos após o processo de inferência bayesiana. Em suma, os resultados foram suficientes aos desígnios deste trabalho, pois foi permitido compreender o relacionamento entre a problemática e as variáveis, sob a premissa do uso de uma metodologia objetiva, eficiente e de fácil entendimento ao usuário inserido no domínio.

6. Conclusões

Este trabalho teve como objetivo identificar quais as variáveis socioeconômicas e quais estruturas educacionais presentes nas escolas em que os alunos do ensino médio frequentaram tem alguma ligação\influência no seu rendimento escolar com base nas notas obtidas no ENEM e dados fornecidos pelo Censo Escolar de 2016. Proporcionou o entendimento a respeito da problemática dos aspectos socioeconômicos em relação a educação.

A metodologias empregada aos propósitos estabelecidos foi o processo de KDD, fazendo uso das redes bayesianas e do PCA, tendo um grande volume de dados acadêmicos registrado nas bases de dados do INEP. Dessa forma pôde-se depreender novos padrões relativos aos fatores dificultadores ao um bom aprendizado no ensino médio.

Em adição, primou-se pela capacidade de expressar esses padrões de maneira concisa em uma estrutura onde fosse possível também aplicar e, concomitantemente, extrair conhecimento de especialistas. Diante do exposto, priorizou-se a escolha da Rede Bayesiana em razão de atender a essas exigências, além disso é permitida uma representação hierárquica entre variáveis e atribuição de probabilidades a determinado evento por meio de evidências em relação a algo que se sabe (ou pretende descobrir) sobre o domínio.

A grande maioria dos alunos está concentrada na esfera estadual de ensino, ou seja, escola pública, uma parcela em escolas particulares e uma baixa porcentagem em escolas federais. Sendo que nas escolas federais os alunos são os que contém as melhores notas em todas as regiões analisadas.

Pode-se analisar quais variáveis são influentes por região, uma vez que os trabalhos presentes na literatura em relação ao domínio escolhido, apenas realizam analises a nível estadual, a pesquisa evidencia que cada região possui um conjunto de variáveis influentes tendo apenas duas variáveis que se repetem em todas as regiões, o tipo de dependência administrativa escolar e a renda familiar.

Na região Norte, Nordeste e Centro-Oeste pode se observar que fatores a respeito do aluno possuir computador em sua residência e ter acesso a internet influenciam nas suas notas obtidas no exame do Enem, evidenciando que nas regiões consideradas

menos desenvolvidas no país, poucos estudantes têm acesso a esse recurso como auxilio para os seus estudos, recursos esses que atualmente são considerados básicos. Mostrando que acesso à internet e possuir computador em casa como fatores que ajudam a adquirir conhecimento de forma continua, onde todas as dúvidas escolares podem ser esclarecidas pelo fato de que há facilmente acesso a esses recursos.

Os resultados obtidos com a base do Censo Enem exibem possíveis indícios de que existem falhas no sistema educacional em sua missão de educar e que ainda são muitos os alunos que progridem lentamente, mesmo que a escola tenha a estrutura básica oferecida.

A contribuição mais significativa deste trabalho ao domínio de aplicação consistiu na utilidade dos padrões obtidos, importantes ao processo de tomada de decisão por parte de gestores e pesquisadores em educação, sobretudo aqueles empenhados no entendimento de fatores extraescolares que afetam o rendimento educacional do jovem, buscando refletir e pensar em formas de minimizar o impacto dos efeitos das diferenças sociais nos rendimentos dos alunos. Em adição, a aplicação de rede bayesiana em cenários educacionais, especificamente no caso de dados fornecidos pelo INEP, micro dados do ENEM e o Censo Escolar, evidenciou a flexibilidade e robustez dessa metodologia, fato que permitirá uma gama de possibilidades voltadas ao aprimoramento e expansão deste estudo.

6.1. Trabalhos Futuros

Para trabalhos futuros, planeja-se a aplicação da técnica para o restante da base do Censo Escolar realizando mais análises mais ampla a respeito dos fatores intraescolares que influenciam no desempenho do estudante do ensino médio.

