Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática

Descoberta de Conhecimento Utilizando Mineração de Dados Educacionais Abertos

Tancicleide Carina Simões Gomes

Recife

Dezembro de 2015

Tancicleide Carina Simões Gomes

Descoberta de Conhecimento Utilizando Mineração de Dados Educacionais Abertos

Orientadora: Maria da Conceição Moraes Batista
Coorientadora: Roberta Macêdo. M. Gouveia

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

Recife

Dezembro de 2015

Agradecimentos

Aos meus familiares pelo apoio e incentivo, torcedores afetuosos que vibram à cada conquista minha, por mais boba que ela seja. Sou grata em especial ao meu pai, pela torcida silenciosa, e à minha mãe, que embora não entenda a razão de tudo o que eu faço, ainda assim me sobrepuja de amor. Meu querido irmão Carlos Júnior, obrigada por chamar minha atenção e me distrair o tempo todo, me proporcionando eventuais retornos ao mundo real. Eu realmente amo vocês!

Aos meus amigos que conheci no âmbito acadêmico, há tantos a agradecer, mas gostaria de elencar algumas que impactaram minha vida de maneira muito única: Elizângela Lucena, Fabiana Almeida, Ingrid Costa, Jéssica Nunes, Mirela Souza, Mônica Padilha, Rotsen Albuquerque, vocês foram grandes companheiros ao longo do curso, foi maravilhoso compartilhar meus dias de aula, muitas vezes extremamente cansada, ao lado de vocês. Dyego Morais, meu querido irmão, foi deslumbrante compartilhar com você inquietações e angústias típicas de pecadores ansiosos por transformar o mundo ao seu redor, usando inclusive, computação. Agradeço ainda a Manoela Oliveira, Tatyane e Amanda Calixto, vocês foram incríveis e somaram de maneira fantástica à minha vida, enriquecendo-a com muitas trocas de figurinhas.

Queridos, mesmo os não citados aqui, vocês tornaram toda esta trajetória em um momento singular em minha vida, encheram-na de sorrisos e alegrias, tornaram-na mais suave e mais feliz. Obrigada.

Agradeço à minha primeira orientadora Silvana Bocanegra, por ser sempre tão fantástica. Nunca vou esquecer do seu *"Tanci, cê tá estudano?"*, obrigada por todo o apoio e por vibrar por mim em todos os momentos e em todas as conquistas, é muito bom poder compartilhá-las com você. Agradeço ainda ao professor Jones Albuquerque, pela constante consultoria gratuita (!) ao longo de toda a minha trajetória.

Agradeço à minha amiga e querida orientadora, professora Jeane Melo, por realizar um dos meus sonhos, pela paciência e pelo companheirismo incansável. Obrigada por fazer de mim a acadêmica que eu sou. Agradeço também à minha querida professora Taciana Pontual - nos faltou tempo para tantos planos e ideias - obrigada pelo carinho, pela amizade e pelas oportunidades maravilhosas que você me proporcionou.

Obrigada professor Alex Sandro Gomes, pelos vislumbres na área de IHC proporcionados pela curta, porém valiosa, experiência com o Redu. Agradecimentos especiais ao projeto DEMULTS, na pessoa da orientadora professora Flavia Peres, foram três anos de profundo aprendizado. Obrigada aos queridos amigos colaboradores

do projeto, em especial a Juanna Pessoa, Débora Capezzera, Glauci Oliveira e Rafael Santos e aos membros mais recentes que enriquecem minha vida.

Às professoras Conceição Moraes e Roberta Macêdo, que me orientaram ao longo deste trabalho de conclusão de curso que foi extenuante e desafiador para mim. Obrigada pela paciência e disposição em me orientar, mesmo diante de tantos percalços, obrigada por não desistirem de mim. Agradeço a professora Andreza Leite, pela paciência I-ME-RE-CI-DA. Um agradecimento especialíssimo à professora Pompéia Villachan, por me auxiliar a redirecionar meus esforços para outros rumos de maneira mais assertiva e a querida madrinha, professora Cláudia Araújo. Agradeço também aos colegas do CCTE que eu tive o prazer de conhecer e que eu importuno até os dias de hoje, em especial o querido Ivanildo Melo.

Agradeço ainda à minha amiga Énery Melo, obrigada pelo incentivo e por todas as risadas ao seu lado, por ser um exemplo inspirador enquanto pessoa e profissional e agradeço, inclusive, por me levar à Comissão Própria Avaliação (CPA). Agradeço à CPA, na pessoa das professoras Fátima Brandão e Giselle Nanes, por me permitir ampliar os horizontes do meu olhar e me permitir ver a universidade de maneira mais ampla. Meus agradecimentos se estendem aos amigos membros desta comissão, Edilene Vilaça, Fabiana Costa e Rodolpho Belarmino.

No âmbito profissional, agradeço a todas empresas pelas quais passei pela oportunidade de aprendizado e aos amigos que conquistei ao passar por cada uma delas. Agradeço aos amigos da *Joy Street*, recém chegados em minha vida e que me proporcionaram os meses mais alegres que eu já tive mesmo em meio a toda correria. Agradeço em especial a Marielle Azoubel, Gabriela Oliveira, Kristiana Oliveira, André, Marcone Marinho, Matheus Sales, Carla Alexandre, Vinícius Fabrino e carinhosamente agradeço à gerente mais sensível e amável que eu já tive, Andrea Pinto. Obrigada à direção da Escola Geração do Futuro, pela compreensão com minha rotina agitada e aos meus alunos, por tornarem minha vida mais leve e cheia de amor. Agradeço ainda a todos que fizeram parte de minha vida universitária contribuindo para que eu me tornasse a pessoa e profissional que sou hoje. Agradeço por todas as oportunidades que vocês me deram, projetos e experiências, pela conversa agradável, pelo cafezinho, pelo acolhimento.

Finalmente, agradeço à pessoa mais especial, o meu Deus. De fato Pai, o diploma foi apenas uma consequência, obrigada por nunca ter ido embora, mesmo eu não merecendo, você foi sempre presente. Obrigada por tudo o que me aconteceu ao longo desta jornada e que juntamente concorreu para o meu bem. À você, dedico minha vida e tudo o que eu sou.

Resumo

A adoção de técnicas de mineração de dados tornou-se uma realidade para varejistas, bancos, fabricantes, seguradoras, dentre outros, proporcionando-lhes a descoberta de relações entre padrões não visíveis em grandes volumes de dados. A mineração de dados tem ganho destaque também no cenário educacional, oferecendo aos educadores e gestores educacionais subsídios para a tomada de decisão, além de perspectivas para mitigar desafios tradicionais do processo de ensino-aprendizagem. Neste sentido, o presente trabalho versa sobre a aplicação de técnicas e métodos de mineração de dados a fim de descobrir padrões e regras de associação em dados estatísticos educacionais oriundos do Exame Nacional do Ensino Médio nos anos de 2013 e 2014 no âmbito da região Nordeste.

Palavras-chave: Mineração de dados educacionais, Regras de associação, Árvores de decisão, Dados abertos

Abstract

The adoption of data mining techniques has became a reality for retailers, banks, manufacturers, insurers, among others, providing them the discovery of relationships among invisible patterns in large databases. Data mining has also gained prominence in the educational setting, providing educators and education manager's insights for decision-making, as well as rich perspectives to mitigate traditional challenges in the teaching and learning process. Thus, the present work addresses the application of techniques and data mining methods in order to discover patterns and association rules in educational statistical data from the Brazilian National High School Exam in the years 2013 and 2014 in the Northeast extent.

Keywords: Educational data mining, Association rules, Decision tree, Open data

Lista de Figuras

Figura 1. Processo de KDD. 14
Figura 2. Ferramenta Weka Explorer para Windows. Fonte: Autora 21
Figura 14. Passos da execução do algoritmo Apriori 25
Figura 3. Script para leitura do arquivo com baixo consumo de memória. 35
Figura 4. Script para criação dos estratos 36
Figura 5. Script para cálculo do percentual e quantidade de registros para cada estrato 37
Figura 6. Procedimento de amostragem aleatória simples para a escolha dos registros de cada um dos estratos 37
Figura 7. Seleção de registros apenas da região Nordeste 38
Figura 8. *Script* para leitura do arquivo de parâmetros 40
Figura 9. Trecho de arquivo com a codificação dos atributos. 41
Figura 10. Trecho de arquivo com todas as capitais do NE 41
Figura 11. Script para leitura do arquivo de capitais e regiões metropolitanas do Nordeste 42
Figura 12.Trecho de arquivo gerado em formato *arff*. 42
Figura 13. Software Weka com os dados carregados 43
Figura 15. Trecho dos resultados da execução do Apriori com todos os atributos da base NE 45
Figura 16. Execução do *Apriori* com os dados do QSE na base de Pernambuco 47
Figura 17. Gráficos dos atributos `NOTA_MT`, `QSE_RENDA_FAMILIA` e `CAPITAL` em Pernambuco 50
Figura 18. `QSE_RENDA_FAMILIA` – NE (2014) e `QSE_RENDA_FAMILIA` – PE (2014) 51
Figura 19. Árvore de decisão correlacionando sexo e desempenho no exame 53

Lista de Quadros

Quadro 1. Categorias do atributo QSE_RENDA_FAMILIA 39

Sumário

_Toc439168023Introdução 10
1.1. Objetivos 12
1.2. Organização do Trabalho 12
Fundamentação Teórica 13
2.1. Mineração de Dados 13
2.2. Dados Abertos 18
2.3. Ferramentas Computacionais de Mineração de Dados 20
2.4. Algoritmos de Aprendizagem de Máquina 22
2.4.1. Algoritmo Apriori 23
2.4.2. Algoritmo J48 26
2.5. Trabalhos Relacionados 27
Metodologia 32
3.1. Pré-Processamento dos Dados 32
3.1.1. Seleção dos dados: Os microdados do ENEM 32
3.1.2. Limpeza dos dados 33
3.1.3. Transformação dos dados 39
3.2. Mineração dos Dados 43
3.2.1. Seleção dos Algoritmos 44
Resultados e Discussões 45
4.1. Análise de desempenho com o algoritmo *Apriori* 45
4.2. Análise de desempenho utilizando o algoritmo *J48* 52
4.3. Análise de redações com notas zero 54
Considerações Finais 56
Referências 59
Apêndice A – Atributos selecionados e respectivos valores nominais 64
Anexo A – Recorte do Dicionário de Dados do ENEM 2013 66

Capítulo 1

Introdução

Atualmente, o volume de dados produzido pelos mais diversos tipos de contextos, em especial com advento da web 2.0 – *e-commerce*, redes sociais, ambientes virtuais de aprendizagem, dentre outros - dobra a cada dois anos, de tal forma que dados não estruturados compõem, por si só, um total de aproximadamente 90% do universo digital; no entanto, mais dados não necessariamente indicam mais informação, mais conhecimento. Neste direcionamento, desponta a mineração de dados (MD), que propõe a filtragem destes dados de maneira a identificar correlações nos dados, bem como o que é relevante, a fim de avaliar os possíveis resultados (LUAN, 2007; USDE, 2012).

As ferramentas e técnicas de mineração e análise de dados, anteriormente apenas confinados a sofisticados laboratórios de investigação, agora têm sido crescentemente adotados por: varejistas, bancos, fabricantes, operadoras de telecomunicações, seguradoras, os quais têm descoberto o potencial de descobrir relações entre padrões não visíveis em vastos bancos de dados e até em redes sociais (LUAN, 2007; WEST, 2012).

Nos últimos anos tem crescido também a adoção do processo de descoberta do conhecimento em contextos educacionais, dentre o qual a mineração faz parte e assim surge a denominada mineração de dados educacionais (MDE), em que diversas instituições de ensino estão começando a utilizar a análise de dados para auxiliar na tomada de decisão (USDE, 2012). Estes padrões são então construídos em modelos de mineração de dados e usados para prever comportamentos individuais com alta precisão e como resultado direto disto as instituições podem alocar recursos e pessoal de forma mais eficaz.

A MDE desenvolve métodos e aplica técnicas de estatística, aprendizado de máquina e banco de dados para analisar dados coletados durante o processo de ensino-aprendizagem, congregando métodos de computação para compreender como os estudantes aprendem (USDE, 2012).

Considerando, por exemplo, o cenário do ensino superior, um dos maiores desafios consiste em prever as trajetórias dos alunos. Instituições de ensino gostariam de saber, por exemplo: *(i) Quais estudantes precisarão de assistência para se formar? (ii) Quais estudantes são mais propensos a evadir que os outros? (iii) Quais alunos estão cumprindo o maior número de créditos? (iv) Quem são os alunos mais persistentes da universidade? (v) Quais tipos de cursos atraem mais estudantes? (vi) Quais os perfis de estudantes propensos a reprovações/retenções?*

Além disso, as questões tradicionais como gerenciamento de matrículas e o tempo para graduar-se são propulsores para a busca por melhores soluções (LUAN, 2007; USDE, 2012; WEST, 2012).

