Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática

A Influência das Características das Escolas do Nordeste Brasileiro na Obtenção das Notas do ENEM

Philippe Cesar dos Santos Oliveira

Recife
2017

Philippe Cesar dos Santos Oliveira

A Influência das Características das Escolas do Nordeste Brasileiro na Obtenção das Notas do ENEM

Orientador: Rafael Ferreira Leite de Mello

Monografia apresentada ao curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco como requisito para obtenção do título de Bacharel em Ciência da Computação.

Recife
2017

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido, no dia 22 de agosto de 2017 às 14 horas, no Auditório do CEAGRI-02 - Sala 07, por Philippe Cesar dos Santos Oliveira como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **A Influência das Características das Escolas do Nordeste Brasileiro na Obtenção das Notas do ENEM**, orientado por Rafael Ferreira Leite de Mello e aprovado pela seguinte banca examinadora:

Rafael Ferreira Leite de Mello
DEINFO/UFRPE

Roberta Macêdo Marques Gouveia
DEINFO/UFRPE

Péricles Barbosa Cunha de Miranda
DEINFO/UFRPE

Agradecimentos

A Jesus, sem o qual eu nem sequer existiria. Toda força, paciência, perseverança, inteligência, raciocínio e sorte que tive vieram Dele, porque Dele, por Ele e para Ele são todas as coisas.

A minha mãe, Isis, minha melhor amiga e maior apoiadora em todo esse trajeto não só durante esse tempo em BCC mas na vida inteira, me dando tanto amor e suporte que nem mereço.

Ao meu orientador, Rafael Ferreira, por ter sido o primeiro a me empurrar para o mundo acadêmico, incentivando a pesquisar, escrever artigos, fazer experimentos. A disciplina de Reconhecimento de Padrões foi um marco, um ponto de virada.

Ao professor George Cabral, que talvez nem sabe, mas, ao ministrar a disciplina de Inteligência Artificial, deu um real sentido a um estudante que mal sabia o que estava fazendo no curso de Ciência da Computação; sentido esse reforçado na disciplina de Redes Neurais (mesmo este trabalho não tendo nenhuma relação com RNs).

Ao corpo docente do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, pela imensa contribuição (e paciência) na minha formação acadêmica ao longo dos anos.

A meus familiares: irmãos, primos, tios. Todos tiveram uma parcela de contribuição. A família é grande e fica difícil lembrar todos os nomes, mas sou muito grato a todos.

Aos amigos e colegas de curso. O companheirismo, as risadas, os puxões de orelha, os conselhos, as brincadeiras, tudo foi útil.

A toda a igreja de Cristo em Recife, independente da denominação (ou da falta dela), por todas as orações e por cada momento de comunhão (que ainda quero ter muito mais).

Resumo

Nos últimos anos, o ENEM vem sendo uma das principais métricas de avaliação do Ensino Médio no Brasil. No entanto, desde sua utilização como seleção unificada a partir de 2009, os dados disponibilizados se expandiram bastante. Além disso, esses dados não estão totalmente estruturados e são divulgados em formatos dos mais variados, e outros dados referentes à educação nacional, relacionados com informações contidas nos microdados do ENEM, são publicados em outras bases, como os Censos Escolares, sem que haja ligação entre esses dados (como *links* ou referências), dificultando a produção de conhecimento. Diante deste cenário, a partir de dados obtidos do portal Educação Inteligente, que coleta dados abertos educacionais brasileiros de diferentes fontes e os organiza e republica, esse trabalho propõe a utilização de extração de conhecimento em base de dados. O intuito é descobrir conjuntos de características que se repitam e indiquem padrões dentro da educação brasileira para análise das características de escolas participantes do ENEM a partir dos dados abertos disponibilizados pelo INEP. Após selecionar e transformar os dados obtidos, a seleção automática de atributos definiu 52 atributos como sendo as principais características das escolas (incluindo tanto fatores de infraestrutura quanto de opções oferecidas aos estudantes) que influenciam na nota do ENEM obtida por cada escola. A utilização de árvores de decisão e regras de associação para classificar os dados extraiu regras explícitas para auxílio de gestores de escolas e administradores públicos, e os valores de Acurácia e *F-Measure* dos melhores algoritmos ficaram entre 79% e 82%. É ainda apresentada uma revisão da literatura sobre o tema, a fim de prospectar mais trabalhos relacionados, dada que a área de pesquisa ainda é recente no Brasil.

Palavras-chave: Árvores de decisão, Regras de associação, Weka, características de escolas, dados abertos educacionais, KDD, ENEM.

Abstract

In recent years, the ENEM has been one of the main evaluation metrics for High School in Brazil. However, after its use as a unified selection since 2009, the available data has expanded greatly. In addition, these data are not fully structured and are publicized in a variety of formats, and other data related to national education, related to information contained in the ENEM microdata, are published in other databases, such as the School Census, without any link between these data (such as URLs or references), making it difficult to produce knowledge. Given this scenario, based on data obtained from the Intelligent Education portal, which collects brazilian open educational data from different sources and organizes and republishes it, this paper proposes the use of knowledge extraction in a database. The intention is to find sets of characteristics that repeat themselves and indicate patterns within the brazilian education to analyze the characteristics of ENEM participating schools, based on the open data provided by INEP. After selecting and transforming the obtained data, the automatic attribute selection defined 52 attributes as being the main features of schools (including both infrastructure factors and options offered to students) that influence the ENEM grade obtained by each school. The usage of decision trees and association rules to classify data extracted explicit rules to help school administrators and public administrators, and Accuracy and F-Measure values of the best algorithms ranged from 79% to 82%. It is also presented a literature review on the subject, in order to prospect more related works, given the research area is still recent in Brazil.

Keywords: Decision trees, Association rules, Weka, school characteristics, open educational data, KDD, ENEM.

Lista de figuras

Lista de tabelas

Sumário

Capítulo 1

1. INTRODUÇÃO

Com a disseminação da Internet, há uma grande facilidade em disponibilizar conteúdo em diversos meios, como redes sociais, e-mails, fóruns de discussão, entre outros, o que ocasiona um crescimento exponencial dos dados gerados, de tal maneira que dados não estruturados compõem cerca de 90% do universo digital (BAUER e KALTENBÖCK, 2012). No entanto, mais dados não necessariamente indicam mais informação ou mais conhecimento (BAUER e KALTENBÖCK, 2012). Muitos desses dados não são acessíveis ao público, e há casos em que são incompreensíveis mesmo por quem está autorizado a acessar e manipular seu conteúdo. Isto resulta em ineficácia e lentidão ao tentar obter conhecimento útil para a sociedade (ISOTANI e BITTENCOURT, 2015).

Para lidar com esse problema, estão surgindo iniciativas para liberar dados em formato processável por máquina que, ao serem processados, devolvam conhecimento aplicável e valorizem essas informações disponibilizadas na Internet. Dentre essas iniciativas, existem os Dados Abertos (BAUER e KALTENBÖCK, 2012; COLPAERT, 2013). Dados digitais são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais (ISOTANI e BITTENCOURT, 2015). Dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los, estando sujeito, no máximo, à exigência de creditar a sua autoria e compartilhar pela mesma licença (BAUER e KALTENBÖCK, 2012).

Juntamente a tais formatos abertos, existem processos para extração de conhecimento desses dados, e nesse sentido se destaca a Mineração de Dados (MD), que oferece suporte a esse tipo de dados, de modo a identificar padrões e correlações e apontar relevâncias. A MD é uma parte do processo de descoberta de conhecimento em banco de dados, em inglês denominado *Knowledge Discovery in Databases*, ou KDD (FRAWLEY, PIATETSKY-SHAPIRO e MATHEUS, 1992), que compreende uma série de etapas para extrair informações relevantes dentro de grandes quantidades de dados (RATH, JONES, *et al.*, 2016).

Nos últimos anos, as ferramentas e técnicas para análise de dados e descoberta de conhecimento têm sido largamente adotadas por outros setores da sociedade além de cientistas e pesquisadores. Bancos, lojas, fabricantes, agricultores e outros “têm descoberto o potencial

de descobrir relações entre padrões não visíveis em vastos bancos de dados e até em redes sociais" (GOMES, 2015). O contexto educacional não tem ficado de fora, e muitas instituições de ensino têm se utilizado da Mineração de Dados Educacionais (MDE) a partir de dados abertos de contexto educacional para tomar decisões estratégicas e aprimorar sua estrutura e o oferecimento de produtos:

> A publicação de dados tem gerado benefícios nas diversas áreas em que ocorre, como a transparência em órgãos governamentais que é utilizada como uma importante ferramenta no combate a corrupção e a ampliação da participação da sociedade no processo de desenvolvimento de novas soluções. Além do setor governamental, a educação também se torna uma beneficiária, pois a grande quantidade de dados disponíveis pode auxiliar na tomada de decisão de professores e gestores escolares. (ALCANTARA, 2015)

Além de instituições de ensino, outras pessoas podem utilizar a MDE em seus contextos. Podemos conjecturar que pais e responsáveis podem buscar aplicações que utilizem esses dados educacionais para compreender a estrutura, as condições de pagamento de mensalidades, os recursos materiais e humanos e as avaliações do último ENEM para levá-los a matricular seu filho em uma determinada escola. Um educador (ou mesmo um grupo deles) que seja docente de um determinado curso com eventual alta evasão de alunos pode buscar maneiras para cessar tal fato através de alternativas utilizadas por outros educadores em outras instituições.

A própria instituição de ensino pode repensar opções de encaminhamento para o mercado de trabalho de acordo com outras instituições; seu gestor pode também mensurar os tipos e características de cursos que mais atraem novos alunos, ou ainda reavaliar e reorganizar seus recursos estruturais, lógicos, financeiros e humanos, beneficiando a qualidade do ensino oferecido. E de fato, de acordo com diversos estudos publicados, a infraestrutura das escolas impacta a qualidade do ensino das escolas. Um estudo de Castro e Fletcher (1986 apud SANTOS, CLARO, *et al.*, 2014) aponta a eficiência trazida quando o governo investe na educação e a implicação no aprendizado dos alunos a partir da qualidade de ensino resultante da infraestrutura das escolas. Lee *et al*. (2014) também afirmam que há impacto positivo na aprendizagem a partir de uma melhor infraestrutura.

No entanto, Soares Neto *et al*. (2013), utilizando dados do Censo Escolar de 2011, apontaram que apenas 0,6% das escolas brasileiras dispõem de infraestrutura considerada ideal para ensino. O mesmo estudo revela que 44% das escolas básicas apresentaram estrutura tida como elementar, apenas com água, sanitário, esgoto, energia e cozinha, e que há grande desigualdade regional na infraestrutura das escolas brasileiras. Uma publicação de Santos *et al*. (2014), utilizando KDD e regras de associação em base de dados do Censo Escolar 2011,

evidencia que escolas da região Nordeste trazem infraestrutura deficitária, onde grande parte das escolas não tem biblioteca ou sala de leitura.

Além dessas questões educacionais, a própria divulgação e obtenção dos dados representa uma adversidade. Pouco conhecimento sistemático é extraído de muitas das fontes de dados educacionais do Brasil, tanto pelo volume grande de dados (acima de 10 Gb/ano) quanto pela diferença na granulação: os dados são referentes a escolas, alunos e professores de escolas, mas são publicados separadamente e praticamente sem links/referências entre um grão e outro (quando existem) (ADEODATO, SANTOS FILHO e RODRIGUES, 2014). A formatação dos dados também não é padronizada, e o acesso às informações pode ser em páginas web, planilhas *csv* ou *xls*, PDFs ou ainda arquivos compactos, dentre outros.

Dessa maneira, é verificável a necessidade de processamento e reorganização desses dados educacionais, através de mineração de dados (ADEODATO, SANTOS FILHO e RODRIGUES, 2014). Assim, compilando seu conteúdo de maneira legível e aproveitável por pessoas, torna-se viável seu aproveitamento por gestores de escolas e administradores públicos, diretamente envolvidos no âmbito educacional. Além disso, há a possibilidade de produzir conhecimento útil e abertamente divulgável para a sociedade, a fim de lidar com diversas questões, sejam educacionais, sociais, políticas, econômicas ou outras (ISOTANI e BITTENCOURT, 2015).

1.1. JUSTIFICATIVA

Como já foi dito, a partir de dados abertos pode ser extraído conhecimento importante para gestores educacionais. Por exemplo, a relação aprendizagem-infraestrutura é verificável ao analisar a presença ou ausência de laboratórios de informática. É senso comum afirmar que mais computadores (ou dispositivos tecnológicos em geral) representem uma melhoria na aprendizagem, e nesse aspecto Löbler *et al*. (2012) ponderaram a relação entre a presença de laboratórios de informática com o desempenho das escolas. Na pesquisa, os autores escolheram duas escolas, identificando uma delas como sendo de alto desempenho e outra de baixo desempenho. Para as duas escolas, houve questionários que deveriam identificar relação entre o desempenho escolar dos alunos e o uso do computador. Ao concluir a pesquisa, não foi encontrada pelos autores uma relação forte entre um e outro, e na verdade observou-se que apenas usar os computadores não se mostrou suficiente para aumentar o desempenho.

Isto não deve ser motivo para retirar computadores de escolas em definitivo. Aguiar e Nascimento (2014) basearam-se em outros trabalhos de Informática na Educação de diferentes autores, que apontaram que o uso de tecnologia da informação favorece o ensino de qualidade, e utilizaram dados do Censo Escolar do Ensino Médio e do ENEM, ambos de 2009 a 2011. Com isso, produziram um estudo de caso onde, dentre outras conclusões, percebeu-se que o número de computadores e o desempenho da escola estão ligados, isto é, mais computadores resultam em melhor desempenho, e vice-versa. Entretanto, o estudo também aponta que, por deficiência das fontes de dados utilizadas, mesmo sabendo que uma escola tenha laboratório de informática e internet, não há como afirmar "se os professores realmente utilizam estes recursos com os alunos e se esse uso é ou não de qualidade, incluindo planejamento, softwares educativos, entre outras informações".