Outro passo é o de integrar outras bases que possam servir para descobrir mais fatores ligados a esse índice tão baixo de aprendizagem no setor público paraense. Base que contenham dados sobre índices de violência e saneamento básico no bairro das escolas. Vendo a necessidade de analisar as diferentes interações entre as variáveis, sejam do mesmo nível ou não, intensificando assim a descrição e análise das diferentes dimensões que representam o dia-a-dia dos alunos e, principalmente, permitindo a identificação de variáveis que interferem no aprendizado e, ao mesmo tempo, são

passíveis de intervenção por meio de políticas institucionais da própria escola ou de políticas públicas.

Ainda em relação aos resultados alcançados, outro fator relevante a ser analisado como fator influente entre grande diferença de notas entre as escolas estaduais, federais e privadas evidenciadas na pesquisa seria a formação do corpo docente presente na escola, suas condições de trabalho e sua renda.

Outro trabalho possível é o de analisar os fatores do aluno não ter acesso a internet e não possuir computador tanto na sua residência como na escola, e como essas variáveis influenciam no seu desempenho. Além de possibilitar um estudo a respeito da escolaridade da mãe como fator relevante, uma vez que a pesquisa mostrou ligação dessa variável com a questão do estudante possuir computador em sua residência.

6.2.Dificuldades Encontradas

Vale destacar que a principal dificuldade encontrada para a realização da pesquisa foi a falta de programas computacionais que conseguissem rodar uma base gigantesca inteira, por esse motivo a pesquisa foi realizada com a divisão da base em partes, no caso por regiões.

O pré-processamento dos dados utilizados para a geração das redes bayesianas, que demanda uma grande parcela de tempo. O auxílio de um especialista em educação básica para realizar mais interpretações e analises a respeito das redes geradas uma vez que elas podem nos fornecer mais análises.

Outra dificuldade seriam a existência de poucos trabalhos que fazem uso das bases do Inep e que utilizem as redes bayesianas para realizar análises.

6.3. Publicações

Os resultados obtidos no estudo de caso I foram publicados na:

— **International Conference on Data Science and Management (ICDSM - 2019).** In Association with: Springer. Educational Data Mining: A study on socioeconomic indicators in Education in INEP Database.

REFERÊNCIAS

ADEODATO, Paulo JL; SANTOS FILHO, Maílson M.; RODRIGUES, Rodrigo L. Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2014. p.891.

ALVES, Rafael Damiani. PREDIÇÃO DO DESEMPENHO DA REDAÇÃO DO ENEM UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS. 2018. 67 f. TCC (Graduação) - Curso de Tecnologias da Informação e Comunicação, Universidade Federal de Santa Catarina, Araranguá, 2018.

ANDRIOLA, W. B. Doze motivos favoráveis à adoção do Exame Nacional do Ensino Médio (ENEM) pelas instituições de Ensino Superior (IFES). Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 19, n. 70, p. 107-126, jan./mar. 2011. Disponível em:<<http://www.scielo.br/pdf/ensaio/v19n70/v19n70a07.pdf&gt>. Acesso em: 03 agosto, 2018.

BAKER, R. S.; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. JEDM - Journal of Educational Data Mining, 2009. v. 1, n. 1, p. 3–17, out. 2009.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidadespara o Brasil. In: . [S.l.: s.n.], 2011. v. 19, n. 3–13.

Baker, R.S.J.d., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19(02), 3.

BERNERS-LEE, Tim. Putting government data online. 2009. Disponível em:<https://www.w3.org/DesignIssues/GovData.html>. Acesso em: 29 maio 2018.

BINGHAM, E.; MANNILA, H. Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2001), August 26–29 2001, ACM: San Francisco, CA, USA. pp. 245–250, 2001.

BORUNDA, M. et al. Bayesian networks in renewable energy systems: A bibliographical survey. Renewable and Sustainable Energy Reviews, 2016. v. 62, p. 32 – 45, 2016. ISSN 1364-0321.

CALIXTO, K. E. A.; SEGUNDO, C. V. N.; GUSMÃO, R. P. Mineração de dados aplicado a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In VI Congresso Brasileiro de Informática na Educação, p. 1447 - 1456. Recife, PE, 2017.