Novos métodos e ferramentas de aprendizagem apoiados por computador, sistemas tutores inteligentes, simulações, jogos, abriram oportunidades para a coleta e análise de dados dos estudantes para descobrir padrões e tendências a partir destes dados, permitindo novas descobertas e o teste de hipóteses sobre como os alunos aprendem.

Os dados obtidos a partir de sistemas como estes podem ser agregados de uma grande quantidade de alunos e podem conter muitas variáveis que algoritmos de mineração podem explorar na construção de novos modelos. Por exemplo, com olhares individualizados e voltados para os dados de progresso de cada aluno, os educadores poderiam prever o desempenho dos alunos e desenvolver estratégias para mantê-los no caminho adequado (LUAN, 2007). Ou seja, a MDE tem o potencial de tornar visíveis dados até então ocultos e despercebidos.

O presente trabalho versa sobre a aplicação de métodos e técnicas de mineração de dados em dados estatísticos educacionais provenientes dos anos de 2014 e 2013 do

Exame Nacional do Ensino Médio (ENEM) utilizando algoritmos de regras de associação e de classificação.

1.1. Objetivos

O presente trabalho visa contribuir, através da descoberta de conhecimento, oferecendo subsídios para a criação e/ou fortalecimento de iniciativas que visem melhorias na diminuição da evasão e retenção dos estudantes universitárias, avaliação e diagnóstico do corpo discente.

Além disso, pretende-se auxiliar, por exemplo, os gestores públicos na criação/consolidação de ações afirmativas dentro da universidade, que vão desde o estabelecimento de cotas, programas de nivelamento a programas de bolsas de estudo e de apoio acadêmico.

Os objetivos específicos deste trabalho são:

- Descobrir padrões ocultos e identificar regras de associação;
- Implementação de *scripts* para as fases de limpeza, transformação e codificação dos dados;
- Utilização de algoritmos de associação e classificação *J48* e *Apriori*.

1.2. Organização do Trabalho

Este trabalho está organizado da seguinte forma:

- O capítulo 2 apresenta um conjunto de definições relevantes para o entendimento deste trabalho, assim como os principais trabalhos relacionados.
- O capítulo 3 apresenta os métodos e técnicas utilizados para o desenvolvimento deste trabalho.
- O capítulo 4 apresenta os resultados e discussões.
- O capítulo 5 apresenta as considerações finais.

Capítulo 2

Fundamentação Teórica

2.1. Mineração de Dados

A mineração de dados é um campo de pesquisa interdisciplinar, tendo como os principais e mais expressivos os esforços em pesquisa em três áreas: *(i) Estatística, (ii) Aprendizado de Máquina, (iii) Banco de Dados.* Deste modo as definições de MD perpassam por algumas variações conforme o campo de atuação dos autores (RODRIGUES et al., 2014). A partir destas perspectivas são apresentadas três definições consideradas relevantes e, de certo modo, complementares, para o conceito de mineração de dados:

Hand et al. (2001 apud CAMILO; SILVA, 2009, p.9), considerando um viés estatístico apresenta a seguinte definição:

> *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.*

Cabena *et al.* (1998 apud CAMILO; SILVA 2009, p.9), por sua vez, definem mineração de dados como um campo interdisciplinar que agrupa técnicas de máquinas de conhecimento, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados.

A definição de Fayadd *et al.* (1996 apud CAMILO; SILVA, 2009, p.9) é ligeiramente distinta:

> *Mineração de dados é um passo no processo de descoberta de conhecimento que consiste na realização da análise dos dados e*

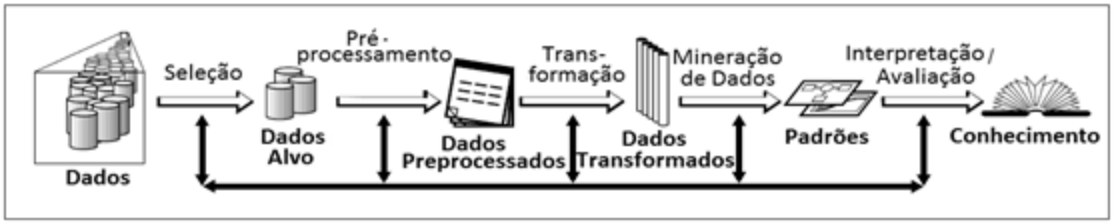
aplicação de algoritmos de descoberta que, sob certas condições, produzem um conjunto de padrões de certos dados

Neste trabalho consideraremos MD como o processo de descoberta automatizado de informações novas em um contexto previamente definido a partir de grandes volumes de dados, objetivando o apoio à tomada de decisão: um passo essencial no processo de descoberta de conhecimentos envolvendo o uso de variadas técnicas, tarefas e algoritmos (RODRIGUES et al., 2014).

A descoberta de padrões é um processo que se inicia pela escolha dos dados que documentam de alguma maneira as questões que os especialistas desejam responder. Os dados são integrados e pré-processados para que sejam entregues estruturados, limpos, selecionados e padronizados à tarefa de mineração de dados. Na tarefa de mineração aplica-se alguma técnica inteligente que possibilite o encontro de soluções auxiliem os especialistas na descoberta de respostas. Os resultados desta tarefa devem ser pós-processados para se apresentem análises qualitativa/ quantitativa dos elementos encontrados e, quando possível, apresentados de maneira que possa ser interpretada de modo a auxiliar na tomada de decisão (RODRIGUES et al., 2014).

Este processo é denominado Descoberta de Conhecimento em Bases de Dados (do inglês *Knowledge Discovery in Databases* - KDD), que referenciaremos apenas por KDD ao longo deste trabalho. O KDD abrange diversas fases que são potencialmente o caminho que os dados percorrem até tornaram-se conhecimento que seja útil. Ou seja, consiste em um processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente uteis e interpretáveis (FAYADD, 1996 apud CAMILO; SILVA, 2009), conforme explicita a Figura 1.

Figura 1. Processo de KDD.



Em um primeiro momento, é necessário definir qual o tipo de conhecimento a descobrir, pressupondo-se uma ampla compreensão do domínio da aplicação assim com do tipo de decisão que tal conhecimento pode contribuir para melhorar. A etapa Seleção (Figura 1) consiste em selecionar um conjunto de dados ou mesmo focar em um subconjunto onde a limpeza deve ser realizada.

A etapa seguinte é o Pré-Processamento, que tem como funções básicas a remoção de ruídos, seleção de atributos relevantes, formatação dos dados, tratamento de campos ausentes, dentre outros. A Transformação é a etapa seguinte em que ações como a redução da dimensionalidade dos dados ou mesmo o enriquecimento dos dados.

A Mineração de Dados é a fase logo a seguir, em que os dados já foram transformados para identificar padrões através de algoritmos de aprendizagem. Posteriormente, há a interpretação dos padrões encontrados, em que comumente se revisitam as fases iniciais do processo. A fase final consiste na incorporação do conhecimento ou na sua documentação e *report* para os responsáveis.

Os novos padrões (informações) descobertos necessitam possuir algum grau de certeza para que sejam consideradas válidas e precisam ser descritas em alguma linguagem que possa ser facilmente compreendida pelos usuários finais, a fim de que eles possam realizar uma análise mais profunda (WEST, 2012).

No contexto educacional, a MD assume uma nova roupagem e seus métodos são adaptados, surgindo uma nova área de pesquisa denominada *Mineração de Dados Educacionais* (MDE), que tem como foco o desenvolvimento de métodos para a exploração de conjuntos de dados advindos de contextos educacionais (BAKER, 2011). Deste modo, tem-se subsídios para uma melhor compreensão de diversos aspectos, tais como: *(i)Como os alunos aprendem? (ii)Qual o papel do contexto da aprendizagem?; (iii)Quais fatores influenciam a aprendizagem?*, dentre outras questões, que permitem ensaiar estratégias e métodos que promovam melhorias no processo de ensino-aprendizagem.

Assim é possível encontrar indicadores sobre a motivação do aluno, oferecer personalização do ambiente e dos métodos de ensino proporcionados no intuito de ofertar melhores condições de aprendizagem.

Os estudiosos da área têm definido alguns tópicos como sendo os principais objetivos de pesquisa (BAKER, 2011; BAKER; YACEF, 2009 *apud* USDE, 2012):

- A predição do possível comportamento da aprendizagem através da criação de modelos que incorporam informações tão detalhadas como: o conhecimento dos alunos, a motivação, a metacognição e as atitudes;
- A descoberta ou melhoria de modelos de domínio que caracterizam o conteúdo a ser aprendido e as sequências instrucionais;
- O estudo dos efeitos de diferentes tipos de suportes pedagógicos que podem ser fornecidos por softwares de aprendizagem; e
- O conhecimento científico avançado sobre aprendizagem e aprendizes através da construção de modelos computacionais que incorporem modelos de estudantes, do domínio e da pedagogia do software.

Segundo Baker (2011) apud USDE (2012), a fim de atingir estes objetivos, a pesquisa em mineração de dados educacionais contempla o uso de cinco categorias de métodos: predição, *clustering*, *relationship mining*, *distillation for human judgment* e descoberta de modelos.

A predição pressupõe o desenvolvimento de um modelo que infere um aspecto singular dos dados a partir de alguma combinação de outros aspectos. Exemplos de uso da predição incluem a detecção do comportamento dos estudantes, assim como quando eles estão jogando no sistema ou falhando ao responder uma questão corretamente apesar de possuírem habilidade para tal. Modelos preditivos têm sido usados, por exemplo, para a compreensão de quais comportamentos em um ambiente online podem indicar quais estudantes de uma turma podem falhar.

A *clustering* refere-se a encontrar marcadores (*datapoints*) que naturalmente se agrupam e que podem ser usados para dividir um conjunto mais amplo em categorias.

Exemplos de aplicação de *clustering* são o agrupamento de estudantes baseados em suas dificuldades de aprendizagem e padrões de interação, assim como o quanto e como eles usam ferramentas em um sistema de gestão de aprendizagem. Podem ser analisados entrevistas cognitivas dos alunos, postagens em fóruns de discussão e dados tão variados quanto, usando técnicas para trabalhar com dados não estruturados, e, em seguida agrupando os resultados. A *clustering* pode ser usado em qualquer domínio que envolve a classificação, até mesmo para determinar o quanto os usuários colaboram com base nas postagens em fóruns de discussão.

As regras de associação envolvem a descoberta entre variáveis em conjunto de dados e a codificação deles com regras para uso posterior. Por exemplo, as regras de associação podem identificar as relações entre os produtos adquiridos em compras on-line (ROMERO; VENTURA 2010). No cenário educacional, as regras de associação são muito relevantes.

Elas podem ser utilizadas para encontrar erros de estudantes que ocorrem concomitantemente, associando conteúdos com os tipos de usuários para construir recomendações para conteúdos que sejam suscetíveis a serem interessantes ou ainda para realizarem alterações nas abordagens de ensino. Tais técnicas podem ser usadas para associar a nota de um aluno, em um sistema de gerenciamento da aprendizagem, com atividades e/ou fóruns de discussão, como por que o uso de testes práticos pelos alunos diminui ao longo de um semestre de estudos.

A *sequential pattern mining* constrói regras que capturam as conexões entre as ocorrências de eventos sequenciais, encontrando sequências temporais, como erros de estudantes seguidos pela busca de ajuda. As aplicações educacionais de regras de associação incluem a descoberta de associações entre a performance dos estudantes e as sequências de cursos, e, descobrir quais estratégias pedagógicas levam a uma aprendizagem mais robusta e eficaz.

A *distillation for human judgment* é uma técnica que envolve descrever dados de uma forma que permite que um ser humano possa identificar ou classificar

rapidamente características dos dados. Esta área de MDE melhora os modelos de aprendizado de máquina, porque os seres humanos podem identificar padrões (ou características de aprendizagem) em ações, comportamentos do estudante ou dados que envolvam colaboração dos alunos.

Ao utilizar estas técnicas, os pesquisadores podem construir modelos para responder questões como:

- Qual a sequência de tópicos mais eficaz para um estudante em específico?
- Quais as ações do estudante estão associadas com mais aprendizagem?
- Quais ações do estudante indicam satisfação, envolvimento, progresso na aprendizagem, etc.?
- O que prediz o sucesso do estudante?

O delineamento de respostas para estas perguntas perpassa também por percorrer vastas e ricas bases de dados. Comumente, as principais bases de dados utilizadas para a mineração de dados educacionais abrangem desde sistemas tutores inteligentes, simulações, jogos, sistemas virtuais de aprendizagem até sistemas de gerenciamento acadêmico.

No entanto, embora tenham recebido menos destaque que as outras bases mais comuns, os dados abertos (o que inclui dados estatísticos, como censos, por exemplo) possuem uma ampla gama de informações relevantes para a tomada de decisão e para a criação de políticas públicas. A seção a seguir apresenta uma breve visão sobre o que são dados abertos e como eles podem ser ricas fontes para a aplicação de mineração e análise de dados.