Essa última informação corrobora com o trabalho de Adeodato *et al*. (2014). A partir de dados educacionais abertos do ENEM 2011 e do Censo Escolar do mesmo ano, e utilizando mineração de dados, regressão logística e árvore de decisão, concluiu-se que, dentre muitos outros fatores, possuir infraestrutura de laboratório de computadores pode ser tanto bom quanto ruim, porque sua eficácia depende de alocação, controle e maneira de uso, podendo trazer distrações para a aprendizagem ao serem usados, por exemplo, para interação em redes sociais. Por outro lado, o mesmo estudo concluiu que a presença de laboratório de ciências contribui para um bom êxito da escola.

Por fim, em experimento que averiguou motivações determinantes para a conclusão ou não do Ensino Fundamental na cidade de Porto Alegre, Rio Grande do Sul, Ferreira (2015), utilizando dados do Censo Escolar da Educação Básica de 2014, concluiu que "os recursos de internet banda larga, laboratório de ciências, auditório na escola e ensino privado apresentaram forte ligação com maiores chances de o aluno apresentar o Ensino Fundamental concluído".

Todos os estudos citados também concluíram outros resultados, como a forte influência do fator socioeconômico. Diante da importância de tais fatores, percebe-se a importância do aprimoramento na abordagem do tema e no beneficiamento da infraestrutura nas escolas brasileiras. Com base nisso, este trabalho traz a hipótese de que, a partir da obtenção de dados abertos educacionais referentes a características de escolas, e processando esses dados utilizando KDD e técnicas de reconhecimento de padrões, pode-se avaliar as escolas do Nordeste brasileiro que ofereçam Ensino Médio.

Ao descrever os atributos estruturais e não-estruturais de uma escola, e ao classificar escolas reais em grupos a partir de suas notas médias obtidas no ENEM, obtém-se regras que denotam um conjunto de atributos que efetivamente levam àquele resultado, para uma escola

ou para uma coleção delas. Serão utilizados diferentes algoritmos para executar essa classificação, e tendo sido escolhidos os com melhor desempenho (dentro de uma série de critérios), os resultados destes serão utilizados para tentar responder à seguinte pergunta de pesquisa: “Quais características estruturais e não-estruturais de uma escola influenciam no resultado do ENEM obtido por ela?”.

1.2. OBJETIVOS

1.2.1. Objetivo Principal

Analisar a influência dos atributos de escolas do Nordeste brasileiro nas notas obtidas no ENEM por meio de aplicação do processo de descoberta de conhecimento em dados abertos educacionais.

1.2.2. Objetivos Específicos

Dentre os objetivos específicos deste trabalho, podem ser destacados os seguintes:

- Extrair informações de dados abertos educacionais, criando base de dados de escolas possível de ser reutilizada em outros experimentos;
- Descrever quais os atributos estruturais e não-estruturais de uma escola utilizando algoritmos de árvore de decisão ou algoritmos de classificação que gerem regras de associação, a fim de buscar melhor compreensão do domínio do problema;
- Contribuir com os estudos de análise de avaliações de ensino através de mineração de dados, apresentando uma solução com a utilização de técnicas de aprendizagem de máquina para extrair características de dados.

1.3.ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma:

- O capítulo 2 apresenta um conjunto de definições importantes para este trabalho, bem como uma contextualização sobre o tipo de dados utilizado e ainda as ferramentas utilizadas e os algoritmos escolhidos;
- O capítulo 3 apresenta uma revisão de literatura, efetuada de maneira sistemática, além dos trabalhos relacionados;
- O capítulo 4 apresenta técnicas e métodos utilizados no desenvolvimento deste trabalho;

- O capítulo 5 apresenta os resultados obtidos e as discussões;
- O capítulo 6 apresenta as considerações finais.

Capítulo 2

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção trata dos conceitos e temas necessários para melhor entendimento dos assuntos abordados neste trabalho

2.1.DADOS ABERTOS

Apesar do termo Dados Abertos ter sido concebido e ganhado força nos anos recentes, ele é um fenômeno inspirado em outros movimentos anteriores, como o *Open Source* (WEBER, 2004), que cresceu a partir de um interesse dos indivíduos, até chegar a produtos comercialmente viáveis. Os Dados Abertos são importantes bases de dados para aplicação de técnicas de mineração, consistindo em uma iniciativa amplamente fomentada pela *Open Knowledge Foundation* (OPK)[1], e seu uso denota diversos benefícios. Podemos conjecturar inicialmente esses benefícios da seguinte maneira:

> [Imaginando uma determinada cidade], caso os dados sobre os transportes públicos [desta cidade] estejam disponíveis livremente na Web em formato aberto, um cidadão poderia ter acesso às informações contidas nestes dados e utilizá-las a seu favor. Por exemplo, poderia planejar uma rota simples da sua casa ao trabalho, utilizando diferentes meios de transporte; poderia também comparar o custo/benefício de diferentes rotas e tipos de transporte. Da mesma forma, um funcionário de um município também usufruiria de benefícios, pois poderia facilmente acessar dados de outros municípios e do estado para realizar suas atividades. Por exemplo, poderia adequar os horários dos transportes locais; alocar mais veículos em horários com maior demanda; e também comparar os dados locais com os de municípios similares para analisar a eficiência do serviço. (ISOTANI e BITTENCOURT, 2015)

É importante ressaltar os oito princípios que pautam os dados abertos em geral (ISOTANI e BITTENCOURT, 2015):

1. *Completos*. Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão

[1] https://okfn.org

sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.

2. *Primários*. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada.
3. *Atuais*. Os dados são disponibilizados o quão rapidamente seja necessário para preservar o seu valor.
4. *Acessíveis*. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.
5. *Processáveis por máquina.* Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.
6. *Acesso não discriminatório*. Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.
7. *Formatos não proprietários*. Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.
8. *Livres de licenças*. Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

2.1.1. Legislação e Dados Abertos Governamentais

Em muitos países, o Estado tem um papel fundamental na distribuição de dados abertos, a partir de legislações que prescrevem acesso às informações governamentais. Nessas leis, é descrito que os cidadãos devem ter acesso aos dados na forma mais ampla possível, o que abrange o uso de dados abertos. No Brasil, a Lei nº 12.527, de 18 de novembro de 2011, conhecida como Lei de Acesso à Informação (LAI), regulamenta o direito constitucional de acesso às informações públicas e cria mecanismos que possibilitam a todos o recebimento de informações públicas dos órgãos e entidades (BRASIL, 2011). A LAI deve ser cumprida pelos três poderes da União, em níveis Federal, Estadual e Municipal.

Sendo o Governo um importante divulgador de dados abertos, e com a crescente utilização da Internet como meio de participação popular, é coerente que o cidadão possa ter acesso a dados públicos, controlar as ações dos agentes públicos, cobrar providências e participar de assuntos de interesse geral, pois “abre-se caminho para uma Administração Pública pautada pela eficiência, eficácia, efetividade e com postura ética” (ATRICON, 2015).

De maneira análoga, o Tribunal de Contas da União (TCU) criou uma cartilha denominada “5 motivos para a abertura de dados na Administração Pública” (TRIBUNAL DE CONTAS DA UNIÃO, 2015). As justificativas utilizadas pelo TCU para disponibilizar dados públicos de maneira aberta são:

1. *Transparência na gestão pública*. Existindo transparência, a sociedade pode avaliar ações e decisões governamentais, dando à população o papel de agente transformador, através de fiscalização do desempenho do governo.
2. *Contribuição da sociedade com serviços inovadores ao cidadão.* Organizações, cidadãos, acadêmicos e mesmo instituições públicas podem empregar bases de dados públicos a fim de criar e distribuir novos conhecimentos e serviços em coparticipação entre o ente privado e o governo na oferta de serviços públicos à sociedade.
3. *Aprimoramento na qualidade dos dados governamentais*. A abertura dos dados governamentais permite que os próprios cidadãos possam apontar inconsistências e erros e sugerir correções nos próprios dados, contribuindo ativamente para a melhoria de coleta e publicação dos dados e diminuindo os esforços da Administração Pública.
4. *Viabilização de novos negócios.* A iniciativa privada pode utilizar dados abertos governamentais para elaborar produtos e serviços que até então eram inexistentes e comercializá-los à população, o que possibilita criar novos postos de trabalho e por consequência aumentar a receita pública mediante o recolhimento de tributos.
5. *Obrigatoriedade por lei.* Mesmo antes da existência da LAI, a Lei Complementar 101/2000, conhecida como a Lei de Responsabilidade Fiscal, já tratava da transparência, do controle e da fiscalização da gestão fiscal. A abertura de dados governamentais não é apenas uma opção para viabilizar a transparência pública, mas também é um dever a ser cumprido pelo administrador público.

2.1.2. Dados Abertos Educacionais, ENEM e INEP

A aprovação da Lei de Acesso à Informação levou à criação do Portal Brasileiro de Dados Abertos[2], que tem o objetivo de ser o ponto único referencial para a busca e o acesso à dados públicos brasileiros de todo e qualquer assunto ou categoria. Por esse motivo, suas atribuições vão desde divulgar iniciativas – como aplicativos que utilizem dados abertos – a agregar, republicar e buscar atualização das bases de dados públicos em geral; dentre eles,

[2] http://dados.gov.br/

existem os dados abertos relativos à educação, denominados Dados Abertos Educacionais (DAE).

Existe na literatura um termo semelhante a Dados Abertos Educacionais: são os Recursos Abertos Educacionais (RAE). É importante diferenciar um do outro: ambos são dados relativos à educação, porém, os RAE são materiais que podem ser usados em ensino e pesquisa, tendo como principal foco ajudar professores a criar aulas mais iterativas e ricas (DUTRA e TAROUCO, 2007). Eles estão focados no conteúdo propriamente dito, e não em estatísticas de professores, escolas e alunos, como é o caso dos DAE.

Os DAE, foco principal deste trabalho, se referem especificamente aos dados abertos vindos das instituições de ensino, como escolas, universidades, bibliotecas, entre outras (GUY; ALCANTARA, 2015), e também de avaliações e pesquisas do próprio Governo. Estes dados podem ser de vários tipos, como:

- Dados Internos: dados de staff, infraestrutura, recursos disponíveis, orçamento;
- Dados Pedagógicos: dados de cursos, grades curriculares, objetivos de aprendizagem;
- Dados de Usuários: estatísticas de aprendizagem, avaliações, dados de performance.

Uma importante fonte de dados abertos educacionais no Brasil é o ENEM – Exame Nacional do Ensino Médio (BRASIL, 1998). Criado em 1998, o Exame foi a primeira iniciativa de avaliação do sistema de ensino a ser implantada em nível nacional no país, sendo uma ferramenta extra vestibular de auxílio ao Ministério da Educação para elaborar políticas de melhoria do ensino escolar, através do cruzamento de pesquisas e dados dos resultados no ENEM com os Parâmetros Curriculares Nacionais do Ensino Médio (BRASIL, 2000), literalmente tendo como objetivo “avaliar o desempenho da educação básica, buscando contribuir para a melhoria da qualidade desse nível de escolaridade” (GOMES, 2015).

Em 2009, o Ministério introduziu um novo formato para o Exame, que continuaria a fornecer dados para a avaliação do ensino médio no Brasil, e unificaria as provas de vestibular e o processo de seleção e ingresso nas universidades federais brasileiras a partir do Sistema de Seleção Unificada (SiSU), servindo como fase única do vestibular. Poderia servir também como parte da composição da nota de ingresso, combinando-se ao processo seletivo próprio da universidade. Também passou a servir de certificação para alunos que cursaram a Educação de Jovens e Adultos (EJA), bem como é ou foi utilizado em outros programas oferecidos pelo Governo Federal, tais como o Ciência sem Fronteiras (CsF), Programa Universidade para Todos (PROUNI), Fundo de Financiamento Estudantil (FIES), dentre outros. Desde sua

criação, o ENEM ocorreu em abrangência nacional e de maneira anual e ininterrupta, o que demonstra a importância dada pelo Governo ao Exame.

Nem todos os DAE são necessariamente de fontes governamentais, entretanto, no Brasil, os dados educacionais governamentais, ainda que publicados no Portal Brasileiro de Dados Abertos, são organizados e fornecidos pelo INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira[3], sendo o maior divulgador de DAE no país. Além dos dados do ENEM, o INEP divulga diferentes resultados de avaliações nacionais de forma aberta, como os dados do Censo Escolar, do Índice de Desenvolvimento da Educação Básica (IDEB), da Prova Brasil, e outros.

2.2.MINERAÇÃO DE DADOS

A Mineração de Dados (MD) dados é um processo de pesquisa interdisciplinar cujo foco e definições internas dependem bastante do campo de atuação dos autores, podendo ser relacionado a domínios de pesquisa como Estatística, Aprendizagem de Máquina, Banco de Dados, entre outros, sejam tais campos utilizados isoladamente ou (mais costumeiramente) em conjunto. Fayadd *et al*. (1996) apontam que "Mineração de dados é um passo no processo de descoberta de conhecimento que consiste na realização da análise dos dados e aplicação de algoritmos de descoberta que, sob certas condições, produzem um conjunto de padrões de certos dados". Já Hand *et al*. (2001) define da seguinte maneira: "Mineração de dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis aos donos dos dados".

O processo de descoberta de conhecimento em banco de dados (do inglês *Knowledge Discovery in Databases* - KDD) compreende uma série de etapas para extrair informações relevantes dentro de grandes quantidades de dados (RATH, JONES, *et al.*, 2016). Ou seja, consiste em um processo não trivial de identificação de padrões válidos, desconhecidos (a priori), potencialmente uteis e interpretáveis (FAYYAD, PIATESKY-SHAPIRO e SMYTH, 1996). Inicialmente, o processo dependerá da escolha dos dados pelos pesquisadores, pressupondo-se uma ampla compreensão do domínio da aplicação, bem como do tipo de decisão que tal conhecimento pode trazer como contribuição. O ideal é que a base de dados contenha uma grande quantidade de dados relacionados ao conhecimento que se quer descobrir.