COOPER, G. F.; HERSKOVITS, E. A Bayesian Method for the Induction of Probabilistic Networks from Data. Mach. Learn., 1992. Kluwer Academic Publishers, Hingham, MA, USA, v. 9, n. 4, p. 309–347, out. 1992. ISSN 0885-6125.

DETONI, Douglas; CECHINEL, Cristian; MATSUMURA ARAÚJO, Ricardo. Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. Revista Brasileira de Informática na Educação, v. 23, n. 3, 2015.

DINIZ, Vagner. Como conseguir dados governamentais abertos. 2010.

DUNTEMAN, G. H. Principal components analysis. Sage University paper series on quantitative applications in the social sciences, Newbury Park, CA, USA, 1999.

DUTRA, Claudio C. & LOPES, Karen M. G. (2013). Dados Abertos: uma forma inovadora de transparência. In: VI Congresso Consad de Gestão Pública. Brasília,2013.

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

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.

Ferreira, G. (2015). Investigação acerca dos fatores determinantes para a conclusão do Ensino Fundamental utilizando Mineração de Dados Educacionais no Censo Escolar da Educação Básica do INEP 2014. In: Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação.

García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. The Internet and Higher Education,14(2), 77-88.

Gottardo, E., Kaestner, C., and Noronha, R. V. (2012). Avaliação de desempenho de estudantes em cursos de educação a distância utilizando mineração de dados. In Anais do Workshop de Desafios da Computação Aplicada à Educação, pages 30–39.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3. ed. [S.l.]: Morgan Kaufmann, 2012.

HLEL, E.; JAMOUSSI, S.; HAMADOU, A. B. Bayesian network for discovering the interests of authors. In: 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA). [S.l.: s.n.], 2016. p. 1–6.

HOLMES, S.; et al. Visualization and statistical comparisons of microbial communities using R packages on phylochip data. Bioscomputing 2011: Proceedings of the Pacific Symposium. Hawaii, USA, p. 142-153, 2011.

INEP Conheça o Enem. Disponível em:<https://enem.inep.gov.br/#/antes?_k=l13hcw&gt>. Acesso em: 12 agosto 2018.

Jindal and M. D. Borah. A survey on educational data mining and research trends. International Journal of Database Management Systems, 5(3):53–73, junho, 2013.

JOLLIFFE, I.T. Principal Component Analysis (2. ed.) Springer, New York, 2002.

KO, S.; KIM, D.-W. An efficient node ordering method using the conditional frequency for the k2 algorithm. Pattern Recogn. Lett., 2014. Elsevier Science Inc., New York, NY, USA, v. 40, p. 80–87, abr. 2014. ISSN 0167-8655.

KORB, K. B.; NICHOLSON, A. E. Bayesian Artificial Intelligence, Second Edition. 2nd. ed. Boca Raton, FL, USA: CRC Press, Inc., 2010. ISBN 1439815917, 9781439815915.

Laisa, J., & Nunes, I. (2015). Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. In: Anais do XXVI Simpósio Brasileiro de Informática na Educação.

MARKOS, A. I.; VOZALIS, M. G.; MARGARITIS, K. G. An optimal scaling approach to collaborative filtering using categorical principal component analysis and neighborhood formation. In H. Papadopoulos, A. S. Andreou, & M. Bramer (Eds.), Artificial intelligence applications and innovations (AIAI 2010), October 6–7 2010. Larnaca, Cyprus: Proceedings. IFIP Advances in information and communication technology (v. 339, p. 22-29). Springer, 2010.

OCDE (2016). PISA 2015 Results: Excellence and Equity in Education. Volume I. Paris: OECD Publishing.

OECD – ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. OECD Strategy on Development. Paris: OECD, 2016.

OPEN KNOWLEDGE BRASIL. Índice de Dados Abertos da Open Knowledge indica pouco progresso por parte dos governos em abrir dados chave. 2014. Disponível em: . Acesso em: 27 maio 2018.

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems. San Francisco, Calif.: Morgan Kaufmann.

QUEIROGA, Emanuel Marques. Geração de modelos de predição para estudantes em risco de evasão em cursos técnicos a distância utilizando técnicas de mineração de dados. 2017. Dissertação de Mestrado. Universidade Federal de Pelotas.