2.2. Dados Abertos

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]. São considerados dados abertos os dados provenientes de fontes diversas, tais como[2]: *(i) Cultura, (ii) Ciência, (iii) Finanças, (iv) Estatísticas, (v)Tempo, (vi) Ambiente, (vii) Transporte*. O conceito de dados abertos é proveniente da possibilidade de promover transparência, participação da sociedade e engajamento, bem como agregar valor social e comercial, de modo que os dados possam ser livremente utilizados, modificados e compartilhados por quaisquer pessoas para qualquer finalidade[3].

Neste sentido, anualmente, é publicado o Índice Global de Dados Abertos, que oferece uma avaliação independente da abertura dos dados de diversos lugares (não necessariamente países) nas seguintes áreas: horários dos meios de transporte, orçamento governamental, gastos governamentais, resultados eleitorais, registros de empresas, mapas nacionais, estatísticas nacionais, legislação, códigos postais e até mesmo emissão de poluentes.

Na última edição publicada, em 2014, o Brasil alcançou o 26º lugar, tendo sofrido uma queda, já que em 2013 alcançou a 24º posição, mesmo tendo avançado de 48% para 54% de dados abertos. Aparentemente, embora tenha havido um aumento no apoio a implantação da política de dados abertos nos governos, o progresso, conforme apontam os índices, parece caminhar mais lento do que a retórica[4], pois a disponibilização de dados abertos perpassa não apenas pela simples disponibilização dos dados, mas estes dados precisam disponibilizados em formatos acessíveis também.

Tim Berners-Lee[5] propôs um esquema de cinco estrelas para a publicação de dados abertos, de modo que os dados tornem-se progressivamente, à medida que avançam na escala, mais poderosos e de uso mais fácil, estes níveis são descritos a seguir:

1. Disponível na web em qualquer formato (*pdf,* imagens) sob licença aberta;

[1] https://okfn.org/
[2] https://okfn.org/opendata/
[3] http://opendefinition.org/
[4] http://index.okfn.org/dataset/
[5] http://www.w3.org/DesignIssues/LinkedData.html

2. Dados estruturados de maneira legível por uma máquina;
3. Formato não proprietário (*csv*)
4. Padrões W3C (RDF e SPARQL)
5. Dados conectados com dados de outras fontes para fornecer contexto

Se considerada a classificação proposta por Berners-Lee, é possível afirmar que os dados abertos utilizados neste trabalho (ENEM 2013-2014), se encontram na classificação 3 estrelas: disponíveis na web, sob licença aberta, estruturados de maneira legível para máquinas e utilizando um formato não proprietário (*csv*). Ou seja, ainda não é possível integrar as informações obtidas com outros dados estatísticos educacionais tais como: Prova Brasil, Censo Escolar e Censo da Educação Superior.

Atualmente, os dados abertos fornecidos pelo governo federal brasileiro estão disponíveis, primordialmente no Portal de Dados Abertos Brasileiros[6], embora alguns estejam desatualizados se comparados às suas fontes originais. Por exemplo, o site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) já disponibiliza os dados do ENEM de 2014, enquanto o Portal apresenta os dados apenas até o ano de 2012.

2.3. Ferramentas Computacionais de Mineração de Dados

Atualmente, existem diversas ferramentas disponíveis para auxiliar no processo de mineração de dados, algumas, inclusive algumas voltadas a usuários finais. Umas foram desenvolvidas visando a utilização com bancos de dados específicos, tais como:

- *IBM Intelligent Miner*[7] (uma ferramenta de mineração para bancos de dados DB2),
- *Oracle Data Mining*, o *SQL Server Analysis Services* (suplemento de recursos para apoiar a mineração de dados em bancos de dados SQL Server).

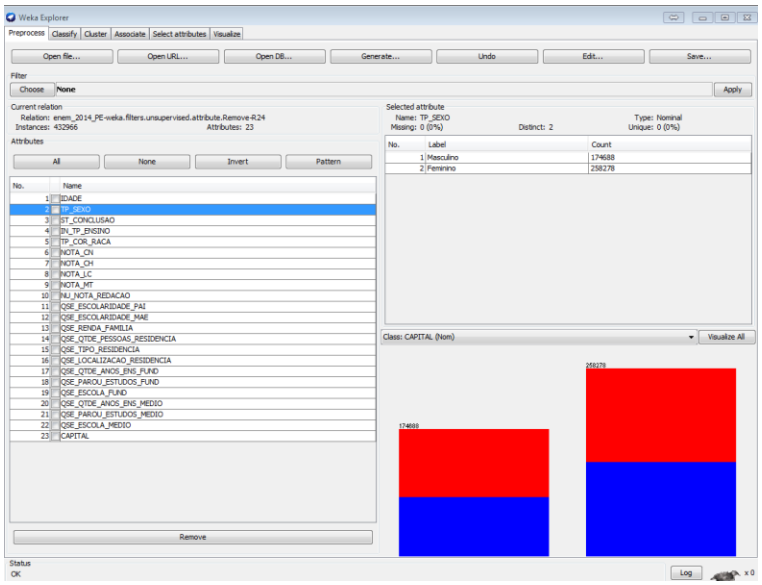
[6] http://dados.gov.br/dataset
[7] http://www.ibm.com/developerworks/data/library/tutorials/iminer/iminer.html

Outras ferramentas possuem interface gráfica do usuário, dentre elas podemos citar: a *Rapid Miner, SAS Enterprise Miner* e a *Weka*. A *RapidMiner*[8] é uma plataforma de software que fornece um ambiente integrado para aprendizagem de máquina, mineração de dados, mineração de textos, análise preditiva e de negócios. A *RapidMiner* suporta o desenvolvimento da aplicação e suporta todos os passos do processo de mineração de dados incluindo resultados de visualização, validação e otimização

A *Statiscal Analysis System (SAS) Enterprise Miner*[9] é uma suíte de software que permite análises avançadas: análise multivariada, inteligência de negócio, gerenciamento de dados e análise preditiva. Ele permite minerar, alterar, gerenciar e recuperar dados de uma variedade de fontes.

A *Weka*[10] consiste em uma coleção de algoritmos de aprendizagem de máquina destinado a possibilitar a realização de tarefas de mineração de dados. Estes algoritmos podem ser aplicados diretamente ao conjunto de dados ou serem *chamados* via código Java.

Figura 2. Ferramenta *Weka Explorer* para *Windows*. Fonte: Autora



[8] https://rapidminer.com/
[9] http://www.sas.com/en_us/software/analytics.html
[10] http://www.cs.waikato.ac.nz/ml/weka/

A *Weka* congrega ferramentas para pré-processamento dos dados, classificação, regressão, *clusterização*, regras de associação e visualização, sendo adequado ainda para o desenvolvimento de novos esquemas de aprendizagem de máquina.

2.4. Algoritmos de Aprendizagem de Máquina

O aprendizado de máquina é um subcampo da Inteligência Artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitem ao computador aprender, ou seja, permitem ao computador aperfeiçoarem o seu desempenho em alguma tarefa específica.

Algumas das principais técnicas de Inteligência Artificial possuem três características principais: *(i)* busca (a fim de explorar as distintas possibilidades em problemas em que os passos não são claramente definidos); *(ii)* emprego do conhecimento (permite explorar a estrutura, relações do mundo ou domínio à que pertence o problema bem como a redução de possibilidades a considerar, tal qual os humanos); *(iii)* abstração (proporciona a maneira de generalizar nos passos intrinsecamente similares).

Embora o aprendizado de máquina seja uma ferramenta poderosa para a aquisição automática de conhecimento, deve ser observado que não existe um único algoritmo que apresente o melhor desempenho para todos os problemas (MONARD; BARANAUSKAS, 2005).

No entanto, é relevante salientar que não existe um algoritmo único que apresente o melhor desempenho na resolução de todos os problemas. Deste modo, se faz necessário compreender as limitações dos diversos algoritmos de aprendizado de máquina utilizando alguma metodologia que permita avaliar os conceitos induzidos por estes algoritmos em determinados problemas (MONARD; BARANAUSKAS, 2005).

De acordo com os padrões a serem aprendidos e a disponibilidade de dados para treinamento, pode-se separar em dois tipos de aprendizado, os quais são conhecidos como paradigmas de aprendizado de máquina: aprendizado supervisionado e não

supervisionado. Um dos principais algoritmos de aprendizagem de máquina não supervisionado é o *Apriori* e um dos mais conhecidos algoritmos de aprendizagem de máquina supervisionado é o *J48*, amplamente utilizados em estudos envolvendo descoberta de conhecimento, ambos são apresentados a seguir (MONARD; BARANAUSKAS, 2005).

2.4.1. Algoritmo Apriori

As regras de associação permitem encontrar regras para descrever a ocorrência de um item baseado na ocorrência de outros itens em uma mesma transação. De modo que é possível encontrar todos os conjuntos de itens que frequentemente ocorrem de forma conjugada na base de dados e formar regras a partir destes conjuntos. Formalmente, é possível descrever o problema de mineração envolvendo regras de associação é da seguinte maneira (AGRAWAL, 1996) *apud* (GOUVEIA, 2009): Considere $I = \{i_1, i_2, i_3, ..., i_n\}$ um conjunto de n itens distintos e D uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens, chamado *itemset*, tal que $T \subseteq I$.

Uma regra de associação é uma expressão na forma $X \Rightarrow Y$, onde $X \subseteq I$, $Y \subseteq I$, $X \neq \emptyset$, $Y \neq \emptyset$, $X \cap Y \neq \emptyset$. X é denominado antecedente e Y denominado consequente da regra. Cada uma das regras geradas possui dois atributos que determinam sua validade no conjunto dos dados: suporte e confiança, os quais limitam a quantidade de regras a serem extraídas, garantindo ainda o descarte das regras julgadas de pouco interesse, uma vez que são menos frequentes e por conseguinte menos confiáveis.

A função do suporte é determinar a frequência com que um determinado *itemset* (conjunto de itens) ocorre em todas as transações da base de dados, logo um *itemset* é considerado frequente se o seu suporte for maior ou igual a um suporte mínimo estabelecido previamente, já o parâmetro confiança refere-se o percentual de ocorrência da regra.

A regra $X \Rightarrow Y$ é válida no conjunto de transações D com grau de confiança "c", se c% das transações em D que contêm X também contêm Y. E a regra $X \Rightarrow Y$ tem suporte "s" em D, se s% das transações em D contêm $X \cup Y$.

Ou seja, os parâmetros confiança e suporte são essenciais para o funcionamento do algoritmo. Eles determinam diretamente tanto a quantidade como a qualidade das regras trabalhadas, assim a utilização correta destes parâmetros é fundamental para a geração de regras de associação significativas para a análise.

Esse algoritmo emprega busca em profundidade e utiliza os conjuntos de itens de tamanho k para gerar os conjuntos de itens de tamanho $(k + 1)$. O primeiro passo do algoritmo é encontrar os conjuntos de itens frequentes com 1 item. Este conjunto é denominado de L_1. O conjunto de L_1 é usado para gerar L_2, que representa os conjuntos de itens frequentes com 2 itens, e assim por diante, até que nenhum conjunto de itens frequentes possa ser gerado.

O algoritmo *Apriori* usa o princípio de que cada subconjunto de um conjunto de itens frequentes também deve ser frequente. Esta regra é utilizada para reduzir o número de candidatos a serem comparados com cada transação no banco de dados. Todos os candidatos gerados que contém algum subconjunto que não seja frequente são eliminados, ou utilizando a terminologia do algoritmo, podados.

Cada passo inicia com um conjunto semente de itens, e esse conjunto semente gerará novos conjuntos potenciais, chamados conjunto de itens candidatos. Enquanto o conjunto de itens candidatos não ficar vazio, o algoritmo armazena esses conjuntos e para cada tupla do banco de dados testa se um conjunto candidato está ou não contido na tupla.

Caso um conjunto candidato esteja contido na tupla, então incrementa um contador. Se ao final do teste para cada tupla da base de dados uma regra candidata tiver um suporte mínimo especificado, então ela é inserida no novo conjunto semente, que são os itens candidatos.

Suponha um banco de dados cujo conjunto de itens I = {a, b, c, d, e} e um conjunto de transações T = {1, 2, 3, 4, 5, 6}, conforme mostra. O objetivo do algoritmo *Apriori* é determinar os *itemsets* com *MinSup* igual a 50%, ou seja, que ocorram em pelo menos três transações, haja vista que 50% de 6 transações corresponde a 3 transações.

Tabela 1. Exemplo de uso do algoritmo Apriori

Itens da Base de Dados						
T	1	2	3	4	5	6
Itens	abde	bce	abde	abce	abcde	bcd

O algoritmo prossegue sua execução, conforme mostra a figura abaixo.