A partir disso, a primeira etapa do KDD consiste na Seleção de um conjunto de dados,

[3] http://inep.gov.br

ou subconjunto do banco original, que vai ser trabalhado. A segunda etapa é o Pré-processamento, onde tais dados são tratados de maneira a remover ruídos, formatar os dados, tratar campos ausentes, selecionar atributos relevantes, entre outros. A próxima etapa é a Transformação, responsável por redimensionar, normalizar e modificar os tipos dos dados, entre outros, a fim de padronizar a etapa do processamento. Por fim, utilizam-se técnicas de MD para extração de padrões e relações entre os dados já transformados nas etapas anteriores. A área de MD possui diversas técnicas automáticas para classificação, agrupamento e associação de dados que podem ser usadas para gerar conhecimento a partir de grandes quantidades de dados (WITTEN, FRANK, *et al.*, 2016).

A fase final do KDD consiste na interpretação e avaliação dos resultados da MD, seja incorporando o conhecimento ou documentando e reportando aos responsáveis. Estes resultados precisam ser descritos em alguma linguagem que possa ser facilmente compreendida pelos usuários finais, a fim de que eles possam realizar uma análise mais profunda. Como, dentre as várias técnicas de MD, os algoritmos de árvore de decisão e regras de associação são os mais utilizados quando a interpretação humana dos resultados é necessária, por gerarem regras explícitas sobre a associação entre os atributos que levaram a uma determinada classificação dos dados (WITTEN, FRANK, *et al.*, 2016), os resultados de tais algoritmos facilitam a interpretação destes padrões, e por isso foram utilizados neste trabalho.

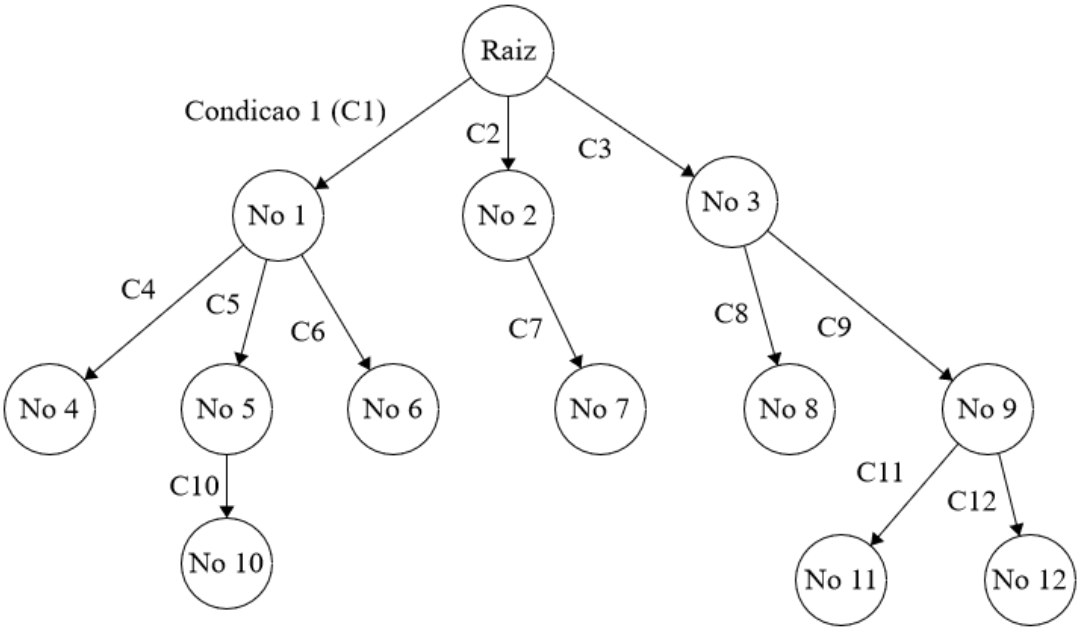
2.3. ÁRVORES DE DECISÃO E REGRAS DE ASSOCIAÇÃO

Dentro da Inteligência Artificial, existe uma área denominada Aprendizagem de Máquina. Ela é responsável por definir sistemas que possam obter conhecimento utilizando e interpretando um conjunto de dados. A partir daí, é possível tomar decisões apoiadas nos exemplos solucionados com sucesso (WEISS e KULIKOWSKI, 1991). Árvore de Decisão é uma dentre as técnicas de aprendizado de máquina, e se utiliza de uma modelagem em formato de grafo ou árvore para tomar decisões e encontrar suas possíveis consequências, como efeitos, custos ou utilidades.

A Figura 1 mostra a estrutura de uma árvore de decisão. Em cada nó da árvore, incluindo a raiz, estão localizados os atributos da amostra. Cada aresta que parte de um nó é uma condição, indicando qual o valor do atributo que foi designado àquela instância, de acordo com a predição do algoritmo, já treinado. O algoritmo percorre a árvore até chegar na folha, que será uma resposta com a classe. A interpretação da árvore por um humano aconteceria ao considerar um trajeto único da raiz até as folhas. Por exemplo, de acordo com a figura x, imaginemos o

caminho Raiz (C3) – Nó 3 (C9) – Nó 9 (C11) – Nó 11. Um exemplo, considerando nossa base de dados, poderia ser alimentacao (False) – computadoresAlunos (40-59) – laboratorioCiencias (True) – 600-699, o que indicaria que uma escola ou um grupo de escolas que ofereçam alimentação, disponibilizem entre 40 e 59 computadores especificamente para os alunos e que possuem laboratório de ciências foram classificadas como tendo obtido uma nota do ENEM entre 600 e 699. Outros caminhos sugerem outras interpretações, que contribuem para o modelo do sistema como um todo.

Figura 1 - Estrutura de uma árvore de decisão



Fonte: o autor

Algoritmos de regras de associação seguem exatamente o mesmo princípio acima, porém, não com um grafo em formato de árvore. As regras são demonstradas de maneira individual, como sendo os diversos trajetos de uma árvore, considerando um conjunto de afirmações e uma interpretação da classe. Desde que a precisão de acerto da classificação do sistema seja satisfatória, isso pode trazer mais facilidade para a interpretação final. Por exemplo:

- televisores = 0-9 AND retroprojetores = 0-4 AND computadoresAdm = 10-19 AND fax = True => enemMediaObjetiva = 500-599
- energiaGerador = False AND televisores = 20-29 => enemMediaObjetiva = 500-599

- energiaGerador = False AND impressoras = 0-9 AND televisores = 20-29 AND areaVerde = False => enemMediaObjetiva = 400-499
- impressoras = 0-9 AND computadoresAdm = 40-49 AND quadraDescoberta = True AND despensa = False => enemMediaObjetiva = 600-699

Devido a este conhecimento mais explícito vindo diretamente da saída dos algoritmos, o processamento dos dados deste trabalho, descrito no capítulo 4, utilizará tais algoritmos. Com esse conhecimento claro, os gestores podem tomar decisões mais diretamente e mais facilmente.

2.4. SELEÇÃO DE ATRIBUTOS

A seleção de atributos é uma tarefa, executada a partir de um grupo de técnicas, cujo principal objetivo é escolher um subconjunto das variáveis de entrada ao eliminar características que sejam julgadas irrelevantes ou que não tenham informações preditivas. Esta área tem se mostrado efetiva em aumentar a eficiência da aprendizagem, a precisão da classificação (em especial, na área de aprendizagem de máquina supervisionada) e reduzindo a complexidade do modelo em geral (ALMUALLIM e DIETTERICH, 1994; KOLLER e SAHAMI, 1996).

Esta etapa, normalmente executada antes do processamento dos dados, busca o espaço dos subconjuntos de atributos, avaliando cada um deles. Isto é obtido ao combinar um avaliador de um subconjunto de atributos com um método de busca que devolverá tais subconjuntos. Um dos métodos de busca do melhor subconjunto envolve estabelecer um ranking dos atributos (*Ranker*). Este método determina a importância de cada característica individual baseando-se em estatísticas, teoria da informação ou em algumas funções da saída do classificador (DUCH, WINIARSKI, *et al.*, 2003). Outros métodos utilizam abordagens baseadas em algoritmos genéticos (GOLDBERG, 1989) e em busca gulosa (*GreedyStepwise*), seja *forward* ou *backward*, cujo ponto de parada seja um decréscimo na avaliação ao adicionar ou retirar atributos do subconjunto atual (BUTTERWORTH, 2004; GEVREY, DIMOPOULOS e LEK, 2003; FREITAG e CARUANA, 2017).

2.5. FERRAMENTAS

Existem diversas ferramentas disponíveis para auxiliar no processo de descoberta de conhecimento através de mineração de dados. Algumas inclusive não possuem um foco científico, sendo mais voltada ao usuário final do que ao pesquisador, com interface gráfica

intuitiva. Outras se voltam a formatos específicos (e muitas vezes proprietários) e afirmam ser o mais abrangente possível em cálculos e amostragens estatísticas e na demonstração dos resultados.

Um deles é o SPSS Statistics, criado em 1968 e originalmente chamado de *Statistical Package for the Social Sciences*, ou Pacote estatístico para as Ciências Sociais[4]. As primeiras versões utilizavam processamento em lotes em mainframes (grandes computadores) com cartões perfurados, e em 1984 saiu a primeira versão para computadores pessoais. Além de ser concebido e ainda muito utilizado para análise estatística em Ciências Sociais, é também usado por pesquisadores de mercado, saúde, educação, e ainda empresas de marketing, mineração de dados e por governos. Foi adquirido pela IBM em 2009. O SPSS é distribuído numa licença *Trialware*, onde o usuário pode utilizar o programa completo até o fim do período de testes, e então é revertido para um modo com funcionalidades reduzidas, ou até mesmo não funcional, caso o usuário não pague a licença.

Outra conhecida ferramenta é o SAS[5]. Inicialmente conhecido como *Statistical Analysis System*, é uma suíte de software lançada em 1976 pela *North Carolina State University*, e hoje desenvolvido pelo *SAS Institute*, para análise avançada de diversos tipos, como análise multivariada, análise preditiva, inteligência de negócio e gerenciamento de dados. O programa permite minerar, alterar, gerenciar e recuperar dados de diferentes fontes, e ainda operar análise estatística nesses dados. Ele também possibilita analisar mídias sociais, minerar texto, e definir variáveis de gerenciamento de risco para bancos e empresas de serviços financeiros. O SAS é distribuído sob uma licença proprietária. Os microdados do ENEM são divulgados em formato *csv*, bem como (em alguns anos) em formatos de uso do SAS e do SPSS, permitindo aos usuários de tais plataformas utilizá-los e extrair conhecimento.

O RapidMiner[6] é uma plataforma de software de *data science* que provê um ambiente integrado para pré-processamento de dados, aprendizagem de máquina, *deep learning*, mineração de texto, análise preditiva e outros. É utilizado para aplicações comerciais e de negócios bem como para pesquisa, educação, treinamento, prototipação e desenvolvimento de aplicações e suporta todas as etapas do processo de aprendizagem de máquina, incluindo a visualização de dados, validação do modelo e otimização (HOFMANN e KLINKENBERG, 2013). Sua utilização pode ainda ser estendida por meio de scripts R e Python. Ele possui diferentes versões dependendo do uso, onde a versão gratuita (distribuída numa licença AGPL)

[4] http://www.redbooks.ibm.com/abstracts/sg248057.html
[5] http://www.sas.com/en_us/company-information.html
[6] htpp://rapidminer.com

permite o uso de 1 processador lógico e 10.000 linhas de dados. As outras versões são proprietárias.

O Weka (*Waikato Environment for Knowledge Analysis*) é uma suíte de software de aprendizagem de máquina, que contém uma coleção de ferramentas de visualização e algoritmos para análise de dados e modelagem preditiva, além de interfaces gráficas para o usuário acessar tais funções com mais facilidade (WITTEN, FRANK, *et al.*, 2016). A versão inicial foi projetada como uma ferramenta para analisar dados de agricultura (GARNER, CUNNINGHAM, *et al.*, 1995), mas desde o Weka 3 (de 1997) é possível utilizá-lo em diversas áreas, em particular para propósitos educacionais e de pesquisa. O Weka é distribuído gratuitamente sob licença GNU *General Public License* e é bastante portátil, por ser implementado em Java. Ele dá suporte a diversas tarefas que são padrão em mineração de dados, como pré-processamento, agrupamento, classificação, regressão, visualização e seleção de atributos. O software se baseia em processar um arquivo único que contém os dados descritos num número fixo de atributos, mas também permite conexão com bancos de dados SQL e processar o resultado devolvido por uma query no banco de dados.

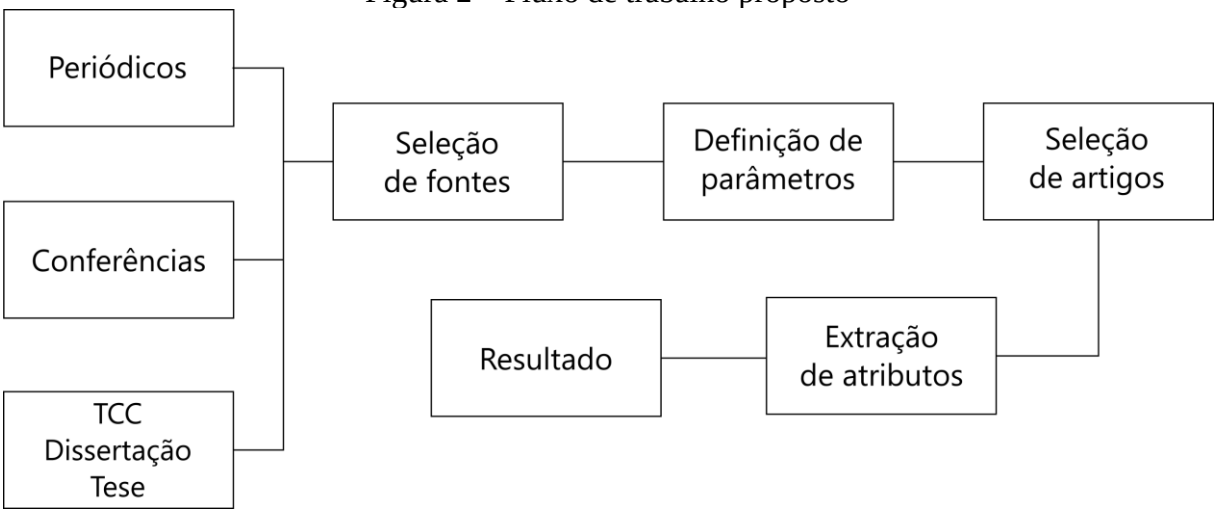
Por conter uma série de algoritmos de aprendizagem de máquina, incluindo diversos algoritmos de árvores e regras, além de vários filtros de seleção de atributos, o presente trabalho utiliza o Weka.