RIBEIRO, Claudio Jose Silva; ALMEIDA, Reinaldo Figueiredo de. Dados Abertos Governamentais (Open Government Data): instrumento para exercício de cidadania pela sociedade. XII Enancib-Políticas de Informação para a Sociedade-Anais. Brasília: Thesaurus, p. 2568-2580, 2011.

RUSSEL, J. S.; NORVING, P. Artificial Intelligence: A Modern Approach. 3. ed. [S.l.]: Pearson, 2010.

SANTANA Ádamo Lima de. Estratégias para a melhoria da modelagem e interpretabilidade de redes bayesianas. Doutorado em Engenharia Eletrica, 2008.

SARDINHA, R. de L.; PAES, A.; ZAVERUCHA, G. Aprendizado Local da Estrutura de Redes Bayesianas a partir de Dados Incompletos - Bayes Ball Structure Learning (BBSL). 2009.

SCAICO, Pasqueline Dantas; QUEIROZ, Ruy José G. B. de; SCAICO, Alexandre. O conceito big data na educação. 3º Congresso Brasileiro de Informática na Educação (CBIE 2014) – 20º Workshop de Informática na Escola (WIE 2014). Disponível em: < http://www.br-ie.org/pub/index.php/wie/article/view/3115>. Acesso em: 30 mar. 2018.

SHLENS, J. A Tutorial on Principal Component Analysis. La Jolla, California, USA: Systems Neurobiology Laboratory, Salk Institute for Biological Studies, 2005.

SILVA, L. A.; MORINO, A. H.; SATO, T. M. C. (2014) Prática de Mineração de Dados no Exame Nacional do Ensino Médio. Anais dos Workshops do Congresso Brasileiro de Informática na Educação, v. 3, n. 1, p. 651–660.

SIMON, Augusto; CAZELLA, Sílvio. Mineração de Dados Educacionais nos Resultados do ENEM de 2015. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2017. p. 754.

SONG, M.; YANG, H.; SIADAT, S.; PECHENIZKIY, M. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. Expert Systems with Applications, v. 40, pp. 3722–3737, 2013.

SRIVASTAVA, Jaya; SRIVASTAVA, Abhay Kumar. Data mining in education sector: a review. International Journal of Advanced Networking Applications, Special Conference Issue, National Conference on Current Research Trends in Cloud Computing & Big Data, p. 184–190, 2013. Disponível em;<https://pdfs.semanticscholar.org/d91e/ee207844e8d3a2dfc7308fa243bbc1a8a3ae.pdf&gt> . Acesso em: 31 Setembro. 2018.

Stearns, B., Rangel, F., Rangel, F., Firmino, F., and Oliveira, J. (2017). Scholar performance prediction using boosted regression trees techniques. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).