Figura 3. Passos da execução do algoritmo Apriori

$C_1 = L_1$

Itemset	Suporte	
a	67%	4/6
b	100%	6/6
c	67%	4/6
d	67%	4/6
e	83%	5/6

C_2

ab	67%	4/6
ac	33%	2/6
ad	50%	3/6
ae	67%	4/6
bc	67%	4/6
bd	67%	4/6
be	83%	5/6
cd	33%	2/6
ce	50%	3/6
de	50%	3/6

L_2

ab	67%	4/6
ad	50%	3/6
ae	67%	4/6
bc	67%	4/6
bd	67%	4/6
be	83%	5/6
ce	50%	3/6
de	50%	3/6

C_3

abc	17%	1/6
abd	50%	3/6
abe	67%	4/6
acb	33%	2/6
acd	17%	1/6
ace	33%	2/6
ade	67%	4/6
bce	50%	3/6
bde	50%	3/6
cde	17%	1/6

L_3

abd	50%	3/6
abe	67%	4/6
ade	50%	3/6
bce	50%	3/6
bde	50%	3/6

C_4

abcd	17%	1/6
abce	33%	2/6
acde	17%	1/6
abde	50%	3/6
bcde	17%	1/6

L_4

abde	50%	3/6

RESULTADO – Algoritmo Apriori

Suporte		itemset
100%	6/6	b
83%	5/6	e, be
67%	4/6	a, c, d, ab, ae, bc, bd, abe.
50%	3/6	ad, ce, de, abd, ade, bce, bde, abde.

O algoritmo gera os conjuntos candidatos (C_1, C_2, C_3 e C_4) e a partir destes descobrindo os *itemsets* frequentes (L_1, L_2, L_3 e L_4) com suporte mínimo de 50%, como

mostra a tabela RESULTADO. O conjunto candidato (Cn) é formado por todas as combinações do *itemset*. O conjunto do "*itemset freqüente*" (L_n) é formado pelos valores do Cn que possuem suporte mínimo.

2.4.2. Algoritmo J48

O algoritmo J48 permite a criação de modelos de decisão em árvore. Utiliza uma tecnologia *greedy* para induzir árvores de decisão para posterior classificação. O modelo de árvore de decisão é construído pela análise dos dados de treino e o modelo utilizado para classificar dados ainda não classificados. O J48 gera árvores de decisão, em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão são construídas do topo para a base, através da escolha do atributo mais apropriado para cada situação. Uma vez escolhido o atributo, os dados de treino são divididos em sub-grupos, correspondendo aos diferentes valores dos atributos e o processo é repetido para cada sub-grupo até que uma grande parte dos atributos em cada sub-grupo pertençam a uma única classe. A indução por árvore de decisão é um algoritmo que habitualmente aprende um conjunto de regras com elevada acuidade. Este algoritmo é escolhido para comparar a percentagem de acerto com outros algoritmos.

Librelotto e Mozzaquatro (2014) afirmam que classificação consiste no processo de encontrar um conjunto de modelos que descrevem e distinguem classes, com o objetivo de utilizar o modelo final (refinado) para predizer a classe de objetos que ainda não foram classificados, tendo como um de seus algoritmos mais conhecidos o *J48* que gera árvores de decisão. A construção do modelo é baseada na análise prévia de um conjunto de dados de amostragem ou dados de treinamento, contendo objetos corretamente classificados.

O J48 é um indutor *top-down* de árvores de classificação. A seleção da melhor partição dos nós e o critério de parada são baseados na entropia de Shannon, como é usual em parte da família de indução de árvores de classificação. O J48 também possui

uma fase de pós-poda da árvore após a expansão, na qual são convertidas para folhas as sub-árvores que não representam ganhos de informação acima de um limiar especificado. O algoritmo é capaz de lidar com classes binárias, nominais e valores faltantes de classe. Suporta atributos binários, de data, nominais, numéricos e valores faltantes (BRILHADORI; LAURETTO, 2013).

As árvores de decisão utilizam a ideia de segmentar recursivamente o conjunto até encontrar uma partição que represente casos pertencentes à mesma classe. O algoritmo *J48* tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, de maneira que este modelo é usado para classificar as instâncias no conjunto de teste, permitindo o uso de variáveis qualitativas contínuas e discretas.

Ao montar a árvore de decisão, este algoritmo usa a abordagem de dividir para conquistar, em que um problema complexo é decomposto em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles determinada classe.

2.5. Trabalhos Relacionados

A mineração de dados educacionais (MDE) têm despontado nos últimos anos enquanto uma relevante área de pesquisa que permite a análise de grandes volumes de dados, proporcionando, inclusive novos vislumbres e perspectivas para a resolução de problemas no âmbito educacional.

Os esforços em pesquisa no panorama nacional envolvendo esta temática têm sido crescentes ao longo dos últimos anos, que evoluíram de um tímido cenário com menos de 20 trabalhos relatados nos principais periódicos e conferências de 2006 a 2010

(RODRIGUES *et al.* 2014), para mais de 10 trabalhos publicados apenas nos anais do Congresso Brasileiro de Informática na Educação (CBIE) em 2015[11].

Através de uma recente revisão da literatura, considerando 68 artigos envolvendo MDE publicados nas principais conferências e periódicos do país, Rodrigues *et al.* (2014) relatam que os trabalhos concentram-se nas seguintes temáticas: *(i) mineração de dados, (ii) mineração de textos e (iii) visualização de dados.*

Ainda segundo Rodrigues *et al.* (2014), os principais objetivos educacionais abordados são: 1. Estimativa ou modelagem de desempenho de estudantes (12 trabalhos publicados), 2. Modelagem de grupos ou aprendizagem colaborativa (10 artigos), 3. Mediação ou recomendação pedagógica, 4. Apoio ao estudante e *feedback*, 5. Detecção ou previsão de evasão, 6. Avaliação ou modelagem do estudante, 7. Teórico, 8. Objetos de aprendizagem, 9. Apoio ou gestão institucional, 10. Detecção de plágio, 11. Avaliação de sentimento.

Comumente, os trabalhos têm seus dados oriundos de diversos contextos educacionais: *(i) ambientes virtuais de aprendizagem* (QUEIROGA; CECHINEL; ARAÚJO, 2015; SANTOS; BERCHT; WIVES, 2015, SILVA, L. *et al.*, 2015, SILVA, R. *et al.*, 2015); *(ii) instituições de ensino* (BERNARDINI; COSTA; ARTIGAS, 2015; SILVA; NUNES, 2015); *(iii) dados estatísticos públicos* (FERREIRA, 2015; SILVA, MORINO; SATO, 2014). Uma busca introdutória no Banco de Teses e Dissertações (BDTD), utilizando os termos mineração de dados educacionais, apresenta 07 trabalhos, dentre os quais 06 deles são relacionados a ambientes virtuais de aprendizagem (FALCI JÚNIOR, 2010; FONSECA, 2014; GOTTARDO, 2012; KAMPFF, 2009; KIPPES, 2010; MORO, 2015).

No cenário de dados oriundos de ambientes virtuais de aprendizagem, Queiroga, Cechinel e Araújo (2015) descrevem um trabalho voltado à predição precoce de evasão de acadêmicos de um curso de educação à distância utilizando mineração dos dados

[11] Estes dados foram obtidos através de uma revisão introdutória da literatura realizada pela própria autora nos anais do CBIE nos anos de 2014 e 2015.

relativos às suas interações nas 04 semanas iniciais do curso. Foram utilizados dados extraídos diretamente do ambiente *Moodle,* oriundos das interações de 84 alunos do curso técnico em dois pólos distintos e considerando duas disciplinas introdutórias. Os resultados obtidos demonstraram a viabilidade da prática e o seu potencial enquanto instrumento preditivo para auxiliar em estratégias na diminuição da evasão.

No contexto de um curso de bacharelado em Administração Pública na modalidade EAD, também a partir da técnica de análise de agrupamentos, Silva, L. *et al.* (2015) analisaram os dados dos registros das interações dos alunos no ambiente virtual *Moodle*. Foram formados grupos distintos de alunos que possuíam características de interação similares, os padrões obtidos foram utilizados na determinação do conhecimento acerca dos grupos identificados, o que possibilitou uma percepção sobre suas interações e respectivos desempenhos na disciplina.

Visando propor indicadores para o monitoramento da qualidade da gestão educacional, Silva, R. *et al.* (2015a) realizaram mineração de dados em fóruns de discussão do ambiente *Moodle* através da técnica de agrupamento de dados. Foram identificados os principais tópicos mencionados permitindo a definição de um plano de ação e o elencamento de prioridades a partir dos resultados obtidos.

Utilizando dados de um sistema de gerenciamento acadêmico de uma escola de ensino médio da rede privada, ao longo dos anos de 2011 a 2014, Silva e Nunes (2015), realizaram análise de padrões e características de alunos que estivessem em risco de reprovação. Os resultados obtidos demonstraram que a quantidade de alunos aprovados é maior do que a de reprovados, sobretudo no 3 º ano, mas que o índice de reprovação é maior na cidade de Campina Grande. Foi possível notar ainda que há um baixo quantitativo de estudantes bolsistas desistentes.

Visando definir estratégias de capacitação docente no intuito de melhorar a qualidade do ensino, Reis e Angeloni (2010) descrevem uma experiência de descoberta de conhecimento em que esperava-se estabelecer relações entre a capacitação docente e o desempenho dos estudantes. Buscava-se, a princípio, avaliar a qualidade das

capacitações efetuadas pelo Secretaria de Educação do governo de Santa Catarina, tanto para garantir a melhoria do ensino, quanto para justificar os investimentos realizados.

Os dados foram coletados em 2004, sendo utilizados os dados de desempenho nas disciplinas de Português e Matemática de alunos do ensino fundamental e médio. Os resultados obtidos demonstraram que o processo de capacitação docente produziu o aumento do desempenho escolar dos alunos, representando, ainda que indiretamente melhoria na qualidade de ensino. Outro resultado relevante apontou uma queda de desempenho na disciplina de Português no 3º ano do ensino médio, indicando que a necessidade de reformulações estratégicas da capacitação docente com vistas a melhoria de desempenho dos alunos.

No contexto de dados abertos, Silva, Morino e Sato (2014) relatam uma experiência de mineração de dados do ENEM do ano de 2010, buscando analisar relações de causa e efeito entre o relacionamento entre o desempenho no ENEM e fatores socioeconômicos com dados de alunos apenas das capitais da região Sudeste. Este trabalho buscou analisar os seguintes aspectos: *(i) A quantidade de pessoas que mora como o aluno possui algum tipo de interferência no desempenho dele? (ii) O grau de escolaridade da mãe tem relação com a nota do aluno na prova objetiva? (iii) O valor da renda familiar mensal contribui para o desempenho do aluno? (iv) O tipo de escola em que o aluno estudo no Ensino Médio afeta o seu desempenho na prova?* A partir dos resultados obtidos através do conhecimento extraído, se pode observar que a rendar familiar baixa, a escolaridade dos pais de nível primário e uma grande quantidade de pessoas que residem com os estudantes são aspectos que contribuem para o baixo desempenho do aluno.

Ferreira (2015), por sua vez, utilizando os microdados do Censo Escolar da Educação Básica fornecido pelo INEP, descreve uma experiência de mineração utilizando árvores de decisão com o objetivo de identificar os fatores relacionados à conclusão do ensino fundamental. Os resultados obtidos demonstraram que recursos

como internet banda larga, laboratório de ciências, auditório na escola e ensino privado estão associados à maior chance do aluno concluir o ensino fundamental, bem como evidenciou necessidades especiais estão estreitamente relacionadas à não conclusão do ensino fundamental.

Namen e Soares (2011) relatam a aplicação do processo de mineração de dados para a identificação de associações entre variáveis relacionadas ao ensino de Língua Portuguesa para alunos do 5º ano do ensino fundamental do estado do Rio de Janeiro, utilizando os microdados da Prova Brasil[12] do ano de 2007 disponibilizados pelo INEP. Os resultados obtidos indicaram que alguns fatores como: falta de incentivo dos pais, reprovação prévia do aluno e atuação do aluno em trabalho doméstico e/ou fora de casa, entre outros, exerceram influência negativa sobre o aprendizado do estudante.

Mediante o exposto, nota-se a diversidade de aplicações de mineração de dados em contextos educacionais e como os resultados obtidos foram significativos para o cenário em que estavam inseridos.

[12] A Prova Brasil consiste em uma prova realizada bianualmente para todos os alunos do 5º e 9º ano do ensino fundamental.

Capítulo 03

Metodologia

Esta seção descreve em detalhes o processo metodológico aplicado no desenvolvimento deste trabalho, abrangendo desde a seleção dos dados, das tecnologias utilizadas: os algoritmos e a ferramenta de mineração utilizadas, bem com a descrição, passo a passo, de todas as fases do processo de descoberta de conhecimento.