Capítulo 3

3. REVISÃO DA LITERATURA

Este capítulo discute acerca da revisão da literatura realizada para este trabalho. É importante notar que a revisão buscou apenas trabalhos brasileiros e que tenham relação direta com o tema deste trabalho, por se tratar de descoberta de conhecimento em dados relacionados à educação brasileira. Para a realização da revisão de literatura, a fim de prospectar trabalhos relacionados para dar suporte a este trabalho, foram definidas algumas etapas formais que caracterizam um fluxo de atividades sequenciais, seguindo o modelo proposto em Peña-Ayala (2014). Este fluxo adaptado é mostrado na Figura 2.

Figura 2 – Fluxo de trabalho proposto



Fonte: o autor

3.1. SELEÇÃO DAS FONTES

A etapa de seleção das fontes foi baseada principalmente em duas categorias: periódicos e conferências. Além disso, também houve busca de referências de teses, dissertações e trabalhos de conclusão de curso. Em todas as categorias, foram buscados trabalhos publicados no Brasil e que tenham alguma relação com educação ou sistemas computacionais.

Na etapa de parâmetros de busca, foi definido que a pesquisa seria realizada em páginas de conferências e periódicos da área de Informática na Educação, no Portal Periódicos da

CAPES[7] e na ferramenta *Google Scholar*. Além disto, as palavras chaves utilizadas para as buscas foram: “Dados abertos educacionais”, “dados abertos + educação”, “INEP”, “IDEB” e ”ENEM”. As duas primeiras são termos mais genéricos; por outro lado as últimas representam os principais tipos de dados abertos utilizados em trabalhos de educação. Em particular, vale salientar que foram considerados os trabalhos publicados nos últimos 10 anos (2008-2017).

Na terceira etapa foram selecionados aqueles artigos considerados relevantes. Foram excluídos artigos que: (i) estavam fora do período da pesquisa; (ii) não utilizam dados abertos educacionais, por exemplo artigos relacionados a recursos educacionais; e (iii) não utilizam dados abertos para aplicações educacionais.

3.2.RESULTADOS DA SELEÇÃO

A busca inicial resultou em 47 artigos, e após a etapa de seleção 33 trabalhos foram considerados relevantes para o estudo. A Tabela 1 mostra a distribuição dos artigos em relação ao veículo no qual os artigos foram publicados. O grande número de artigos publicados em conferências mostra que a exploração de dados abertos educacionais na literatura brasileira ainda está em fase inicial.

Tabela 1 – Distribuição dos artigos por categorias

Categoria	**Quantidade**
Conferências	16
Periódicos	8
Workshop	5
Outros	4
Total	**33**

Fonte: o autor

Poucos trabalhos foram publicados, e menos ainda utilizam técnicas computacionais para extrair informações dos dados. Isso explica a grande concentração de artigos em eventos como Workshop de Informática na Escola (WEI) e *Taller Internacional de Software Educativo* (TISE) que aceitam mais análises qualitativas e estudos de caso. Como se pode ver na tabela 2, a conferência que teve mais artigos relevantes foi o Simpósio Brasileiro de Informática na

[7] http://periodicos.capes.gov.br/

Educação, enquanto o periódico com maior número de publicações foi a Revista Brasileira de Informática na Educação. Por fim, foram encontrados 4 trabalhos classificados como outros, estes incluem os TCCs e dissertações.

Tabela 2 – Distribuição dos artigos por tipo de fonte

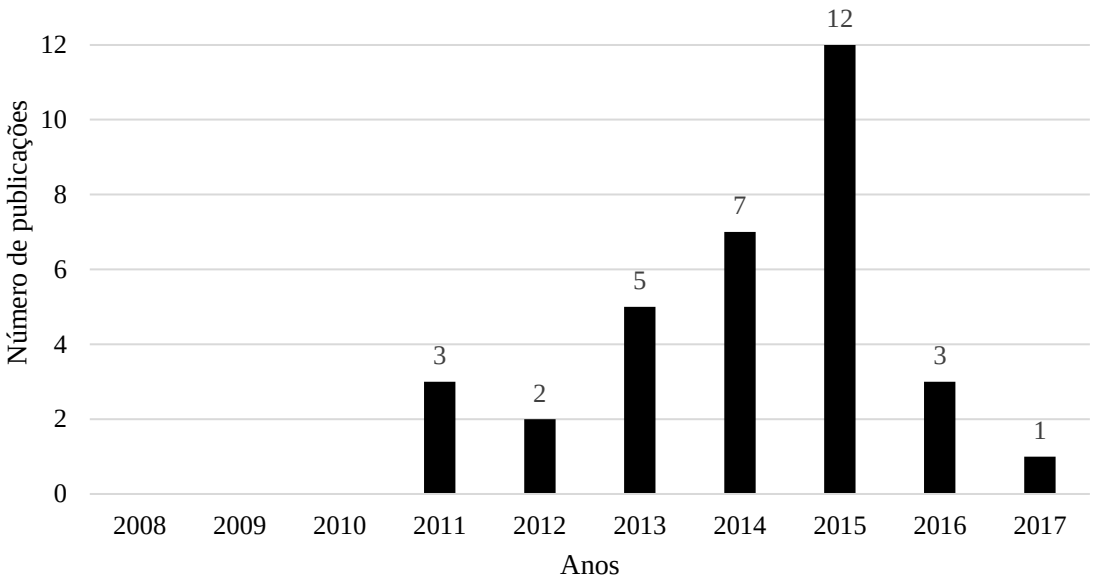
Fonte		**Quantidade**
Workshop	WIE	5
Conferências	TISE	3
	SBSI	1
	SBIE	7
	DESAFIE	2
	SBSR	1
	SBSeg	1
	KDMiLe	1
Periódicos	RBIE	2
	ISys	1
	InCID	1
	TCU	1
	Revista Mediação	1
	Revista Percurso	1
	Revista EAE	1
Outros		4

Fonte: o autor

Outro dado para confirmar que a área estudada neste trabalho ainda está começando a se desenvolver é a distribuição dos artigos ao longo dos anos, como mostra a Figura 2. Apesar da busca ter sido realizada para artigos a partir do ano de 2008, apenas em 2011 o primeiro

artigo foi encontrado. Além disto, a quantidade de publicações no tema atingiu o maior número em 2015 com 12 artigos.

Figura 3 – Número de publicações ao longo dos anos



Fonte: o autor

Os principais objetivos dos artigos selecionados, apresentados na Tabela 3, são:

- Mineração de dados aplicados a dados abertos educacionais: Dentro desse objetivo encontramos três tipos de artigos, sistemas para apoiar o gestor na tomada de decisão; predição de desempenho de escolas; e análise de algoritmos de aprendizagem de máquina focados para dados abertos educacionais.
- Análise qualitativa: Análise relativa a qualidade de dados abertos educacionais, possibilidades de aplicações e estudos de contextos escolares que se adequam à utilização dos mesmos.
- Teórico: Nesta categoria encontram-se artigos que detalham métodos e desafios sobre a utilização de dados abertos educacionais.
- Sistema para apoio a aprendizagem: Utilização de dados abertos para apoiar a aprendizagem colaborativa.
- Visualização dos dados: Aplicações que possibilitam uma melhor visualização dos dados abertos educacionais de uma determinada região.

Tabela 3 – Objetivo principal do artigo

Categoria	**Quantidade**
Mineração de dados aplicados a dados abertos educacionais	11
Análise qualitativa	9
Teórico	7
Sistema para apoio à aprendizagem	3
Visualização dos dados	3

Fonte: o autor

Por fim, em junho de 2017 foi coletado no Google Scholar o número de citações de cada artigo, para estabelecer um ranking dos trabalhos referenciados na literatura. Esse resultado, descrito na Tabela 4, evidencia mais uma vez que esta área ainda é pouco explorada, pois os principais trabalhos estão classificados como teóricos ou de análises qualitativas. O que mostra que ainda existe muito espaço para a utilização de técnicas computacionais para extrair informações de dados abertos educacionais.

Tabela 4 – Trabalhos mais citados

Título	**Nº Citações**
Mineração de Dados Educacionais: Oportunidades para o Brasil (BAKER, ISOTANI e CARVALHO, 2011)	84
Uma escala para medir a infraestrutura escolar (SOARES NETO, DE JESUS, *et al.*, 2013)	72
Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades (RIGO, CAZELLA e CAMBRUZZI, 2012)	13
O uso das TIC na formação de professores de escolas que obtiveram baixo IDEB (COELHO NETO, BERNARDELLI, *et al.*, 2011)	4
Avaliando as competências escolares através da Prova Brasil usando ferramenta web (PINHEIRO, ELIA e SAMPAIO, 2013)	4

Fonte: o autor

3.3. DISCUSSÃO E TRABALHOS RELACIONADOS

É perceptível que, apesar de existirem muitas fontes de dados, as pesquisas ainda estão em estágio inicial de desenvolvimento. Existem poucos artigos relacionados a este tópico, porém também fica bem evidenciado o potencial de desenvolvimento e interesse principalmente pelo grande crescimento de publicações nos últimos anos. Grande parte dos trabalhos encontrados na literatura são análises qualitativas sobre possíveis aplicações de mineração de dados educacionais (BAKER, ISOTANI e CARVALHO, 2011). Contudo, mesmo sendo estudos preliminares, os artigos apontam vários benefícios da aplicação destes dados. As principais aplicações estudadas são:

- Avaliação do ensino em vários níveis, desde escolas até estados (PINHEIRO, ELIA e SAMPAIO, 2013);
- Utilização de tecnologia na sala de aula (GENEROSO, NETO, *et al.*, 2013);
- Análise dos dados educacionais existentes (DE ASSIS RODRIGUES, SANT'ANA e FERNEDA, 2015);
- Combate à evasão (RIGO, CAZELLA e CAMBRUZZI, 2012).

Outro ponto importante a se destacar é a utilização de técnicas de mineração de dados aplicadas aos dados abertos educacionais. A partir de dados disponibilizados pelo INEP, diferentes trabalhos propõem a utilização de algoritmos de aprendizagem de máquina, como regressão logística e árvore de decisão, para extrair informações úteis para gestores de escolas. Adeodato *et al.* (ADEODATO, SANTOS FILHO e RODRIGUES, 2014) utilizaram os microdados do ENEM 2011 e do Censo Escolar 2011 para extrair conhecimento sobre a educação do ensino médio do Brasil. O trabalho focou em escolas privadas e utilizou a média aritmética das notas dos alunos como medida de qualidade das escolas, a fim de definir um limiar para marcar uma escola como sendo “boa”. Utilizou-se os três grãos de dados (aluno, escola, professor de escola) e foi construído um Sistema de Suporte à Decisão, a partir de regressão logística para pontuar a influência das variáveis do sistema, árvore de decisão – C4.5 (QUINLAN, 1993) – para explicitar sequências lógicas de regras compreensíveis por especialistas humanos, e regras de associação para detectar condições de alta relevância. Observou-se que um dos atributos mais importantes era a quantidade de banheiros que o aluno tinha em casa. Isso demonstra a forte influência dos aspectos econômico-financeiros, seja direta (pela renda familiar, destacada pela árvore de decisão) ou indiretamente (opção do ProUni, destacada pela árvore, e número de banheiros em casa, tanto pela árvore como pela seleção de

atributos) ou ainda em aspectos culturais (nível de educação da mãe ou do pai, evidenciados pela regra de associação) da família.

De maneira semelhante, Ferreira (FERREIRA, 2015) utilizou dados do Censo Escolar da Educação Básica de 2014 num escopo reduzido à cidade de Porto Alegre, RS, e com dados de mais de 19 mil alunos obteve precisão de mais de 95% nas classificações dos dados com o algoritmo C4.5, seja com o conjunto total de atributos (176) ou com o conjunto de atributos selecionados automaticamente (10). Além do que já foi descrito no capítulo 1, onde o trabalho concluiu que a presença de "recursos de internet banda larga, laboratório de ciências, auditório na escola e ensino privado apresentaram forte ligação com maiores chances de o aluno apresentar o Ensino Fundamental concluído" (FERREIRA, 2015), a inversa também é verdadeira, contribuindo negativamente. Acrescentando a isso, turmas que desfrutam de inglês, espanhol, artes e outras disciplinas não obrigatórias têm mais chance de concluir o Ensino Fundamental; entretanto, alunos com necessidades especiais estão mais propensos a não concluir os estudos.

Com intuito de facilitar a visualização de dados abertos educacionais, a plataforma CultivEduca (CARVALHO, NEVES e MELO, 2016) extrai anualmente dados do Censo Escolar da Educação Básica para apresentar informações relativas a perfis de docentes, como o estado e a condição de formação inicial e continuada dos profissionais, em diferentes estados do Brasil. Além disto, os trabalhos (GUERRA, NAKAMURA e HRUSCHKA, 2014) e (DA SILVA, 2014) propõem que a visualização dos dados pode auxiliar gestores de cidades e estados a tomar decisões sobre investimentos em escolas. Os dados abertos também podem ser usados para auxiliar a aprendizagem através de sistemas de apoio à pesquisa e sistemas para auxiliar a aprendizagem colaborativa (FRITZEN, SIQUEIRA e ANDRADE, 2013; DE SOUSA e DA SILVA, 2015)

Por fim, é importante destacar trabalhos mais teóricos e relacionados a características das escolas, disponibilização de dados abertos e avaliação de algoritmos de aprendizagem de máquina. Santos *et al.* (SANTOS, CLARO, *et al.*, 2014), utilizando KDD e regras de associação em base de dados do Censo Escolar 2011, discute que a infraestrutura das escolas impacta a qualidade do ensino das escolas apontando a eficiência trazida quando o governo investe na educação e a implicação no aprendizado dos alunos a partir da qualidade de ensino resultante da infraestrutura das escolas. O trabalho "Dados abertos conectados para a educação" (BANDEIRA, ÁVILA, *et al.*, 2015) apresenta boas práticas de como se publicar dados abertos e detalha vantagens e limitações do cenário atual dos dados abertos educacionais. Em "Comparação de algoritmos do aprendizado de máquina aplicados na mineração de dados

educacionais" (DE SOUZA, 2015) é apresentada a avaliação de algoritmos supervisionados aplicados a dados abertos educacionais.