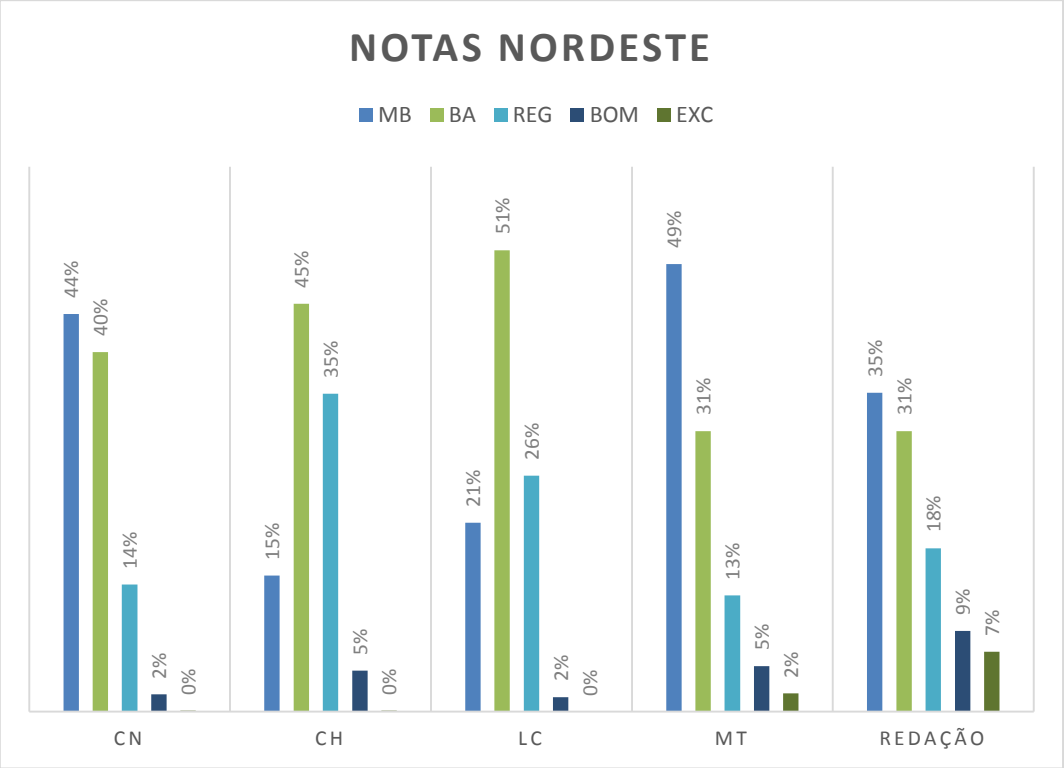
XU, X.; WANG, X. An adaptive network intrusion detection method based on PCA and support vector machines. In X. LI, S.; WANG, Z. Y. Dong (Eds.), Advanced data mining and applications, first international conference (ADMA 2005), July 22–24, 2005. Wuhan, China: Proceedings. Lecture notes in computer science (v. 3584, p. 696–703). Springer, 2005.

YANG, H. HARRINGTON, C. A.; VARTANIAN, K.; COLDREN, C. D.; HALL, R.; CHURCHILL, G. A. Randomization in laboratory procedure is key to obtaining reproducible microarray results. PLoS One, 3:e3724, 2008.

KRAWCZYK, N. O Ensino médio no Brasil. São Paulo: Ação Educativa, 2009. (Coleção Em Questão, 6)

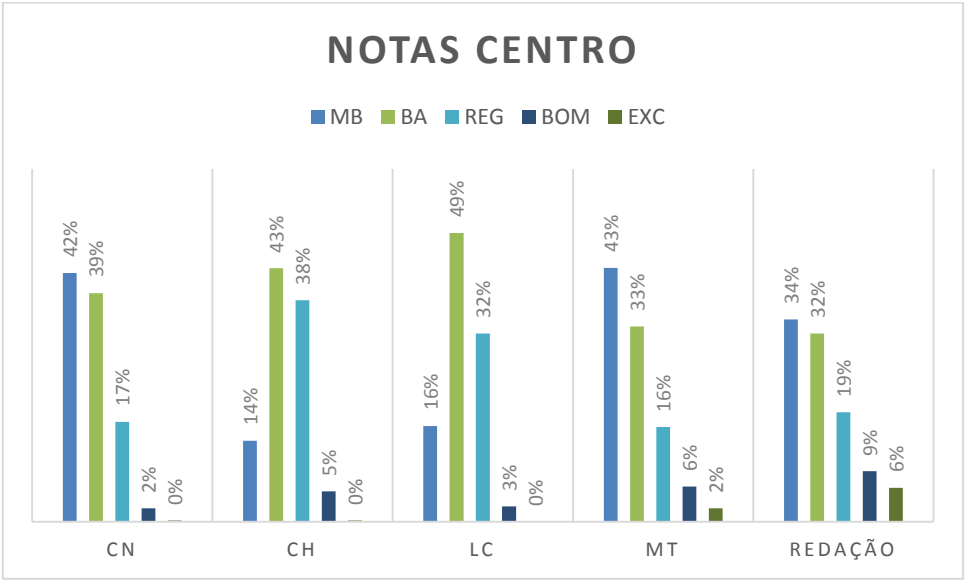
APÊNDICE A – Comportamento das notas na região Nordeste e Centro Oeste

Gráfico demonstrando o comportamento das notas da região Nordeste em cada área do conhecimento.