3.1. Pré-Processamento dos Dados

O pré-processamento contempla as fases de seleção, limpeza e transformação dos dados, sendo fundamental para obtenção de padrões e conhecimentos relevantes, gerados a partir da mineração de dados.

3.1.1. Seleção dos dados: Os microdados do ENEM

Os dados analisados foram oriundos do Exame Nacional do Ensino Médio (ENEM). Criado em 1998, o ENEM possuía como objetivo *"avaliar o desempenho da educação básica, buscando contribuir para a melhoria da qualidade desse nível de escolaridade"*[13]. Em um primeiro momento, o ENEM foi criado apenas para avaliar os alunos do último ano do ensino médio das redes públicas de ensino do país, no entanto o ENEM alcançou ampla visibilidade ao longo dos anos.

Em meados de 2009, passou por reformulações e tornou-se também instrumento de seleção para o ingresso em instituições federais de ensino superior (IFES), ora enquanto fase única através do Sistema de Seleção Unificada (SiSU) ou mesmo combinado aos processos seletivos próprios da universidade. O desempenho no ENEM também é utilizado em outros programas oferecidos pelo Governo Federal, tais como: Programa Universidade para Todos (PROUNI), Fundo de Financiamento Estudantil (FIES),

[13] http://portal.inep.gov.br/web/enem/sobre-o-enem

Ciências sem Fronteiras (CsF) e inclusive para ingressar em vagas gratuitas em cursos técnicos mediados pelo Sistema de Seleção Unificada da Educação Profissional e Tecnológica (Sisutec).

O exame tem abrangência nacional e, até o presente, é realizado anualmente, cujos conteúdos abrangidos se desdobram em quatro áreas de conhecimento, a saber: 1. Ciências Humanas e suas Tecnologias, 2. Ciências da natureza e suas Tecnologias, 3. Linguagens, códigos e suas Tecnologias e Redação, 4. Matemática e suas Tecnologias.

A área de conhecimento 1 abrange quatro componentes curriculares: *(i) História, (ii) Geografia, (iii) Filosofia e (iv) Sociologia*. A área dois envolve Química, Física e Biologia, já área três envolve Língua portuguesa, Literatura, Língua Estrangeira, Artes, Educação Física e Tecnologias da Informação e Comunicação. A quarta e última envolve apenas Matemática.

O INEP disponibiliza os dados de cada participante, sem a possibilidade de identificação do indivíduo, desde o ano de 1998, sendo a versão mais recente o ano de 2014. Atualmente, estão disponíveis desde dados como o desempenho em cada uma das provas objetivas e na prova de redação, dados pessoais como estado civil, cor/raça do indivíduo, além do município e unidade federativa de residência. Estão disponíveis ainda também as informações fornecidas no questionário socioeconômico, que envolvem questões como: renda familiar, escolaridades dos pais, dentre vários outros itens.

Como é possível observar, há uma ampla e valiosa gama de informações sobre cada um dos participantes que pode ser explorada para a descoberta de conhecimento através de técnicas de mineração de dados.

No presente trabalho foram utilizados os dados provenientes dos anos de 2014 e 2013, considerando especificamente a região Nordeste e de Pernambuco, visando oferecer uma descoberta de conhecimento inserida dentro do contexto regional e local.

3.1.2. Limpeza dos dados

A limpeza dos dados realizada neste trabalho consistiu em garantir que informações inconsistentes ou mesmo errôneas fossem corrigidas ou mesmo eliminadas, a fim de não

comprometer o conhecimento a ser extraído ao fim do processo de KDD. Esta etapa compreende três funções:

(i) limpeza de informações ausentes, a qual compreende a eliminação de valores ausentes em conjuntos de dados;

(ii) limpeza de inconsistências, que abrange a identificação assim como a eliminação de valores inconsistentes;

(iii) limpeza de valores não pertencentes ao domínio: compreende a identificação e a eliminação de valores que não pertençam ao domínio dos atributos do problema.

Outro passo importante realizado nesta etapa é a aplicação de métodos para a redução dos dados antes de iniciar a busca por padrões. Isto decorre, geralmente, de restrições de espaço em memória ou mesmo tempo de processamento, devido a quantidade elevada de atributos disponíveis para análise, o que inviabilizaria a utilização de determinados algoritmos de extração de padrões.

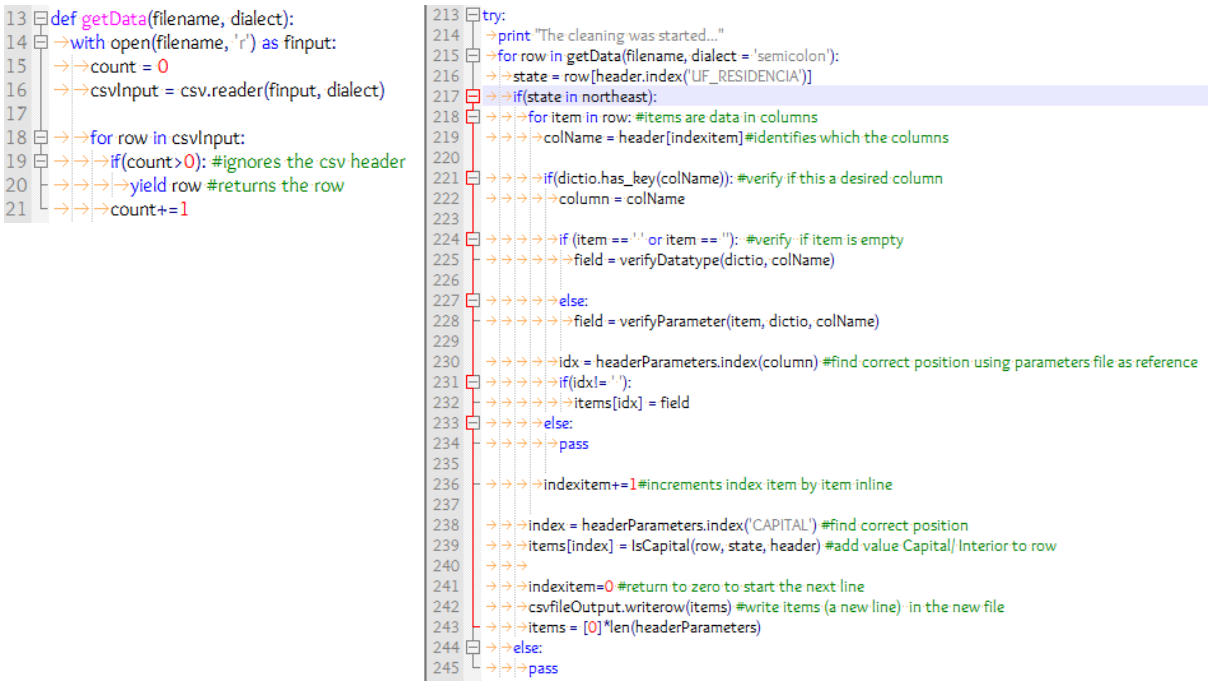
Originalmente, os dados utilizados neste trabalho obtidos do *website* do INEP estavam em formato *csv*, então o primeiro passo foi a seleção aleatória de um subconjunto dos dados reais para que pudessem ser realizados testes dos *scripts* de limpeza. Contudo, uma vez que o arquivo com os dados de 2014, por exemplo, continha mais de 5 gigabytes e mais de 8 milhões de tuplas, os editores de texto e de planilhas eletrônicas comuns, como o *Microsoft Excel, Notepad++* ou mesmo o Bloco de Notas, não tinham capacidade de abri-los. Para suprir esta necessidade, foi utilizado o software para edição de arquivos hexadecimais, *SweetScape 010 Editor*[14], em sua versão de avaliação. Este *software* foi utilizado única e exclusivamente com intuito de visualizar os arquivos de dados e criar as bases de dados mínimas para testes.

Os *scripts* de limpeza e codificação dos dados foram implementados na linguagem de programação *Python*. Um aspecto importante no processo de limpeza dos dados foi

[14] http://www.sweetscape.com/010editor/

a leitura dos arquivos originais, pois comumente, a leitura de arquivos em *Python* consiste em armazenar na memória todo o arquivo como uma lista, em que cada linha é um elemento a ser acessado diretamente por um índice. Isto ocasiona elevado consumo de memória e tempo de execução significativo, sobretudo considerando-se que o arquivo original possuía tamanho superior a vários *gigabytes*.

Figura 4. Script para leitura do arquivo com baixo consumo de memória.



A solução encontrada foi criar uma função (*getData*) que pusesse apenas uma linha por vez na memória à medida que o arquivo era lido (Figura 4): a função *getData* lê o arquivo e retorna, sempre que é chamada, apenas uma linha considerando a última posição acessada anteriormente. Uma desvantagem deste método é perder a capacidade de realizar o acesso direto a um índice específico.

Em um primeiro momento, objetivando atuar com a base de dados em nível nacional, optou-se pela redução da dimensionalidade dos dados a partir de amostragem estratificada aleatória. Este método estatístico de amostragem consiste em três passos: *(i)* Identificar e selecionar subgrupos representativos (denominados estratos) da população; *(ii)* Calcular o peso relativo (%) destes estratos na base de dados; (*iii*)

Realizar, em cada um dos estratos, um procedimento de amostragem aleatória simples para a escolha dos registros que irão compor a amostra.

Os estratos precisam ser definidos considerado a relação deles como objetivo do estudo a ser realizado, além disso todos os registros devem pertencer a um e somente um estrato. Quaisquer variáveis podem ser utilizadas enquanto critério para definir um estrato, podendo inclusive, mais de duas variáveis representarem um único estrato, para este trabalho foram consideradas as seguintes variáveis: 'CAPITAL' e 'RENDA_FAMILIA'.

A variável 'CAPITAL' possui os seguintes valores: 1. Capital/RM, 2. Interior; e a variável 'RENDA_FAMILIA' possui os valores: 1. Até 2 salários mínimos, 2. De 2 até 4 salários mínimos, 3. De 4 até 10 salários mínimos, 5. De 10 até 15 salários mínimos, 6. Acima de 15 salários mínimos.

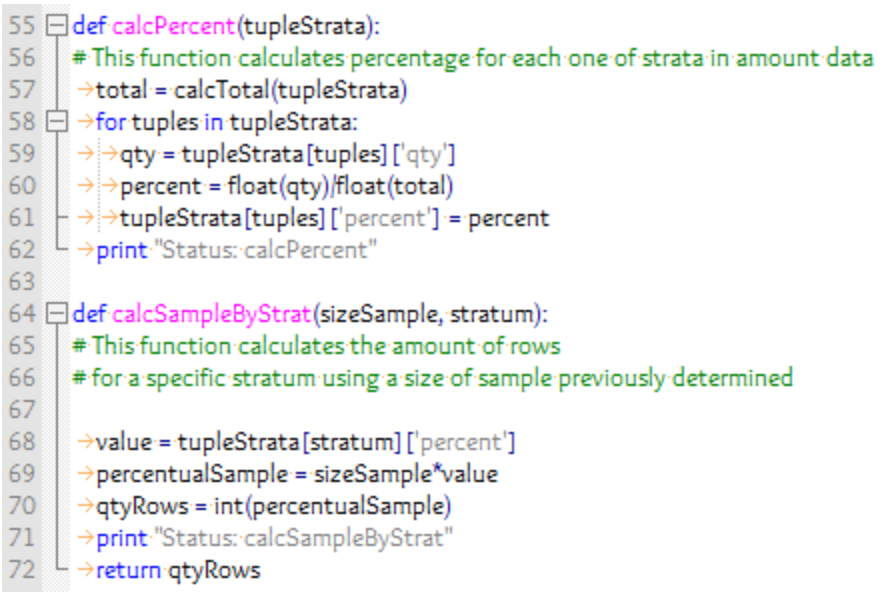
A partir destes valores foram gerados 12 estratos distintos, os quais foram armazenados como chaves de um dicionário *tupleStrata* (vide linha 27, Figura 5), que armazena também a quantidade de registros por estrato, os índices de cada registro do estrato na base de dados original e a proporção do estrato na base de dados original – inicialmente definido como zero (Figura 5). Dentre os estratos gerados se pode citar como exemplo: ('Interior', 'Até 2 salários mínimos') ou ('Capital/RM', 'De 2 até 4 salários mínimos').

Figura 5. Script para criação dos estratos

```
21  def defineTuplesStrata(parameters, strata):
22      # This function creates mix from values of columns selected
23      tupleStrata = {}
24      capitalValues = ['Capital/RM','Interior']
25      for i in parameters[strata[0]]['choices'].values():
26          for j in capitalValues:
27              tupleStrata[i,j] = {'indexes':[], 'percent': 0}
28      return tupleStrata
```

Após identificar-se os estratos, é necessário calcular a proporção de cada um deles na base de dados (vide linha 60, Figura 6). Posteriormente, definido o tamanho da amostra é preciso calcular, considerando a proporção na base de dados original, a quantidade de registros de cada estrato na amostra (vide linha 70, Figura 6).