Todos esses trabalhos utilizam mineração de dados educacionais. Neles se notam a necessidade e a importância de ampliar os experimentos e as pesquisas nesta área. As diferenças entre nosso trabalho e os citados acima estão na utilização de dados específicos a escola, na utilização de mais algoritmos (exceto em de Souza 2015, porém o foco era a comparação de algoritmos e não a interpretação dos dados educacionais) de árvore de decisão e regras de associação, e engloba ensino privado e público das escolas do Nordeste brasileiro.

3.4.OUTRAS INICIATIVAS

Além dos trabalhos científicos, outras iniciativas têm surgido no contexto de dados abertos educacionais no Brasil e no mundo. Em 2014, no Reino Unido, ocorreu um desafio chamado Education Open Data Challenge[8]. O desafio consistia em criar soluções que ajudassem pais e responsáveis a tomar decisões sobre seus filhos, expressando preferência por alguma escola, escolhendo um assunto ou alguma prioridade de aprendizagem, ou ainda praticar com a aprendizagem dos seus filhos. O vencedor foi a aplicação *Skills Route*, que foi projetada para mostrar opções de cursos e instituições de ensino de graduação disponíveis traçando parâmetros para projetar opções de carreira e até salários possíveis.

O QEdu[9] é uma plataforma na web que agrega dados de avaliações e indicadores nacionais, como a Prova Brasil, o Censo Escolar, IDEB e ENEM. Através do cruzamento de informações, a ferramenta auxilia gestores, professores e outros interessados a fazerem melhores escolhas na educação. Para isso, além de disponibilizar os dados, trabalha em outras frentes: Blog, onde geram e discutem ideias; Academia, que explica o funcionamento e o propósito dos indicadores e do próprio QEdu; e Redes, que fornece uma síntese de dados para redes municipais e estaduais.

O Reduca[10] é a Rede Latino-Americana de Organizações da Sociedade Civil pela Educação. O Observatório Educativo do Reduca é uma plataforma que reúne indicadores e informações educacionais dos 14 países das organizações que compõem a rede, o que inclui o Brasil. Seu objetivo é dar visibilidade aos dados educacionais da região para que a sociedade monitore e avalie políticas públicas educativas e promova o intercâmbio de boas práticas,

[8] https://theodi.org/education-open-data-challenge-series
[9] http://www.qedu.org.br
[10] http://www.todospelaeducacao.org.br

auxiliando gestores, pesquisadores e formuladores de políticas públicas a desenvolver soluções para os desafios educacionais. As fontes de dados do Observatório vêm de organismos internacionais como a Unesco e a Comissão Econômica para a América Latina e o Caribe – Cepal, além de censos demográficos, pesquisas domiciliares e avaliações nacionais elaborados por instituições de cada país da região.

Em vista de tais afirmações, nota-se a relevância do tema, e a importância de se estudar a relação de causa e efeito entre os atributos das escolas brasileiras e as notas médias do ENEM obtidas por cada escola.

Capítulo 4

4. METODOLOGIA

Esta seção detalha a metodologia de KDD adotada no fluxo do processo de pesquisa e desenvolvimento do trabalho. A metodologia proposta consiste de três etapas: Definição dos dados (I), Seleção, Pré-processamento e Transformação dos dados (II) e Mineração dos dados (III). Algumas dessas etapas possuem subetapas, que serão descritas em cada tópico. O principal objetivo é encontrar atributos que possam ser relevantes para o desempenho de estudantes das escolas no ENEM. Em outras palavras, a ideia com esse experimento é identificar o que maximiza a média das escolas no ENEM.

4.1. FONTE DOS DADOS

Nem todos os dados do INEP são publicados segundo os conceitos internacionais de dados abertos. Os formatos são dos mais variados, desde consultas a páginas HTML até "sistemas" feitos no Microsoft Excel, passando por PDFs e grandes arquivos compactados. Além disso, há dados que não estão disponíveis para processamento em máquina, como dados de localização de escolas. Esta heterogeneidade leva a dificuldades na utilização desses dados pela população, o que inclui o desenvolvimento de aplicações que processem e consumam tais dados.

A partir da iniciativa do Concurso INOVAPPS, promovido pelo Ministério das Comunicações e financiado pelo Governo Federal, através do edital Nº 11/2014/SEI-MC, surgiu o portal Educação Inteligente[11]. O portal foi um dos vencedores do concurso, e seu objetivo principal é "auxiliar a melhoria da educação básica no Brasil, sendo uma plataforma de captura, análise, manipulação e publicação dos dados governamentais sobre a educação" (LABORATÓRIO DE DADOS ABERTOS BRASIL, 2015).

Todos os dados utilizados no site provêm de fontes públicas, principalmente os dados do INEP, como os Micro Dados do Censo Escolar, do ENEM (incluindo o ENEM por Escola), as Estatísticas do IDEB, mas há dados como os Resultados e Metas do IDEB e o

[11] http://educacao.dadosabertosbr.com

DataEscolaBrasil, que tem portais abrigados no INEP, porém, os dados de cada avaliação não são do INEP, mas sim pertencentes ao domínio na *web* que cada avaliação tem. Todos esses dados coletados foram organizados, correlacionados e republicados, e o portal disponibiliza uma API para acesso aos dados, via HTTP/REST e resposta no formato JSON, para melhorar a disponibilização desses dados, incentivando o uso dos mesmos para geração de novas iniciativas por terceiros.

Os dados e estatísticas disponíveis no Educação Inteligente são referentes ao ano de 2013. Pela importância dada ao ENEM pelo próprio Governo Federal, e também em acordo com alguns dos trabalhos relacionados, será utilizada a nota média da prova objetiva do ENEM obtida por cada escola como classe. A justificativa é utilizar uma única métrica, daí a média; e excluir avaliações subjetivas, no caso, a nota da Redação, daí a utilização da média da prova objetiva e não a média geral, que contém a Redação.

Os dados utilizados são referentes à região Nordeste, devido ao alto tempo de processamento caso utilizássemos todos os dados, em escala nacional. Não houve especificação de estado dentro da base, isto é, mesmo havendo informação de cidade e estado a que a escola pertença, não houve busca por padrões ou informações específicas de um ou outro estado, e sim da região Nordeste como um todo.

4.2. SELEÇÃO, PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS

Para executar esta etapa, foi criado um único script de limpeza, transformação e preparação dos dados. A seguir, serão descritas as diferentes etapas do KDD.

4.2.1. Seleção e pré-processamento dos dados

O pré-processamento dos dados geralmente consiste de três sub-etapas: (i) limpeza de informações ausentes, retirando valores ausentes em conjuntos de dados; (ii) limpeza de inconsistências, que tanto trata de identificar quanto de retirar valores incertos; (iii) limpeza de valores não pertencentes ao domínio, excluindo o que não pertença ao domínio dos atributos do problema. Nesta etapa também se aplicam métodos para reduzir os dados antes do processamento, caso exista restrição de memória ou de tempo de processamento devido a uma quantidade muito grande de atributos para serem correlacionados.

Foi definido que nesta pesquisa seriam selecionados os dados relativos a escolas do Nordeste que tivessem alunos participantes do ENEM. Como já foi citado, os dados são referentes ao ano de 2013, de acordo com a base de dados do Educação Inteligente. Além disto, alguns atributos foram retirados do conjunto de dados. Os critérios para exclusão foram os seguintes:

- Como o principal objetivo do trabalho é encontrar os atributos para maximizar a nota do ENEM levamos em consideração apenas a média final do ENEM como atributo alvo. Outras notas do ENEM que estavam presentes foram removidas. Por exemplo, notas de Ciências Exatas ou de Ciências Humanas e suas Tecnologias. Isso foi necessário, pois existe uma alta correlação entre as notas individuais e a nota média. Com isso, apenas esses atributos eram selecionados pelos algoritmos.
- Atributos Redundantes. Por exemplo, códigos e nome da cidade onde a escola se localiza. De semelhante maneira, com distritos e Unidades Federativas (UFs).
- Foram retirados alguns atributos de localização de escola: nome, endereço, latitude, longitude.
- Atributos relacionados a outras avaliações, como o IDEB, tanto porque mais da metade das escolas sequer tinha índices ou notas de outras avaliações, como porque a proposta é utilizar o ENEM como avaliação única.

4.2.2. Transformação dos dados

A próxima etapa do KDD é a transformação dos dados. A intenção aqui é ajustar os dados ao formato de entrada do algoritmo de MD, ou no caso, ao formato de entrada aceitável pelo WEKA, o software escolhido para a etapa de MD. Além disso, os dados podem ser agregados, generalizados ou até mesmo ter novos atributos adicionados para melhor compreensão do problema. Um processo comum é a transformação de valores reais em intervalos ou categorias.

Neste trabalho, muitas variáveis no banco de dados são numéricas booleanas, contendo apenas 0 ou 1, e o Educação Inteligente os exibe como “possui” ou “não possui”. Por exemplo: uma escola pode possuir ou não uma Biblioteca; pode possuir ou não uma Quadra Coberta. Esses dados foram transformados para True ou False.

Outros dados numéricos, que não são booleanos, foram transformados em faixas de valores, por exemplo: o parâmetro de Número de Computadores na escola está em faixas de 20,

sendo de 0 a 19, 20 a 39, 40 a 59 e assim por diante, mas limitando a 10 intervalos, o que se traduz num intervalo máximo de 199 itens; acima disso, o intervalo é “acima de 200”. É importante dizer que não houve balanceamento, apenas a discretização dos dados.

Por fim, a nota do ENEM, por estar na faixa de 300 a 1000, está dividida em 7 faixas de 100 pontos. Desta forma, todos os atributos numéricos foram transformados para booleanos ou categóricos. A Tabela 5 apesenta alguns exemplos de dados transformados:

Tabela 5 – Exemplos de dados transformados

Atributo	**Exemplo**	**Após transformação**
refeitorio	1	True
salaLeitura	0	False
salasExistentes	126	120-134
enemMediaObjetiva	679.75	600-699

Fonte: o autor

4.3.MINERAÇÃO DE DADOS

Esta etapa é responsável por extrair conhecimento a partir dos dados já processados. Como a proposta do trabalho é encontrar a relação entre os atributos das escolas que influenciam as suas notas no ENEM, foram utilizados algoritmos de árvore de decisão e regras de associação. Essa escolha se deu porque eles geram regras de relacionamento explícitas entre os atributos.

Para esta etapa foi utilizada a ferramenta Weka (*Waikato Environment for Knowledge Analysis*), que provê diferentes técnicas de mineração de dados e seleção automática de atributos (HALL *et al.*, 2009). Seu uso se baseia em ler um arquivo de extensão *.arff*[12]. Um arquivo ARFF (*Attribute-Relation File Format*) é um arquivo de texto plano, sem formatação, que descreve uma lista de instâncias que compartilham um conjunto de atributos. O arquivo ARFF tem duas seções distintas: o *Header* (cabeçalho) e o *Data* (dados), e são definidos por

[12] http://weka.wikispaces.com/ARFF+%28book+version%29

relações, através do uso de @ (arroba). Em qualquer seção, uma linha iniciada por % (porcentagem) é interpretada como comentário.

O *Header* contém a declaração da relação, anotado por:

- @RELATION <nome da relação>

e uma lista de atributos, anotados por:

- @ATTRIBUTE <nome do atributo> <tipo do dado>

onde o <nome do atributo> tem que começar com caractere alfabético. Se houver espaço no nome, então o nome inteiro deve estar entre aspas. O Weka dá suporte a quatro tipos de dados, sendo eles:

- numeric (onde inteiros e reais são tratados como numeric)
- <especificação nominal> (que lista um conjunto de valores possíveis, como {<opção1>, <opção2>, <opção3>,...})
- string
- date [<formato de data>] (como yyyy-MM-dd ou outros, segundo a norma ISO-8601.

Um exemplo de cabeçalho de um ARFF se encontra na Figura 4 abaixo:

Figura 4 – Exemplo de cabeçalho (Header) de arquivo ARFF do Weka.

```
@RELATION escolas-nordeste-enem-2013

% @ATTRIBUTE cod STRING
@ATTRIBUTE aee {True,False}
@ATTRIBUTE aguaCacimba {True,False}
@ATTRIBUTE aguaFiltrada {True,False}
@ATTRIBUTE aguaInexistente {True,False}
@ATTRIBUTE aguaPocoArtesiano {True,False}
@ATTRIBUTE aguaPublica {True,False}
@ATTRIBUTE aguaRio {True,False}
@ATTRIBUTE alimentacao {True,False}
@ATTRIBUTE almoxarifado {True,False}
@ATTRIBUTE alojamentoAluno {True,False}
@ATTRIBUTE alojamentoProfessor {True,False}
@ATTRIBUTE anoCenso {2013}
@ATTRIBUTE aparelhosSom {None,0-4,5-9,10-14,15-19,20-24,25-29,30-34,35-39,40-44,45-49,"acima de 50"}
@ATTRIBUTE areaVerde {True,False}
@ATTRIBUTE atendimentoEspecial {True,False}
@ATTRIBUTE atividadeComplementar {True,False}
@ATTRIBUTE auditorio {True,False}
@ATTRIBUTE bandaLarga {True,False}
@ATTRIBUTE banheiroChuveiro {True,False}
@ATTRIBUTE bercario {True,False}
@ATTRIBUTE biblioteca {True,False}
@ATTRIBUTE ciclos {True,False}
% @ATTRIBUTE codDistrito STRING
% @ATTRIBUTE codMunicipio STRING
% @ATTRIBUTE codUf STRING
@ATTRIBUTE computadores {None,0-19,20-39,40-59,60-79,80-99,100-119,120-139,140-159,160-179,180-199,"acima de 200"}
@ATTRIBUTE computadoresAdm {None,0-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99,"acima de 100"}
@ATTRIBUTE computadoresAlunos {None,0-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99,"acima de 100"}
@ATTRIBUTE copiadoras {None,0-4,5-9,10-14,15-19,20-24,25-29,30-34,35-39,40-44,45-49,"acima de 50"}
@ATTRIBUTE cozinha {True,False}
```

Fonte: o autor

A seção *Data* contém uma declaração @DATA, indicando o início da seção, e as linhas de instâncias de dados, onde cada linha é uma instância. Uma vírgula ou uma tabulação delimitam os valores de atributo de cada instância, e podem ser seguidos por zero ou diversos espaços (dependendo do tipo do valor), mas devem aparecer na ordem em que foram declarados na seção Header, isto é, o n-ésimo valor do dado corresponde aos valores possíveis do n-ésimo item @ATTRIBUTE. Valores em branco, se não forem tratados em pré-processamento, podem ser representados por uma interrogação.