Fonte: Autor, 2019.

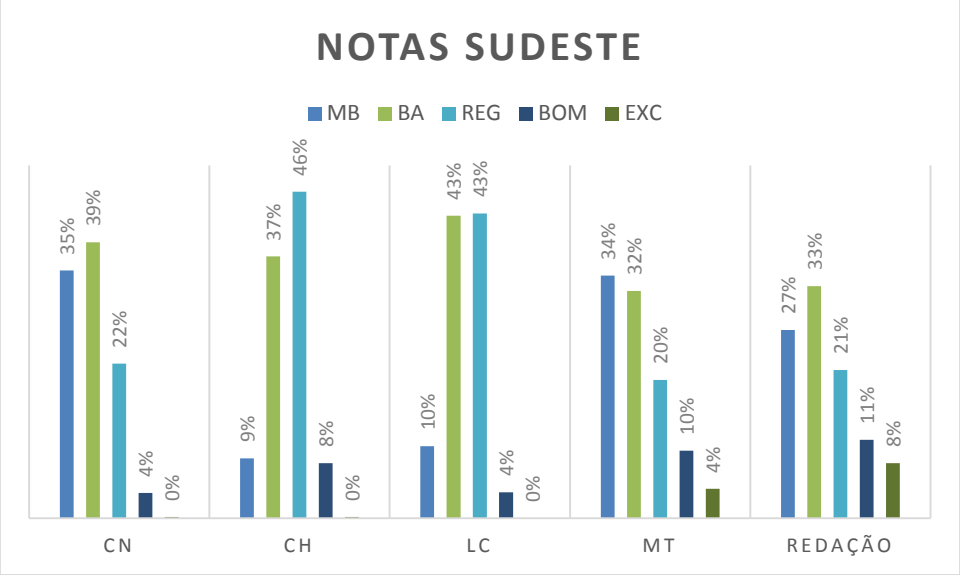
Gráfico demonstrando a concentração das notas da região Centro em cada área do conhecimento.



Fonte: Autor, 2019.

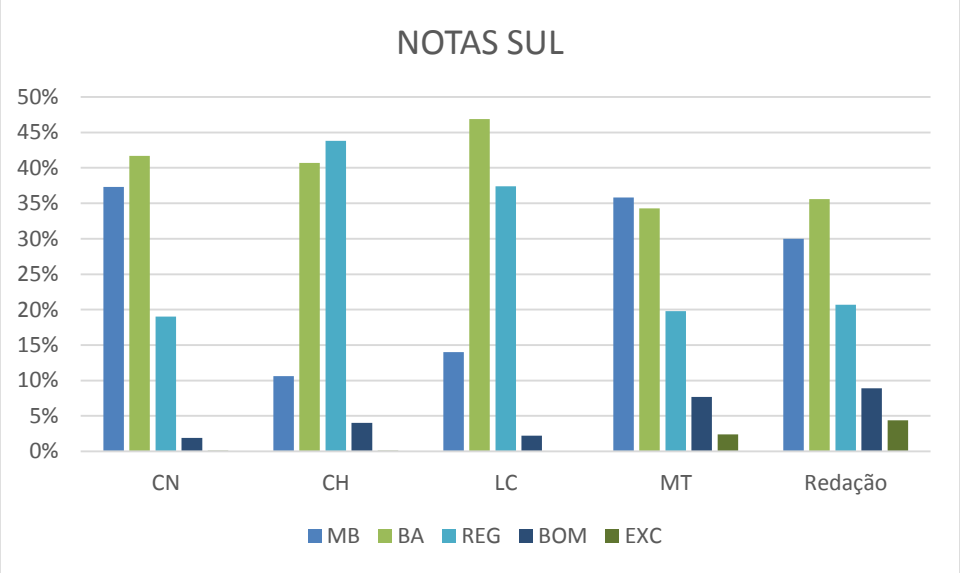
APÊNDICE B - Comportamento das notas na região Sudeste e Sul

Gráfico com as concentrações das notas da região Sudeste em cada área do conhecimento.



Fonte: Autor, 2019.

Gráfico com as concentrações das notas da região Sul em cada área do conhecimento.



Fonte: Autor, 2019