Figura 6. Script para cálculo do percentual e quantidade de registros para cada estrato



```
55  def calcPercent(tupleStrata):
56      # This function calculates percentage for each one of strata in amount data
57      total = calcTotal(tupleStrata)
58      for tuples in tupleStrata:
59          qty = tupleStrata[tuples]['qty']
60          percent = float(qty)/float(total)
61          tupleStrata[tuples]['percent'] = percent
62      print "Status: calcPercent"
63
64  def calcSampleByStrat(sizeSample, stratum):
65      # This function calculates the amount of rows
66      # for a specific stratum using a size of sample previously determined
67
68      value = tupleStrata[stratum]['percent']
69      percentualSample = sizeSample*value
70      qtyRows = int(percentualSample)
71      print "Status: calcSampleByStrat"
72      return qtyRows
```

O próximo passo é realização um procedimento de amostragem aleatória simples para a escolha dos registros que irão compor a amostra. No entanto, a fim de preservar a replicabilidade do estudo, optou-se por gerar aleatoriamente e armazenar um conjunto de *seeds* a serem utilizados com cada estrato (função *createSeedList*, linhas 82 a 93, Figura 7). O módulo *random* é inicializado com um *seed* específico (linha 102, Figura 7) e os registros (os índices de cada um dos registros) de um estrato são selecionados na linha 103:

- *random.sample(tupleStrata[stratum]['indexes'], quantityRows)*

A variável *quantityRows* representa a quantidade de registros por estrato que já foi calculada pela função *calcSampleByStrat* (linhas 64-72, Figura 6), o método *sample* seleciona aleatoriamente um conjunto de valores de tamanho definido pela variável *quantityRows* na lista de índices de cada estrato (*tupleStrata[stratum]['indexes']*).

A variável *sampleIndexes* é uma lista que contém todos os índices de todos os registros de todos os estratos. Ao ler o arquivo da base de dados original, o *script* verifica se o índice atual está contido na lista *sampleIndexes* e se estiver, adiciona a linha ao novo arquivo criado denominado *ENEM<ano>_sample.csv*.

Figura 7. Procedimento de amostragem aleatória simples para a escolha dos registros de cada um dos estratos

```
82  def createSeedList(tupleStrata):
83      seeds = []
84      amountStrat = len(tupleStrata.keys())
85      
86      for i in range(amountStrat):
87          value = random.randint(0, 100)
88          if(value not in seeds):
89              seeds.append(value) #This creates a random seeds list without repeated numbers
90          else:
91              pass
92      print "Status: createSeedList"
93      return seeds
94  
95  def createSampleIndexes(seeds, sizeSample):
96      index = len(seeds) -1
97      sampleIndexes = []
98  
99      for stratum in tupleStrata.keys():
100         quantityRows = calcSampleByStrat(sizeSample, stratum)
101         seedValue = seeds[index]
102         value = random.seed(seedValue) #choice a number created randomly to be a seed
103         indexes = random.sample(tupleStrata[stratum]['indexes'], quantityRows)
104         createRegister(stratum, tupleStrata, seedValue, quantityRows)
105 
106         sampleIndexes = sampleIndexes+indexes #add the list with indexes randomly selected to sampleIndexes
107         index-=1
108     print "Status: createSampleIndexes"
109     return sampleIndexes
```

A partir deste método, foram geradas amostras de até 10 mil tuplas, as quais seriam relevantes para manter a pesquisa dentro do âmbito nacional, mas, no intuito de ampliar as discussões e trazê-las para um contexto mais regional e local, optou-se por atuar com os dados da região Nordeste e do estado de Pernambuco, sem redução de dimensionalidade.

Figura 8. Seleção de registros apenas da região Nordeste

```
212 northeast = ['AL','BA','CE','MA','PB','PE','PI','RN','SE']
213 try:
214     print "The cleaning was started..."
215     for row in getData(filename, dialect = 'semicolon'):
216         state = row[header.index('UF_RESIDENCIA')]
217         if(state in northeast):
```

Neste direcionamento, buscou-se realizar a remoção de valores não pertencentes ao domínio, o que consiste em identificar e selecionar os atributos mais relevantes para compor a base de dados final. Os atributos foram selecionados manualmente, considerando o grau de relevância para encontrar padrões condizentes com os objetivos deste trabalho. Foram selecionados um total de 24 atributos que abrangem tanto dados pessoais do inscrito, como idade, município de residência, como dados advindos do

questionário socioeconômico e o ano de realização do exame, um recorte do catálogo de dados é apresentado no Anexo A e um recorte dos dados brutos é apresentado no Anexo B.

Neste ínterim, foram tratados também os valores ausentes na base de dados, os quais foram preenchidos usando o símbolo '?' que o *Weka* reconhece automaticamente como nulo e desconsidera este valor quando da execução dos algoritmos. Não foi utilizado um valor global, tal qual 'INDEFINIDO', porque para os algoritmos a variável pode ser compreendida como um padrão importante ao invés de compreender de que se tratam apenas de atributos sem valor definido, sobretudo em bases em que há muito valores nulos/ausentes e a variável se repete várias vezes.

3.1.3. Transformação dos dados

No intuito de utilizar os dados como *input* para algoritmos de MD, é necessário codificá-los, transformá-los para uma maneira mais apropriada para a mineração, podendo ser agregados, generalizados ou mesmo podem ser construídos novos atributos.

Neste trabalho, dos 24 atributos selecionados, 8 foram generalizados/agrupados, dentre eles o atributo 'IDADE', em que atribuiu-se categorias como 'Até 20 anos' e 'Acima de 50 anos'. Outro exemplo, é a renda familiar que no questionário socioeconômico (QSE), foi descrita usando várias faixas salariais diferentes, com baixa granularidade, então para garantir a coerência dos dados foi criado um padrão de intervalos considerando a classificação do IBGE[15] (Quadro 1).

Quadro 1. Categorias do atributo QSE_RENDA_FAMILIA

Classificação do QSE	Classes Sociais IBGE	Classificação Utilizada
A = Nenhuma renda B = Até um salário mínimo C = Mais de 1 até 1,5 D = Mais de 1,5 até 2 salários E = Mais de 2 até 2,5 salários F = Mais de 2,5 até 3 salários	Classe A = Acima de 20 salários mínimos Classe B = De 10 a 20 salários mínimos Classe C = De 4 a 10 salários mínimos	Acima de 20 salários mínimos De 10 a 20 salários mínimos De 4 a 10 salários mínimos De 2 a 4 salários mínimos Até 2 salários mínimos Nenhuma renda

[15] http://www.datosmarketing.com.br/listas-detalhes-classes-sociais.asp

G = Mais de 3 até 4 salários H = Mais de 4 até 5 salários I = Mais de 5 até 6 salários J = Mais de 6 até 7 salários K = Mais de 7 até 8 salários L = Mais de 8 até 9 salários M = Mais de 9 até 10 salários N = Mais de 10 até 12 salários O = Mais de 12 até 15 salários P = Mais de 15 a 20 salários Q = Acima de 20 salários	Classe D = De 2 a 4 salários mínimos Classe E = Até 2 salários mínimos	

Este passo é importante também, porque na base de dados os atributos constam como valores sem significação semântica óbvia e que precisam ser relacionados com seus respectivos catálogos de dados para que se possa compreender o que significam, como no Quadro 1 em que a renda familiar acima de 20 salários é representada pela letra Q na base de dados.

Figura 9. *Script* para leitura do arquivo de parâmetros

```
def mappingParameters():
    filename = '../txt/parameters.txt'
    parameters = {}
    dictChoices = {}

    for row in getData(filename, dialect = 'pipes'):
        values = row[2] #choices

        if(len(values)<=1): #Verify if there're multiple choice values
            pass

        else:
            choices = values.rsplit('-') #Something like this: NUM1_VALUE1-NUM2_VALUE2..NUMn_VALUEn

            for i in choices:
                choice = i.rsplit('_') #Numeric aswers and respective values are represented using underline as separator like this: NUM1_VALUE1
                dictChoices[choice[0]] = choice[1].strip()

        # For example: parameters['TP_ENSINO'] = {'datatype': STRING, 'choices': {1: 'Publica', 2: 'Privada'}}
        parameters[row[0]] = {'datatype': row[1], 'choices': dictChoices}
        dictChoices = {}

    return parameters
```

Os atributos selecionados e as respectivas codificações foram descritas em um arquivo de texto (Figura 10) à parte que é lido pelo *script* de limpeza/codificação. Os atributos foram definidos da seguinte forma:

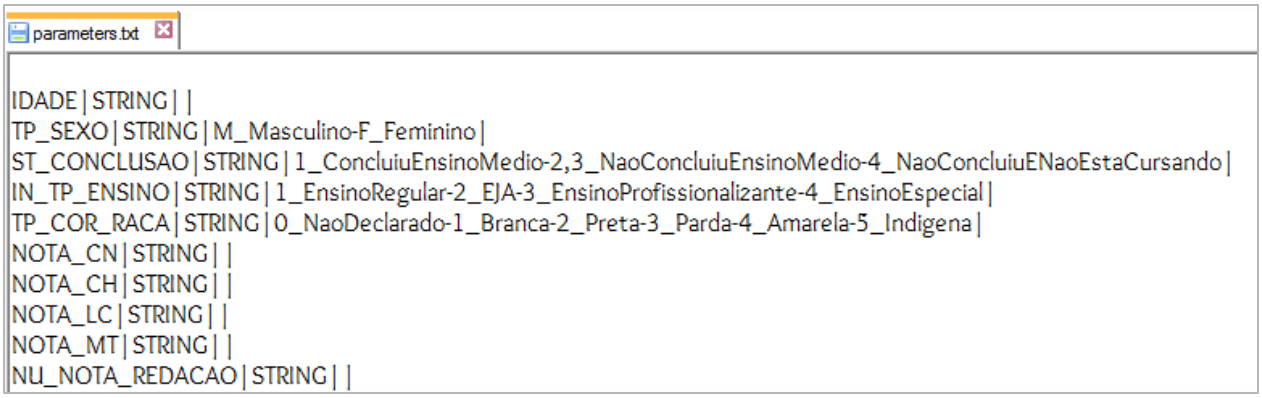
- ATTRIB|DATATYPE|OP1_COD-OP2_COD

O primeiro elemento *attrib* corresponde ao nome do atributo que era descrito manualmente conforme o cabeçalho do arquivo de dados utilizado. O elemento

datatype corresponde ao tipo de dado do atributo após a codificação, por exemplo, o atributo 'IDADE' recebeu *string* como *datatype*, pois embora o valor original fosse um número o valor final atribuído seria uma categoria.

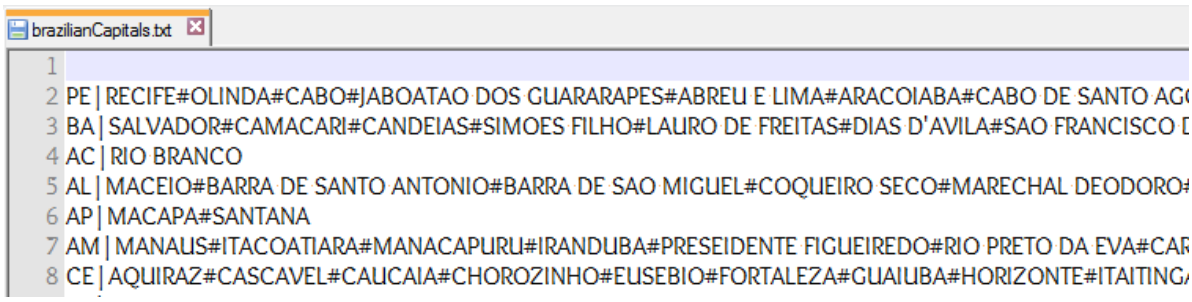
Os elementos restantes correspondem aos intervalos ou categorias, em que *op* corresponde aos valores possíveis e *cod* as codificações definidas. Os atributos de valor numérico, como idade e notas, tiveram suas codificações definidas dentro do código e por isso no arquivo de parâmetros os elementos *op* e cod estão vazios (Figura 10).

Figura 10. Trecho de arquivo com a codificação dos atributos.



Um novo atributo criado, por exemplo, foi o atributo 'CAPITAL', a partir da correlação entre os atributos 'NO_MUNICIPIO_RESIDENCIA' e 'UF_RESIDENCIA', que descrevem respectivamente o nome do munícipio e a sigla da unidade federativa em que o inscrito reside, além de um arquivo de texto contendo todas as capitais e cidades da região metropolitana de cada estado da região Nordeste (Figura 11), que foi criado pela autora para este fim.