Um exemplo de uma seção de dados de um ARFF consiste da seguinte maneira, na Figura 5 abaixo:

Figura 5 – Exemplo da seção de dados (*Data*) de arquivo ARFF do Weka (com quebra de linha).

```
@DATA
True,False,False,False,False,True,False,True,False,False,False,2013,0-4,True,True,False,False,True,
False,False,False,True,False,0-19,0-9,10-19,0-4,True,0-9,Estadual,True,True,0-4,False,False,False,False,F
alse,False,False,True,False,False,True,True,False,False,False,False,False,False,False,False,False,False,F
alse,False,False,False,True,0-4,40-59,0-9,True,True,True,False,False,False,False,False,True,False,"CURURU
PU",True,False,True,False,True,True,False,True,False,False,False,False,True,False,False,False,Não,0
-4,True,False,True,0-14,0-14,True,True,False,False,True,"MA",0,'Em
atividade',0-9,Urbana,0-4,400-499,400-499,500-599,70-79,"Baixo"
False,False,False,False,True,False,False,True,False,False,False,2013,0-4,False,False,False,False,Fa
lse,False,False,False,False,0-19,0-9,10-19,0-4,True,0-9,Estadual,False,True,0-4,False,False,False,F
alse,False,False,False,True,False,False,True,True,False,False,False,False,False,False,False,False,F
alse,False,False,False,False,0-4,20-39,0-9,True,False,True,False,True,False,False,False,True,False,
"MIRINZAL",True,False,True,False,False,False,False,False,False,False,False,False,True,False,False,F
alse,Não,0-4,True,False,True,0-14,0-14,True,False,False,True,False,"MA",0,'Em
atividade',0-9,Rural,0-4,400-499,400-499,400-499,30-39,"Baixo"
False,False,False,False,True,False,False,False,True,False,False,2013,0-4,False,True,False,True,True
,True,False,True,True,20-39,0-9,10-19,0-4,False,0-9,Privada,True,False,0-4,False,False,False,False,
False,False,True,False,False,True,True,False,True,False,False,False,False,False,False,False,False,F
alse,False,True,True,0-4,40-59,0-9,True,True,True,False,True,False,False,False,False,False,"PACO
DO
LUMIAR",True,True,False,True,False,True,False,False,False,False,True,False,True,False,False,True,Não
,0-4,True,False,True,30-44,15-29,True,True,False,True,True,"MA",0,'Em
atividade',0-9,Rural,0-4,400-499,400-499,500-599,60-69,"Médio Alto"
False,False,False,False,False,True,False,True,False,False,False,2013,0-4,False,False,False,False,Fa
lse,False,False,True,False,0-19,0-9,10-19,0-4,True,0-9,Estadual,False,True,0-4,False,True,False,Fal
se,False,False,True,True,False,True,True,False,False,False,False,False,False,False,False,False,Fals
e,False,False,False,False,0-4,20-39,0-9,True,False,True,False,True,False,False,False,False,False,"P
ACO DO
LUMIAR",True,False,True,False,False,False,False,False,False,False,False,False,True,False,False,Fals
e,Não,0-4,True,True,True,0-14,0-14,True,False,False,True,True,"MA",0,'Em
atividade',0-9,Urbana,0-4,400-499,400-499,400-499,60-69,"Médio Baixo"
```

Fonte: o autor

4.3.1. Seleção Automática de Atributos

Apesar de já ter sido feita uma seleção manual, o banco de dados ainda possuía muitos atributos. Por isso, foram utilizados diferentes algoritmos para seleção automática de atributos.

Os algoritmos de seleção que alcançaram melhores resultados quando aplicados a dados educacionais foram: *Chi-Square Attribute evaluation* (CH), *Information-Gain Attribute evaluation* (IG) (RAMASWAMI e BHASKARAN, 2009) e *ReliefF Attribute evaluation* (RF) (KIRA e RENDELL, 1992; KONONENKO, 1994; ROBNIK-ŠIKONJA e KONONENKO, 1997). Todos os algoritmos utilizados ordenam os atributos de acordo com algum critério, sendo chi-quadrados, ganho de informação e proximidade com vizinhança, respectivamente.

Em princípio, todos os algoritmos retornaram resultados similares, sendo que o atributo de maior relevância para a determinação da nota do ENEM foi o perfil socioeconômico do aluno. Além disso, o atributo localização da escola (município) também foi considerado relevante. Esse resultado mostra que as desigualdades sociais são muitas vezes refletidas nas notas do ENEM. Atributos relacionados a escola, como tipo de administração (pública, em diferentes esferas, ou privada), quantidade de salas existentes e utilizadas, e se tem pré-escolar com boa estrutura (parque infantil e sanitários educação infantil) também ficaram bem ordenados nos algoritmos utilizados.

Como o objetivo do trabalho é auxiliar gestores, o atributo do perfil sócio econômico dos alunos e município foi removido, pois ele não pode ser alterado diretamente pelo gestor. Outros atributos, como a dependência administrativa (se a escola é Privada, Federal, Estadual ou Municipal) e o tipo de localização (Urbana ou Rural) foram retirados, bem como os demais atributos de localização. Em testes prévios, foi criado um atributo de Região Metropolitana, com a premissa de que simplificaria a localização, reduzindo um atributo com mais de 1700 cidades em outro com 2 valores: Região Metropolitana ou interior. No entanto, os três algoritmos de seleção de atributos valorizaram o atributo de cidade e atribuíram baixo valor à Região Metropolitana, o que dificultaria a interpretação final das regras geradas. Com isso, o atributo de cidade, bem como o de Região Metropolitana, também foi retirado.

Também retiramos os que obtiveram valores de correlação com vizinhança e ganho de informação menores que 0,03 e chi-quadrados menores que 100, de acordo com os critérios dos algoritmos de seleção automática de atributos. Com isso, o número de atributos passou de 109 para 52 atributos, incluindo a classe, que é a nota do ENEM de cada escola. A Tabela 6 mostra os 20 atributos de maior relevância para cada algoritmo utilizado de forma ordenada.

Tabela 6 – Atributos selecionados por cada algoritmo

Algoritmo	Atributos
CH	alimentacao, salasExistentes, dvds, computadoresAdm, salasUtilizadas, parqueInfantil, sanitarioEducInfant, regPreescola, funcionarios, computadores, ensinoEja, impressoras, datashows, regCreche, ejaMedio, computadoresAlunos, aparelhosSom, quadraCoberta, despensa, fimDeSemana.
IG	alimentacao, parqueInfantil, sanitarioEducInfant, salasExistentes, regPreescola, computadoresAdm, salasUtilizadas, ensinoEja, ejaMedio, regCreche, funcionarios, computadores, impressoras, computadoresAlunos, quadraCoberta, atividadeComplementar, despensa, fimDeSemana, datashows, regFundamental.
RF	alimentacao, ensinoEja, ejaMedio, despensa, computadoresAlunos, computadores, computadoresAdm, funcionarios, cozinha, regMedioMedio, esgotoFossa, esgotoPublico, fax, regMedioIntegrado, regFundamental, formacaoDocente, fimDeSemana, salasExistentes, laboratorioCiencias, atividadeComplementar.

Fonte: o autor

4.3.2. Classificação dos dados

Com esses atributos, foram aplicados diferentes algoritmos baseados em regras e em árvore de decisão para classificar automaticamente as escolas em relação as suas notas do ENEM. Esses algoritmos permitem extrair relações entre os atributos utilizados. Para essa etapa, foram executados os algoritmos de árvore:

- C4.5 (QUINLAN, 1993)
- CART – *Classification And Regression Trees* (BREIMAN, FRIEDMAN, *et al.*, 1984)
- Best-First (SHI, 2007)

e os algoritmos de regras:

- RIPPER – *Repeated Incremental Pruning to Produce Error Reduction* (COHEN, 1995)
- Ridor – *Ripple-Down Rule learner* (GAINES e COMPTON, 1995)
- PART – *Partial C4.5* (FRANK e WITTEN, 1998)

Todas as técnicas implementadas utilizam uma etapa de treinamento e validação. No processo de avaliação aplicamos o método de validação cruzada, ou *cross-validation* (ARLOT e CELISSE, 2010) com *k-folds*. Nesse método, a amostra de dados é dividida em k subamostras

de tamanho igual, e de todas elas, uma é escolhida como o conjunto de teste, e as outras k – 1 subamostras serão o conjunto de treinamento. O processo de validação cruzada é então repetido k vezes (por isso *k-fold*) como se fossem k classificações independentes, onde cada uma das k subamostras é utilizada uma única vez como o conjunto de teste e as outras k – 1 como treinamento. Ao final do processo, a validação se dá pela média dos resultados de cada *fold*. Neste trabalho, utilizou-se 10-fold para esta etapa.

Os algoritmos foram utilizados com as suas configurações padrão disponíveis no Weka. A seção 5 trata dos resultados dessa classificação dos dados, tanto em termos de desempenho dos algoritmos com os atributos selecionados, quanto na interpretação das regras geradas.

Capítulo 5

5. RESULTADOS E DISCUSSÃO

Esta seção discorre acerca dos resultados da classificação dos dados, após as etapas executadas no capítulo 4. Trata-se da avaliação dos algoritmos, dos resultados dos seus processamentos e da discussão sobre tais resultados.

5.1. AVALIAÇÃO DOS ALGORITMOS UTILIZADOS

Conforme anteriormente mencionado, todos os algoritmos foram utilizados com as suas configurações padrão disponíveis no Weka. As medidas de avaliação dos algoritmos foram as seguintes (BAEZA-YATES e RIBEIRO-NETO, 2013):

- Acurácia: Mede a quantidade de instâncias de entrada que foram avaliadas corretamente, de acordo com a associação entre as instâncias e as classes pré-estabelecidas.
- *F-Measure*: é uma média harmônica entre Precisão e Cobertura, Equação 1.

$$F-Measure = 2 * \frac{P*C}{P+C} \quad (1)$$

Onde: Precisão (P) avalia a quantidade de instâncias que foram classificadas corretamente. E Cobertura (C) avalia a porcentagem instâncias de uma determinada classe que não foi classificada como pertencente a essa classe.

A Tabela 3 mostra que os algoritmos CART, Best-First e RIPPER alcançaram os melhores resultados tanto em termos de Acurácia quanto *F-Measure*.

Tabela 7 – Avaliação dos algoritmos

Algoritmo	**Acurácia**	***F-Measure***
CART	82,0776%	81,9%
Best-First	80,8505%	80,5%
RIPPER	79,8516%	79,7%

C4.5	79,0525%	78,8%
Ridor	78,71%	79,1%
PART	77,1119%	77%

Fonte: o autor

Antes de passarmos à interpretação das regras, é importante conhecer a distribuição dos dados em relação à classe, isto é, a maneira que as escolas estão distribuídas segundo as notas do ENEM. A Tabela 8 mostra que a maior concentração de escolas está entre as que obtiveram de 400 a 699:

Tabela 8 – Distribuição das escolas de acordo com a nota média do ENEM

Faixa de valores da nota	**Quantidade de escolas**
300-399	2
400-499	2166
500-599	1205
600-699	128
700-799	3
800-899	0
900-999	0
Total	**3504**

Fonte: o autor

5.2. REGRAS GERADAS

Há duas notações utilizadas nas regras a serem apresentadas. A primeira é a barra vertical (|) , que indica um “ou”: havendo, por exemplo, A|B|C|D, isso significa que as opções A, B, C ou D confirmam a condição mesmo se apenas uma delas for verdadeira e as outras forem falsas, mas a presença de todas as opções indica que instâncias da entrada podem ter sido

identificadas pelo classificador com qualquer uma das opções. A segunda notação é o símbolo != (exclamação seguido de igual), e é apenas uma representação do símbolo de diferença (≠).

Sobre os algoritmos, o classificador melhor avaliado pelas métricas utilizadas foi o CART. Sua execução entregou uma árvore pequena, com 15 folhas. Algumas regras interessantes estão abaixo:

Tabela 9 – Regras do algoritmo CART

Regras (iniciando da raiz)	**Implica em**
alimentacao = False **E** funcionarios = (0-19)|(20-39)|(40-59) **E** laboratorioCiencias = True	enem = 500-599 (301/129)
alimentacao = (False) **E** funcionarios = (0-19)|(20-39)|(40-59) **E** laboratorioCiencias = False **E** formacaoDocente = (0-9)|(20-29)|(30-39)|(40-49)|(10-19) **E** ejaMedio = True	enem = 400-499 (7/1)
alimentacao = True **E** regMedioMedio = True **E** parqueInfantil = False	enem = 400-499 (1710/98)
alimentacao = False **E** funcionarios != (0-19)|(20-39)|(40-59) **E** computadores=(20-39)|(0-19)|(40-59)|(60-79)	enem = 500-599 (355/78)

Fonte: o autor

A partir das regras do CART podemos realizar algumas inferências. Inicialmente, falando da natureza do algoritmo, é de implementação simples (talvez por isso no Weka é denominado *SimpleCart*) e utiliza regressão. Algumas das regras tiveram como alvo uma quantidade grande de dados, mas também obtiveram erros na classificação, conforme a notação (x/y) oriunda dos *outputs* das classificações. A primeira regra deixa claro que, mesmo uma escola sem alimentação e com um máximo de 59 funcionários, com laboratório de ciências obterá nota de 500-599. Justamente nesta regra, 301 instâncias são classificadas dessa maneira, embora a classificação de 129 delas estivesse marcada como errada, sendo 42,85% das instâncias.