Figura 11. Trecho de arquivo com todas as capitais do NE



O arquivo é lido (*createCapitalsList,* linhas 180-187, Figura 12) e é criado um dicionário *isCapital* que tem como chave as siglas de cada uma das unidades federativas do NE e como valor todas as cidades que são capitais ou são pertencentes às regiões metropolitanas dos estados da região Nordeste. Na leitura da base de dados, a função *IsCapital* (linhas 189-196, Figura 12) realiza a verificação.

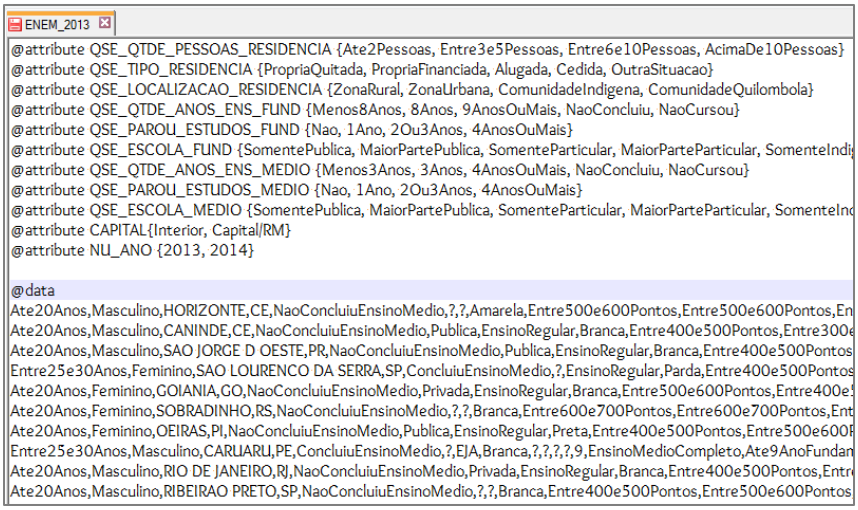
Figura 12. Script para leitura do arquivo de capitais e regiões metropolitanas do Nordeste

```
180 def createCapitalsList():
181     filename = '../txt/brazilianCapitals.txt'
182     isCapital = {}
183     for row in getData(filename, dialect = 'pipes'):
184         state = row[0]
185         cities = row[1]
186         isCapital[state] = cities
187     return isCapital
188
189 def IsCapital(row, state, header):
190     city = row[header.index('NO_MUNICIPIO_RESIDENCIA')]
191     isCapital = createCapitalsList()
192     if(city in isCapital[state]):
193         city = 'Capital/RM'
194     else:
195         city = 'Interior'
196     return city
```

Todos os dados foram organizados considerando como padrão a disposição das colunas do arquivo de dados do ano de 2014. Após a limpeza e transformação dos dados, a base ficou constituída por um total de 24 atributos (incluindo o ano de realização do exame) que são listados no Apêndice A (os atributos com valores iguais foram agrupados em uma mesma linha). Por fim, após a limpeza e a codificação dos dados, foi criado via *script* um arquivo de dados em formato *arff*, que é o formato padrão de arquivos da ferramenta *Weka*.

Neste formato é necessário relatar o domínio do atributo, que pode ser expresso pelo tipo do atributo ou nominalmente, em que se descreve os valores que ele pode representar. Assim, uma vez que todos os dados foram codificados em categorias e com o objetivo de possibilitar a utilização de tipos de algoritmos diferentes, como o *J48* ou *Apriori*, os dados foram descritos nominalmente, conforme ilustra a Figura 13.

Figura 13.Trecho de arquivo gerado em formato *arff*.

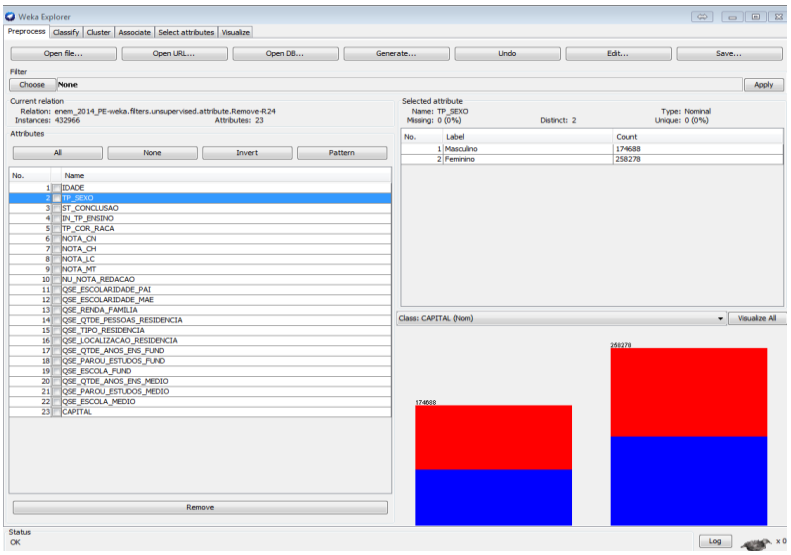


3.2. Mineração dos Dados

A mineração de dados é considerada o núcleo do processo de descoberta de conhecimento em banco de dados. Conforme anteriormente mencionado, este processo de descoberta acontece a partir da extração de informações de conjuntos de dados através de tarefas de mineração de dados.

As principais tarefas de busca de informação implícita são: agrupamento, classificação e associação. Para realizar as tarefas de mineração de dados foi utilizado o *software* livre *Weka*.

Figura 14. Software Weka com os dados carregados



Esta ferramenta foi escolhida por dispor de vários algoritmos supervisionados e não supervisionados, que contemplam tanto métodos descritivos quanto métodos preditivos, além de possuir uma interface gráfica de fácil utilização.

3.2.1. Seleção dos Algoritmos

No intuito de descobrir padrões recorrentes e relacionamentos dentro da base de dados, identificou-se que seria necessário utilizar algoritmos de regras de associação e de classificação, dentre os quais foram selecionados os algoritmos *Apriori* e o *J48*, por serem amplamente utilizados em diversos trabalhos correlatos descritos em detalhes em seções anteriores.

Capítulo 4

Resultados e Discussões

Neste trabalho, os dados do ENEM foram organizados em quatro bases distintas considerando a região Nordeste e o estado de Pernambuco nos anos de 2013 e 2014. A base de dados de 2014, englobando os dados de inscritos da região Nordeste, possui um total de 2.444.754 instâncias e a base de 2013, por sua vez, 2.358.851 instâncias.

4.1. Análise de desempenho com o algoritmo *Apriori*

Inicialmente, buscando encontrar regras de associação foi aplicado o algoritmo *Apriori*. Os parâmetros foram mantidos conforme o padrão usual da ferramenta, com o suporte mínimo de 10% por exemplo, excetuando apenas pelo parâmetro de confiança que foi estabelecido em 80% e para a quantidade de regras a serem geradas, que foi alterada para 30.

Figura 15. Trecho dos resultados da execução do *Apriori* com todos os atributos da base NE

```
Best rules found:

 1. ST_CONCLUSAO=ConcluiuEnsinoMedio QSE_QTDE_ANOS_ENS_MEDIO=3Anos 1103101 ==> IN_TP_ENSINO=EnsinoRegular 1067697
 2. QSE_ESCOLA_FUND=SomentePublica 1057048 ==> QSE_ESCOLA_MEDIO=SomentePublica 1012456    conf:(0.96)
 3. QSE_LOCALIZACAO_RESIDENCIA=ZonaUrbana QSE_QTDE_ANOS_ENS_MEDIO=3Anos 1075024 ==> IN_TP_ENSINO=EnsinoRegular 100
 4. QSE_QTDE_ANOS_ENS_MEDIO=3Anos 1293435 ==> IN_TP_ENSINO=EnsinoRegular 1209234    conf:(0.93)
 5. QSE_PAROU_ESTUDOS_MEDIO=Nao 1120384 ==> QSE_PAROU_ESTUDOS_FUND=Nao 1030894    conf:(0.92)
 6. ST_CONCLUSAO=ConcluiuEnsinoMedio 1438030 ==> IN_TP_ENSINO=EnsinoRegular 1317245    conf:(0.92)
 7. ST_CONCLUSAO=ConcluiuEnsinoMedio QSE_RENDA_FAMILIA=De1Ate2 1126237 ==> IN_TP_ENSINO=EnsinoRegular 1028034    c
 8. ST_CONCLUSAO=ConcluiuEnsinoMedio QSE_LOCALIZACAO_RESIDENCIA=ZonaUrbana 1228140 ==> IN_TP_ENSINO=EnsinoRegular
 9. IN_TP_ENSINO=EnsinoRegular QSE_QTDE_ANOS_ENS_MEDIO=3Anos 1209234 ==> ST_CONCLUSAO=ConcluiuEnsinoMedio 1067697
10. QSE_ESCOLA_MEDIO=SomentePublica 1186355 ==> QSE_PAROU_ESTUDOS_FUND=Nao 1038319    conf:(0.88)
11. QSE_PAROU_ESTUDOS_FUND=Nao 1208245 ==> QSE_LOCALIZACAO_RESIDENCIA=ZonaUrbana 1056473    conf:(0.87)
12. QSE_ESCOLA_MEDIO=SomentePublica 1186355 ==> QSE_LOCALIZACAO_RESIDENCIA=ZonaUrbana 1022504    conf:(0.86)
13. QSE_PAROU_ESTUDOS_FUND=Nao 1208245 ==> QSE_ESCOLA_MEDIO=SomentePublica 1038319    conf:(0.86)
14. ST_CONCLUSAO=ConcluiuEnsinoMedio 1438030 ==> QSE_LOCALIZACAO_RESIDENCIA=ZonaUrbana 1228140    conf:(0.85)
15. QSE_ESCOLA_MEDIO=SomentePublica 1186355 ==> QSE_ESCOLA_FUND=SomentePublica 1012456    conf:(0.85)
```

Nesta execução inicial, que contemplou 23 atributos (o atributo referente ao ano do exame não foi considerado), foram geradas 30 regras, que alcançaram o suporte mínimo de 40%, algumas delas de conhecimento comum como (Figura 15):

- Inscritos advindos do Ensino Regular e que alegaram terem levado três anos para concluir o Ensino Médio já concluíram o Ensino Médio (regra 9);
- Inscritos que declararam ter levado três anos para cursar o Ensino Médio já concluíram o Ensino Médio (regra 17).

No entanto, se pode destacar algumas regras em especial, que comprovaram cientificamente algumas associações, embora algumas delas sejam de senso comum como:

- Inscritos que realizaram o ensino fundamental somente em escola pública apresentam forte tendência de terem realizado o ensino médio apenas em escolas públicas (grau de confiança de 96%);
- Inscritos que cursaram o ensino médio apenas em escolas públicas apresentam forte tendência de serem oriundos de famílias de classe E, ou seja com renda familiar de até dois salários mínimos (grau de confiança de 94%);
- Inscritos que indicaram ter realizado seus estudos na modalidade regular têm forte tendência de residirem na zona urbana (grau de confiança de 83%);
- Inscritos oriundos do interior têm forte tendência de pertenceram a famílias de classe E (grau de confiança de 83%).

A fim de refinar as regras obtidas, foram isolados apenas os dados dos atributos do questionário socioeconômico, mantendo-se o parâmetro de confiança, o suporte mínimo e a quantidade de regras. Foram encontradas regras que asseveraram os resultados acima apresentados, tais como:

- Inscritos oriundos de famílias de classe E apresentam forte tendência de terem cursado todo o ensino fundamental e médio em escolas públicas (grau de confiança de 80%);
- Inscritos que cursaram o ensino fundamental somente em escola pública pertencerem a famílias de classe E (grau de confiança de 80%).

No intuito de comparar com o cenário local, decidiu-se utilizar a base de dados de 2014 contendo apenas dados do estado de Pernambuco (PE). Assim, mantidos os mesmos parâmetros e considerando apenas os dados do QSE, foram encontradas regras muito similares (Figura 16) que alcançaram suporte mínimo de 30%, que demonstram que o estado não está em situação diferenciada do restante da região:

- Inscritos pertencentes a famílias de classe E que concluíram o ensino fundamental somente em escola pública têm forte tendência de terem concluído o ensino médio também apenas em escola pública (grau de confiança de 97%).
- Inscritos que cursaram o ensino fundamental somente em escolas públicas tem forte tendência de terem cursado o ensino médio também apenas em escolas públicas (grau de confiança de 96%).