Na terceira regra, a ideia presente na primeira regra é ainda mais evidente, onde, mesmo oferecendo alimentação e ensino médio comum, mas sem parque infantil, a nota do ENEM será

400-499, mas aqui se torna algo interessante: 1710 escolas de 3504 (48% do total) foram classificadas assim, com um erro de 98, ou 5,7%. Isso se apresenta como um padrão importante dentro dos dados.

Na segunda regra, inferimos que sem alimentação, menos de 60 funcionários, sem laboratório de ciências e com a formação do docente avaliada em até 50[13], mesmo com EJA de nível médio, a avaliação devolveu 400-499. Interpretamos que tanto uma limitação de opções oferecidas quanto o conhecimento do docente (no caso, o que esteja registrado nas avaliações governamentais) tem influência na aprendizagem, e uma má infraestrutura também pode limitar a experiência do ensino passado pelo docente.

Por fim, sem alimentação, mas com mais de 60 funcionários e até 80 computadores indicam 500-599, sendo mais um padrão relevante, com mais de 300 classificações nos dados.

O segundo algoritmo mais bem avaliado foi o Best-First. Seguem algumas regras:

Tabela 10 – Regras do algoritmo Best-First

Regras (iniciando da raiz)	**Implica em**
alimentacao = False E funcionarios = (0-19)|(20-39)|(40-59)	enem = 500-599 (558/343)
alimentacao = False **E** funcionários != (0-19)|(20-39)|(40-59) **E** computadores = (20-39)|(0-19)|(40-59)|(60-79) **E** regPreescola != False	enem = 500-599 (302/45)
alimentacao = False **E** funcionários != (0-19)|(20-39)|(40-59) **E** computadores != (20-39)|(0-19)|(40-59)|(60-79) **E** funcionários != (60-79)|(80-99)|(100-119) **E** auditório = False	enem = 600-699 (11/0)
alimentacao = True **E** regMedioMedio = False **E** computadores != (20-39)|(0-19) **E** formacaoDocente = (0-9)|(20-29)|(30-39)|(40-49)|(10-19)|(50-59) **E** quadraCoberta = False **E** computadoresAdm = (0-9)|(10-19)|(30-39)|(20-29)|(50-59)|(60-69)|(70-79)|(80-89)|(90-99) **E** funcionarios = (20-39)	enem = 400-499(3.0/1.0)
alimentacao = True **E** regMedioMedio = False **E** computadores != (20-39)|(0-19) **E** formacaoDocente = (0-9)|(20-29)|(30-39)|(40-49)|(10-19)|(50-59) **E** quadraCoberta = False **E** computadoresAdm = (0-9)|(10-19)|(30-39)|(20-29)|(50-59)|(60-69)|(70-79)|(80-89)|(90-99) **E** funcionarios != (20-39)	enem = 400-499(10.0/0.0)

Fonte: o autor

[13] Métrica presente nos dados do Censo Escolar; há informações sobre a formação inicial e contínua do docente.

Enquanto árvores de decisão clássicas (como o C4.5) expandem os nós baseados em profundidade, o Best-First se baseia em expandir através de busca gulosa e avaliação dos melhores nós através de redução de impureza. Este algoritmo realiza poda para evitar problemas de *overfitting* (ou sobreajuste, onde a modelagem se ajusta demais aos dados, deixando de representar a realidade, onde há alguns erros e desvios) e é recomendada a execução por validação cruzada para potencializar a poda e reduzir o sobreajuste. A natureza do algoritmo está expressa na primeira regra, em que uma escola sem alimentação e um máximo de 60 funcionários tem nota 500-599. A expansão de nós alcançou 558, mas quase 350 foram incorretos.

No entanto, outros nós possuem uma boa classificação, quase sem erros. A segunda regra possui um bom número resultante de acertos, e infere que sem alimentação, mais de 60 funcionários, até 79 computadores e com pré-escola, uma escola possui nota de 500-599. Esta pode ser uma informação útil às escolas de avaliação mais baixa.

Sobre avaliação mais baixa, as duas últimas regras têm um dado interessante. Notem que todas expansões dos nós são iguais, excetuando a última expansão, funcionários, mas o resultado foi o mesmo, 400-499. Se em termos educacionais isto indica que o número de funcionários não modificará a condição que os itens anteriores trouxeram, em termos computacionais indica a diferenciação na expansão do algoritmo no último nó, classificando 10 escolas de uma maneira e 3 de outra. Por fim, uma regra de resultado de alto nível indica que sem alimentação, com mais de 59 funcionários, mais de 79 computadores, menos de 60 funcionários (ou mais de 119) e sem auditório indicam nota 600-699.

O algoritmo de regras de decisão melhor avaliado foi o RIPPER, e ficou em terceiro na tabela 7. Ele devolveu 10 regras bem simples, as quais se encontram na tabela abaixo:

Tabela 11 – Regras do algoritmo RIPPER

Regras geradas	**Implicam em**
funcionarios = acima de 200 **E** alimentacao = False	600-699 (60.0/29.0)
alimentacao = False **E** formacaoDocente = 70-79 **E** datashows = 10-19	600-699 (15.0/7.0)
alimentacao = False **E** regPreescola = False **E** quadraCoberta = True **E** salasUtilizadas = 30-44 **E** salaLeitura = True	600-699 (9.0/2.0)
alimentacao = False **E** regPreescola = False **E** funcionarios = 80-99 **E** aparelhosSom = 0-4 **E** esgotoFossa = False	600-699 (14.0/4.0)
alimentacao = False	500-599 (1322.0/412.0)

regMedioMedio = False **E** formacaoDocente = 80-89	500-599 (38.0/7.0)
regMedioMedio = False **E** quadraCoberta = True	500-599 (112.0/41.0)
sanitarioEducInfant = True **E** parqueInfantil = True	500-599 (57.0/16.0)
computadoresAdm = 10-19 **E** fimDeSemana = False **E** patioDescoberto = True **E** ensinoEja = False	500-599 (10.0/1.0)
————	400-499 (1867.0/112.0)

Fonte: o autor

Inicialmente, a quinta regra infere apenas que sem alimentação a escola vai de 500 a 599. Obteve uma alta classificação, acima de 1300 registros, mas errou mais de 400 itens. Uma conjectura que fazemos é anterior aos dados que temos hoje: a maioria das escolas que não oferecem alimentação são privadas, provavelmente dando liberdade aos pais e responsáveis de custear ou não a alimentação dos próprios filhos no período que estão na escola. E na distribuição inicial de dados, as escolas privadas estão majoritariamente na classe 500-599, o que nos faz pensar que esta regra foi gerada por motivo que foge da alçada de gestores de escolas, mesmo utilizando um atributo que ele possa modificar. Esta informação é corroborada com a primeira regra, que acrescenta apenas acima de 200 funcionários na escola, e obteve um total de 31 acertos.

Por outro lado, uma escola com quadra coberta, sem pré-escola, 30 a 44 salas utilizadas e sala de leitura indica uma boa infraestrutura, resultando em nota 600-699. Essas características podem representar uma escola maior, mas que aproveita bem o espaço que tem, tanto oferecendo opções de aprendizagem como de lazer para os estudantes.

Por fim, a última regra (que na verdade não aponta nenhuma regra) apenas informa a classe 400-499 em mais de 1800 registros, errando 112 deles. Isto indica que, exceto pelas características citadas anteriormente, qualquer outro registro será marcado com 400-499 independente do que seja inferido. Aos gestores de escolas que estejam nessa faixa de nota, pode ser útil observar as características das regras anteriores geradas e identificar alternativas para beneficiar a aprendizagem.

5.3.DISCUSSÃO

Os resultados obtidos mostram que infraestrutura é relevante para uma escola, confirmando os trabalhos relacionados. Contudo, aspectos não-estruturais também são

importantes para melhorar o desempenho de uma escola em relação ao ENEM. Esse estudo não foi realizado pelos trabalhos anteriores, listados na seção 3. Além disso, em experimentos anteriores, atributos como tipo de administração e o já relatado índice socioeconômico se mostraram bastante influentes na avaliação. No entanto, como não podem ser diretamente modificados por gestores de escola ou mesmo por administradores públicos (pensando na educação formal, dentro dos limites internos da escola), precisaram ser retirados da classificação.

Essa retirada ocasionou uma queda na acurácia e na *F-Measure* de todos os algoritmos. Conforme relatado na tabela 7, o algoritmo com melhor desempenho foi o CART. Em experimentos prévios, contendo todos os atributos ainda antes da etapa de pré-processamento, sua acurácia era de cerca de 86%. Mesmo o C4.5 tinha uma acurácia de 83%. Isso demonstra que alguns fatores de desempenho de escola fogem da alçada dos gestores – embora até seja possível modificar o espaço comunitário ao redor da escola com investimentos em educação não-formal, fazendo da localidade uma extensão da escola. Na educação formal, índices socioeconômicos ainda estão fortemente ligados à qualidade das escolas brasileiras, mesmo quando retirados diretamente da classificação. Atributos como alimentação e quantidade de funcionários costumam refletir tais índices.

O fato da escola ser privada ou pública (nas diferentes esferas) também é um fator presente, seja para melhor ou para pior, e provavelmente indica a atenção e o investimento recebidos pela escola: nos dados obtidos, as escolas de nível socioeconômico acima de Médio-alto são majoritariamente privadas. Esse fato, juntamente com a diversidade de regras encontradas, também reflete a complexidade do ensino nacional: ainda que existam fatores fortes para o beneficiamento do desempenho através de infraestrutura e opções oferecidas aos alunos, e ainda que algumas alternativas possam ser aplicadas através da interpretação das regras obtidas, não há uma fórmula geral. Isto também é algo refletido em muitos dos trabalhos relacionados.

Utilizando métodos manuais e automáticos, a quantidade de atributos que foram considerados relevantes para determinar o desempenho das escolas foi 52 de um conjunto inicial de 109. Isso mostra a necessidade de utilização de algoritmos automáticos para extrair informações em grandes bases de dados, como é o caso dos dados abertos. Por fim, também é importante destacar que a partir da aplicação de algoritmos de mineração de dados é possível gerar relacionamento entre os atributos que podem auxiliar gestores a decidir quais ações tomar para melhorar as condições da escola.

Capítulo 6

6. CONSIDERAÇÕES FINAIS

Ainda é iniciante a área de mineração de dados educacionais no Brasil, mas cada trabalho publicado tem dado passos importantes na descoberta de padrões na educação nacional. Por outro lado, a própria publicação de tais dados precisa ser beneficiada, algo também citado por alguns dos trabalhos publicados. Isto favorece as pesquisas e a divulgação de conhecimento útil. Neste âmbito, o presente trabalho utilizou uma fonte de dados já melhorada, demonstrando que publicar dados em formato estruturado e atualizar as bases de dados abertas não só é possível como de fato enriquece a busca por conhecimento.

A principal proposta deste trabalho foi buscar informações em base de dados educacional, a fim de auxiliar gestores de escola e administradores públicos a incrementar as opções ofertadas por escolas. Para isso, definimos uma lista de atributos que uma escola possa ter, utilizamos métodos automáticos para encontrar aqueles mais relevantes, e averiguamos subconjuntos dessas características de escolas que definissem seu desempenho de alguma maneira verificável. A métrica de desempenho foi a nota média da prova objetiva do ENEM 2013 lograda por cada escola.

6.1. CONTRIBUIÇÕES

Este trabalho trouxe as seguintes contribuições para a literatura de mineração e descoberta de conhecimento em base de dados educacionais:

- Uma revisão da literatura brasileira sobre o tema. A tabela com os resultados está presente no link ***http://tiny.cc/DAERevisaoLiteratura***;
- Uma base de dados em formato ARFF que pode ser utilizado em outros experimentos. O arquivo está presente no link ***http://tiny.cc/DAE-ARFF***;
- O uso de diferentes classificadores em base de dados educacionais, permitindo que diferentes mecanismos avaliem os dados e gerem diferentes regras, expandindo as possibilidades de obter conhecimento;
- Extração de atributos sobre as escolas, descrevendo quais as características de uma escola, de acordo com as bases educacionais do INEP;

- Regras de decisão explícitas, que permitem uma interpretação direta e possibilita auxiliar diretamente gestores de escola e administradores públicos.

6.2. TRABALHOS SUBMETIDOS

No processo do desenvolvimento deste estudo, foram submetidos dois artigos para o XXVIII Simpósio Brasileiro de Informática na Educação (SBIE), no ano de 2017. O primeiro deles é uma revisão da literatura brasileira, e parte dele compõe o capítulo 3 deste trabalho. Este artigo, denominado “Dados Abertos Educacionais: Uma Revisão da Literatura Brasileira” foi escrito em conjunto com os professores Rafael Ferreira e Péricles Miranda.

O segundo artigo, dos mesmos autores, utiliza a mesma fonte de dados deste trabalho, porém uma base mais reduzida, com escolas do Estado de Pernambuco (o que reduz o total de instâncias), um número menor de algoritmos e menos regras geradas. A intenção também é procurar atributos estruturais e não-estruturais de escolas e classifica-los segundo a nota do ENEM.

6.3. TRABALHOS FUTUROS

Há uma grande abertura a ser preenchida com trabalhos futuros. Este trabalho teve como foco contribuir para gestores de escola, utilizando dados identificados como sendo notáveis a eles. Um trabalho futuro consistiria em reutilizar a base de dados com foco em pais e responsáveis, apontando diferenças na educação do ponto de vista externo à escola, sugerindo fatores que equilibrem custo e benefício ao escolher escolas para matricular estudantes. Por exemplo, o nível socioeconômico, o valor da mensalidade caso seja escola paga, a localidade, o tipo de administração (se privada ou pública, nas diferentes esferas) e outros fatores.

Os atributos socioeconômico e de tipo administrativo poderiam ser incluídos em um trabalho futuro. A proposta seria avaliar a utilização de políticas públicas, estabelecendo comparações que permitam verificar itens de sucesso em escolas particulares que possam ser utilizadas em escolas públicas. Deve-se, porém, atentar para o ajuste de parâmetros dos algoritmos, para que não gerem regras muito simples, como por exemplo: socioeconomico = Alto AND dependenciaAdministrativa = Privada AND alimentação = False => enemMediaObjetiva = 600-699. Em experimentos prévios a este trabalho, foram constatadas regras como esta, que podem mascarar a realidade das opções oferecidas por escolas

particulares. É importante que as regras geradas contenham mais atributos, descrevendo melhor as características das escolas para proporcionar uma interpretação mais precisa.

Outra possibilidade é ampliar a base de dados. Aqui utilizamos dados referentes à região Nordeste e apenas do ano de 2013. É possível obter diferentes resultados a partir da utilização de dados do Brasil inteiro, de dados com mais anos, ou mesmo ambos ao mesmo tempo. Há vários cenários viáveis, como estabelecer comparações entre os dados de diferentes anos, verificar a evolução das políticas públicas aplicadas na educação, verificar padrões ao longo dos anos em escolas que modificaram sua estrutura, observar similaridades e diferenças entre estados e regiões diferentes, ou até buscar variações de opções oferecidas ao longo de diferentes anos numa localidade definida, retornando à opção de servir de contribuição para pais e responsáveis.

Também é possível estender o trabalho atual ao modificar os parâmetros dos algoritmos do Weka e buscar uma maior taxa de acerto. Uma opção distinta é utilizar outro algoritmo classificador não existente no Weka e que, para essa base de dados, tenha altíssima taxa de acerto (acima de 95%) na classificação. Aprimorando a acurácia, seria possível criar um sistema com interface gráfica que permitisse a visualização de cada escola na árvore de decisão ou no conjunto de regras, descrevendo com exatidão as características que a fizeram obter aquela nota do ENEM. Dessa maneira, o gestor de escola poderia ver diretamente os atributos de sua escola e trabalhar para mitigar as deficiências, ou ainda mesmo a própria ferramenta poderia se basear em outras escolas registradas no sistema e sugerir melhorias ao gestor.

6.4.LIMITAÇÕES

Uma restrição importante neste estudo é devido ao ENEM ser utilizado como métrica do desempenho das escolas, o que limitou este trabalho a analisar somente escolas que disponibilizem ensino médio, mas não se limitando a este, ou seja, escolas que ofereçam o ensino fundamental e médio entraram na base, mas escolas que ofertam apenas o ensino fundamental, sem o médio, não entraram. Não obstante, mesmo após a retirada das escolas sem ensino médio, foram avaliadas 3504 escolas da região Nordeste para o período de 2013 obtido da base de dados.

Por fim, é importante dizer que este trabalho não teve a intenção de solucionar de vez a problemática da educação nacional. Embora tenhamos tomado dados abertos educacionais de fontes com informações do Brasil inteiro, não só o escopo do trabalho se restringiu à região

Nordeste como também acreditamos que a melhoria da Educação virá através de uma ação conjunta entre governo, educadores, instituições de ensino, pesquisadores e população. Precisamos ter a humildade de reconhecer que há muito a ser feito, e que este trabalho é apenas um esforço, ainda que relevante, em direção à uma educação brasileira de qualidade.

REFERÊNCIAS

ADEODATO, P. J. L.; SANTOS FILHO, M. M.; RODRIGUES, R. L. Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. In: Anais do XXV Simpósio Brasileiro de Informática na Educação. Dourados: XXV SBIE. 2014.

AGUIAR, M. O.; DO NASCIMENTO, E. L. Tecnologia a favor da Educação: Um Estudo de Caso das Escolas do Espírito Santo. In: XX WORKSHOP DE INFORMÁTICA NA ESCOLA, Dourados, 2014. Anais do XX WIE. Dourados: XX WIE. 2014.

ALCANTARA, W. E. A. Desafios no uso de dados abertos conectados na educação brasileira. In: WORKSHOP DE DESAFIOS DA COMPUTAÇÃO APLICADA À EDUCAÇÃO, 2015, Recife. Anais da XXXV CSBC. Recife: UFPE. 2015.

ALMUALLIM, H.; DIETTERICH, T. G. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, v. 69, n. 1-2, p. 279-305, 1994.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. Statistics surveys, v. 4, p. 40-79, 2010.

ATRICON. Cartilha do TCU aborda abertura de dados na administração pública. ASSOCIAÇÃO DOS TRABALHADORES DOS TRIBUNAIS DE CONTAS, 2015. Disponivel em: <http://www.atricon.org.br/imprensa/noticias/cartilha-do-tcu-abordaabertura-de-dados-na-administracao-publica/>. Acesso em: 19 Junho 2016.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca. Bookman Editora, 2013.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o Brasil. Brazilian Journal of Computers in Education, v. 19, n. 2, 2011.

BANDEIRA, J. et al. Dados abertos conectados para a Educação. Jornada de Atualização em Informática na Educação, Uberlândia, v. 4, n. 1, p. 47-69, 2015.

BAUER, F.; KALTENBÖCK, M. Linked Open Data: The Essentials. Viena: Edition mono/monochrom, 2012.

BRASIL. Portaria MEC nº 438, de 28 de maio de 1998. Institui o Exame Nacional do Ensino Médio. Brasília: MEC, 1998.

BRASIL. Parâmetros Curriculares Nacionais – Ensino Médio. Brasília: MEC, 2000.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5, no inciso II do § 3 do art. 37 e no § 2 do art. 216 da Constituição Federal; altera a Lei n 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.1. Brasília: Diário Oficial da União - Seção Extra, 2011. p. 1.

BREIMAN, L. et al. Classification and regression trees. CRC Press, 1984.

BUTTERWORTH, R. E. A. A greedy algorithm for supervised discretization. Journal of biomedical informatics, v. 37, n. 4, p. 285-292, 2004.

CARVALHO, M. J. S.; NEVES, B.; MELO, R. Plataforma CultivEduca. In: Anais dos Workshops do V Congresso Brasileiro de Informática na Educação. Uberlândia, 2016. p. 134.

CASTRO, C. M.; FLETCHER, P. A escola que os brasileiros frequentaram em 1985. Rio de Janeiro: Ipea, 1986.

COELHO NETO, J. et al. O uso das TIC na formação de professores de escolas que obtiveram baixo IDEB. In: Anais do XXII Simpósio Brasileiro de Informática na Educação. Aracaju, 2011. p. 988-996.

COHEN, W. W. Fast effective rule induction. In: Proceedings of the twelfth international conference on machine learning, 1995. p. 115-123.

COLPAERT, P. E. A. The 5 star of open data portals. International Conference on Methodologies, Technologies and Tools Enabling E-Government, Vigo, v. 7, 2013.

DA SILVA, C. F. E. A. Dados abertos: uma estratégia para o aumento da transparência e modernização da gestão pública. Revista do TCU, Brasília, v. 131, p. 22-29, 2014.

DE ASSIS RODRIGUES, F.; SANT'ANA, R. C. G.; FERNEDA, E. Análise do processo de recuperação de conjuntos de dados em repositórios governamentais. InCID: Revista de Ciência da Informação e Documentação, Ribeirão Preto, v. 6, n. 1, p. 38-56, 2015.

DE SOUSA, R.; DA SILVA, L. É. P. BRAVO Sistema Web de Apoio à Pesquisa em Educação. In: Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação. Maceió: 2015. p. 105.

DE SOUZA, M. N. V. Comparação de Algoritmos do Aprendizado de Máquina Aplicados na Mineração de Dados Educacionais. UFRPE. Recife. 2015.

DUCH, W. et al. Feature Ranking, Selection and Discretization. In: Proceedings of Int. Conf. on Artificial Neural Networks (ICANN). 2003. p. 251-254.

DUTRA, R. L. D. S.; TAROUCO, L. M. R. Recursos educacionais abertos (Open Educational Resources). Revista Novas Tecnologias na Educação, Porto Alegre, v. 5, n. 1, 2007.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. KDD, p. 82-88, 1996.

FERREIRA, G. S. 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: CBIE-LACLO 2015. Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação, Maceió, 2015. Maceió: IV CBIE. 2015.

FRANK, E.; WITTEN, I. H. Generating accurate rule sets without global optimization. 1998.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. AI magazine, v. 13, n. 3, p. 57, 1992.

FREITAG, D.; CARUANA, R. Greedy attribute selection. In: Machine Learning Proceedings 1994: Proceedings of the Eighth International Conference. Morgan Kaufmann, 2017. p. 28.

FRITZEN, E.; SIQUEIRA, S. W.; ANDRADE, L. C. Busca Contextualizada Enriquecida com Dados Abertos para Apoiar a Aprendizagem Colaborativa em Redes Sociais. Revista Brasileira de Informática na Educação, v. 21, n. 3, 2013.

GAINES, B. R.; COMPTON, P. Induction of ripple-down rules applied to modeling large databases. Journal of Intelligent Information Systems, v. 5, n. 3, p. 211-228, 1995.

GARNER, S. R. et al. Applying a machine learning workbench: Experience with agricultural databases. In: Proc Machine Learning in Practice Workshop. Anais do Machine Learning Conference. Tahoe City, 1995. p. 14-21.

GENEROSO, A. A. P. et al. Abordagem Qualitativa do uso das TDIC na Educação Básica. In: Anais do Workshop de Informática na Escola. Campinas, 2013.

GEVREY, M.; DIMOPOULOS, I.; LEK, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, v. 160, n. 3, p. 249-264, 2003.

GOLDBERG, D. E. Genetic algorithms in search, optimization, and machine learning. Reading: Addison-Wesley, 1989.

GOMES, T. Descoberta de Conhecimento Utilizando Mineração de Dados Educacionais Abertos. UFRPE. Recife. 2015.

GUERRA, P. C.; NAKAMURA, R. Y. M.; HRUSCHKA, E. R. Estimativa de demanda potencial de matrículas em ensino superior usando dados públicos e múltiplos modelos de regressão. In: Symposium on Knowledge Discovery, Mining and Learning, 2th. São Carlos: Sociedade Brasileira de Computação-SBC. 2014.

GUY, M. What is Open Education Data? European Public Sector Information Platform. Disponivel em: <http://www.epsiplatform.eu/content/what-open-education-data>. Acesso em: 01 Agosto 2015.

HALL, M. et al. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, Nova Iorque, 2009.

HAND, D. J.; MANNILA, H.; SMYTH, P. Principles of data mining. MIT press, 2001.

HOFMANN, M.; KLINKENBERG, R. RapidMiner: Data mining use cases and business analytics applications. Chapman and Hall/CRC Press, 2013.

ISOTANI, S.; BITTENCOURT, I. I. Dados Abertos Conectados. São Paulo: Novatec, 2015.

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: Proceedings of the ninth international workshop on Machine learning. Aberdeen, 1992. p. 249-256.

KOLLER, D.; SAHAMI, M. Toward optimal feature selection. In Proceedings of the Thirteenth International Conference on Machine Learning. Stanford InfoLab. 1996. p. 284-292.

KONONENKO, I. Estimating attributes: analysis and extensions of RELIEF. In: European conference on machine learning. Catania: Springer, 1994. p. 171-182.

LABORATÓRIO DE DADOS ABERTOS BRASIL. Sobre. Educação Inteligente, 2015. Disponivel em: <http://educacao.dadosabertosbr.com/sobre>. Acesso em: 22 Março 2017.

LEE, V. E.; FRANCO, C.; ALBERNAZ, A. Quality and equality in brazilian secondary schools: a multilevel cross-national school effects study. In: ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. San Diego, 2004.

LÖBLER, M. L.; LÖBLER, L. M. B.; NISHI, J. M. Os Laboratórios de Informática em Escolas Públicas e sua Relação com o Desempenho Escolar. Revista Novas Tecnologias na Educação, Porto Alegre, v. 10, n. 3, 2012.

PEÑA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, v. 41, n. 4, p. 1432-1462, 2014.

PINHEIRO, R. G. P.; ELIA, M.; SAMPAIO, F. F. Avaliando as competências escolares através da Prova Brasil usando ferramenta web. In: Anais do XIX Workshop de Informática na Escola. Campinas, 2013.

QUINLAN, R. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers, 1993.

RAMASWAMI, M.; BHASKARAN, R. A study on feature selection techniques in educational data minig. Journal of Computing, v. 1, n. 1, p. 7-11, Dezembro 2009.

RATH, S. et al. Knowledge discovery in databases. In: Database Security IX: Status and prospects. 2016. p. 317.

RIGO, S. J.; CAZELLA, S. C.; CAMBRUZZI, W. Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In: Anais do Workshop de Desafios da Computação Aplicada à Educação. Curitiba, 2012. p. 168-177.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. An adaptation of Relief for attribute estimation in regression. In: Machine Learning: Proceedings of the Fourteenth International Conference. São Francisco: Morgan Kaufmann, 1997. p. 296-304.

SANTOS, M. S. et al. Análise das Infraestruturas do Censo Escolar 2011: uma proposta da disciplina de Tópicos em Banco de Dados. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2014, Dourados. Anais da XXV SBIE. Dourados: XXV SBIE. 2014.

SHI, H. Best-first decision tree learning. Tese de Doutorado. University of Waikato. Waikato. 2007.

SOARES NETO, J. J. et al. Uma escala para medir a infraestrutura escolar. Estudos em Avaliação Educacional, São Paulo, v. 24, n. 54, p. 78-99, 2013.

TRIBUNAL DE CONTAS DA UNIÃO. 5 motivos para a abertura de dados na Administração Pública. TCU. Brasília. 2015. Disponível em: <https://portal.tcu.gov.br/biblioteca-digital/5-

motivos-para-abertura-de-dados-na-administracao-publica.htm> Acesso em 13 de Maio de 2017.

WEBER, S. The Sucess of Open Source. Cambridge: Cambridge University Press, v. 897, 2004.

WEISS, S. M.; KULIKOWSKI, C. A. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. São Francisco: Morgan Kaufmann, 1991.

WITTEN, I. H. et al. Data Mining: Practical machine learning tools and techniques. São Francisco: Morgan Kaufmann, 2016.