Figura 16. Execução do *Apriori* com os dados do QSE na base de Pernambuco

```
Best rules found:

 1. QSE_PAROU_ESTUDOS_FUND=Nao QSE_QTDE_ANOS_ENS_MEDIO=3Anos 139483 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 135532    conf:(0.97)
 2. QSE_ESCOLA_FUND=SomentePublica QSE_PAROU_ESTUDOS_MEDIO=Nao 140663 ==> QSE_ESCOLA_MEDIO=SomentePublica 135773    conf:(0.97)
 3. QSE_RENDA_FAMILIA=De1Ate2 QSE_ESCOLA_FUND=SomentePublica 143287 ==> QSE_ESCOLA_MEDIO=SomentePublica 138282    conf:(0.97)
 4. QSE_PAROU_ESTUDOS_FUND=Nao QSE_ESCOLA_FUND=SomentePublica 148263 ==> QSE_ESCOLA_MEDIO=SomentePublica 142188    conf:(0.96)
 5. QSE_ESCOLA_FUND=SomentePublica 170804 ==> QSE_ESCOLA_MEDIO=SomentePublica 163670    conf:(0.96)
 6. QSE_QTDE_ANOS_ENS_MEDIO=3Anos QSE_PAROU_ESTUDOS_MEDIO=Nao 142945 ==> QSE_PAROU_ESTUDOS_FUND=Nao 135532    conf:(0.95)
 7. QSE_QTDE_PESSOAS_RESIDENCIA=Entre3e5Pessoas QSE_PAROU_ESTUDOS_MEDIO=Nao 142332 ==> QSE_PAROU_ESTUDOS_FUND=Nao 131395    conf:(0.92)
 8. QSE_PAROU_ESTUDOS_MEDIO=Nao 200519 ==> QSE_PAROU_ESTUDOS_FUND=Nao 184249    conf:(0.92)
 9. QSE_RENDA_FAMILIA=De1Ate2 QSE_PAROU_ESTUDOS_MEDIO=Nao 155104 ==> QSE_PAROU_ESTUDOS_FUND=Nao 141384    conf:(0.91)
10. QSE_PAROU_ESTUDOS_MEDIO=Nao QSE_ESCOLA_MEDIO=SomentePublica 166686 ==> QSE_PAROU_ESTUDOS_FUND=Nao 151685    conf:(0.91)
11. QSE_RENDA_FAMILIA=De1Ate2 QSE_PAROU_ESTUDOS_MEDIO=Nao 155104 ==> QSE_ESCOLA_MEDIO=SomentePublica 136925    conf:(0.88)
12. QSE_RENDA_FAMILIA=De1Ate2 QSE_PAROU_ESTUDOS_FUND=Nao 163839 ==> QSE_ESCOLA_MEDIO=SomentePublica 143554    conf:(0.88)
13. QSE_ESCOLA_MEDIO=SomentePublica 200380 ==> QSE_PAROU_ESTUDOS_FUND=Nao 175011    conf:(0.87)
14. QSE_QTDE_PESSOAS_RESIDENCIA=Entre3e5Pessoas QSE_PAROU_ESTUDOS_FUND=Nao 150828 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 131395    conf:(0.87)
15. QSE_RENDA_FAMILIA=De1Ate2 QSE_ESCOLA_MEDIO=SomentePublica 165187 ==> QSE_PAROU_ESTUDOS_FUND=Nao 143554    conf:(0.87)
16. QSE_PAROU_ESTUDOS_FUND=Nao 212018 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 184249    conf:(0.87)
17. QSE_ESCOLA_FUND=SomentePublica QSE_ESCOLA_MEDIO=SomentePublica 163670 ==> QSE_PAROU_ESTUDOS_FUND=Nao 142188    conf:(0.87)
18. QSE_ESCOLA_FUND=SomentePublica 170804 ==> QSE_PAROU_ESTUDOS_FUND=Nao 148263    conf:(0.87)
19. QSE_PAROU_ESTUDOS_FUND=Nao QSE_ESCOLA_MEDIO=SomentePublica 175011 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 151685    conf:(0.87)
20. QSE_RENDA_FAMILIA=De1Ate2 QSE_PAROU_ESTUDOS_FUND=Nao 163839 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 141384    conf:(0.86)
21. QSE_ESCOLA_FUND=SomentePublica QSE_ESCOLA_MEDIO=SomentePublica 163670 ==> QSE_RENDA_FAMILIA=De1Ate2 138282    conf:(0.84)
22. QSE_ESCOLA_FUND=SomentePublica 170804 ==> QSE_RENDA_FAMILIA=De1Ate2 143287    conf:(0.84)
23. QSE_RENDA_FAMILIA=De1Ate2 QSE_ESCOLA_MEDIO=SomentePublica 165187 ==> QSE_ESCOLA_FUND=SomentePublica 138282    conf:(0.84)
24. QSE_ESCOLA_FUND=SomentePublica 170804 ==> QSE_PAROU_ESTUDOS_FUND=Nao QSE_ESCOLA_MEDIO=SomentePublica 142188    conf:(0.83)
25. QSE_ESCOLA_MEDIO=SomentePublica 200380 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 166686    conf:(0.83)
26. QSE_PAROU_ESTUDOS_MEDIO=Nao 200519 ==> QSE_ESCOLA_MEDIO=SomentePublica 166686    conf:(0.83)
27. QSE_ESCOLA_FUND=SomentePublica QSE_ESCOLA_MEDIO=SomentePublica 163670 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 135773    conf:(0.83)
28. QSE_RENDA_FAMILIA=De1Ate2 QSE_ESCOLA_MEDIO=SomentePublica 165187 ==> QSE_PAROU_ESTUDOS_MEDIO=Nao 136925    conf:(0.83)
29. QSE_PAROU_ESTUDOS_FUND=Nao 212018 ==> QSE_ESCOLA_MEDIO=SomentePublica 175011    conf:(0.83)
30. QSE_ESCOLA_MEDIO=SomentePublica 200380 ==> QSE_RENDA_FAMILIA=De1Ate2 165187    conf:(0.82)
```

Mediante os resultados encontrados buscou-se então verificar se o algoritmo indicaria alguma correlação entre a escola em que o inscrito cursou o ensino médio (`QSE_ESCOLA_MEDIO`), as notas obtidas no ENEM (`NOTA_CN`, `NOTA_CH`, `NOTA_LC`, `NOTA_MT`) e a renda familiar (`QSE_RENDA_FAMILIA`). Nesta análise não foi considerada a nota de redação, porque o cálculo desta prova é realizada de maneira

diferenciada das demais provas. Inicialmente, considerou-se a base de dados do NE no ano de 2014, para a qual foram obtidas algumas regras interessantes, dentre as quais se pode destacar:

- Inscritos que concluíram o Ensino Médio em escolas públicas apresentam forte tendência de pertencerem a famílias de classe E (suporte de 40%, confiança de 84%);
- Inscritos com nota entre 400 e 500 pontos em Ciências da Natureza têm forte tendência de terem renda de até 2 salários mínimos (suporte de 30%, confiança de 83%);
- Inscritos com nota entre 400 e 500 pontos em Matemática apresentam forte tendência de pertencerem a famílias de classe E (suporte de 20%, confiança de 82%);
- Inscritos com nota entre 400 e 500 pontos em Ciências Humanas que cursaram ensino médio apenas em escolas públicas apresentam forte tendência de pertencerem a famílias de classe E (suporte de 10%, confiança de 86%).

A mesma análise sobre a base de dados de PE no ano de 2014, por sua vez, não gerou distintas do cenário regional, dentre as quais se pode enumerar ao menos duas que endossam os resultados já encontrados:

- Inscritos que concluíram o ensino médio somente em escola pública possuem forte tendência de pertencerem a famílias de renda familiar de até dois salários mínimos (suporte de 30%, confiança de 82%);
- Inscritos com nota entre 400 e 500 pontos em Ciências Humanas que cursaram ensino médio apenas em escolas públicas apresentam forte tendência de pertencerem a famílias de classe E (suporte de 30%, confiança de 81%).

Observando-se as regras de associação encontradas, nota-se diversas associações entre notas de até 500 pontos com a renda familiar baixa (classe E – até dois salários

mínimos), bem como associações com a conclusão do ensino médio somente em escola pública como indicador dos inscritos serem oriundos de família de classe E.

Neste sentido, buscou-se então verificar se o algoritmo indicaria correlação entre o desempenho apenas na prova de Matemática (`NOTA_MT`), o da renda familiar (`QSE_RENDA_FAMILIA`) e o do local da residência (`CAPITAL`).

O local de residência foi levado em consideração, uma vez que embora no cenário regional as famílias com baixa renda estejam concentradas no interior do estado, em Pernambuco esta distribuição ocorre tanto na capital e regiões metropolitanas quanto em cidades do interior. A nota da disciplina de Matemática foi enfatizada justamente porque é a disciplina que possui mais notas baixas, depois da disciplina de Redação.

Na execução do *Apriori* sobre a base NE em 2014, foram encontradas 5 regras, dentre as quais se pode destacar:

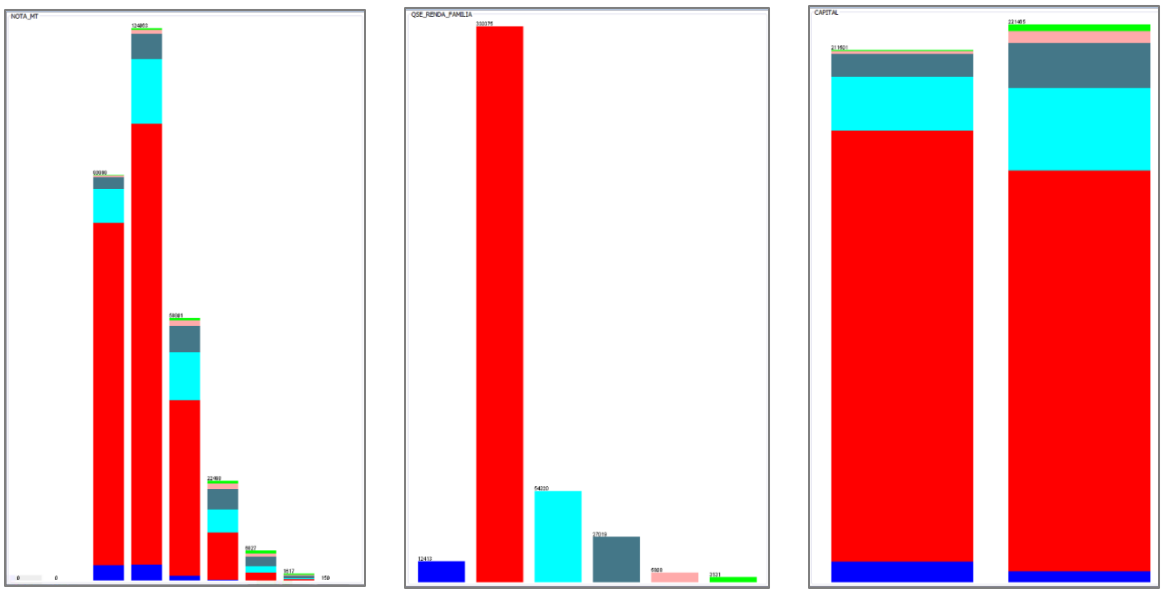
- Inscritos residentes no interior têm forte tendência de pertencerem a famílias com renda de até 2 salários mínimos (83% de confiança, suporte de 40%)
- Inscritos com notas variando entre 400 e 500 pontos no ENEM têm forte tendência de pertencerem a famílias de classe E (82% de confiança, suporte de 20%).
- Inscritos com notas variando entre 300 e 400 pontos no ENEM têm forte tendência de pertencerem a famílias de classe E (87% de confiança, suporte de 10%);

A execução do *Apriori* sobre a base PE resultou em regras similares:

- Inscritos residentes no interior têm forte tendência de pertencerem a famílias com renda de até 2 salários mínimos (81% de confiança, suporte de 30%);
- Inscritos com notas variando entre 300 e 400 pontos no ENEM têm forte tendência de pertencerem a famílias de classe E (84% de confiança, suporte de 10%).

Um outro aspecto importante que é possível observar nos gráficos apresentados na Figura 17, é que há uma expressiva quantidade de inscritos da classe E (cor vermelha) com notas inferiores a 500 pontos, ou seja, inferior ao desempenho médio nacional e até mesmo do desempenho médio do nordeste. À medida que a pontuação aumenta, a expressividade dos inscritos de classe E diminui gradualmente.

Figura 17. Gráficos dos atributos `NOTA_MT`, `QSE_RENDA_FAMILIA` e `CAPITAL` em Pernambuco



Convém mencionar que devido a método psicométrico de correção das provas, a Teoria de Resposta ao Item, nenhum inscrito recebe nota zero, mas sim uma nota mínima que varia todos os anos. As notas mínimas, médias e máximas do ano de 2014 são apresentadas na Tabela 2.

Tabela 2.Proficiência dos participantes do ENEM 2014 - Provas objetivas

	Nota mínima	Nota máxima	Nota média
Ciências Humanas (CH)	324,8	862,1	546,5
Ciências da Natureza (CN)	330,6	876,4	482,2
Linguagens e Códigos (LC)	306,2	814,2	507,9
Matemática (MT)	318,5	973,6	473,5

Embora o Nordeste seja a segunda região do país com maior número de participantes, perdendo apenas para a região Sudeste, no entanto, o desempenho médio

dos inscritos do NE foi inferior à média nacional em todas as áreas no ano de 2014[16] (Tabela 3).

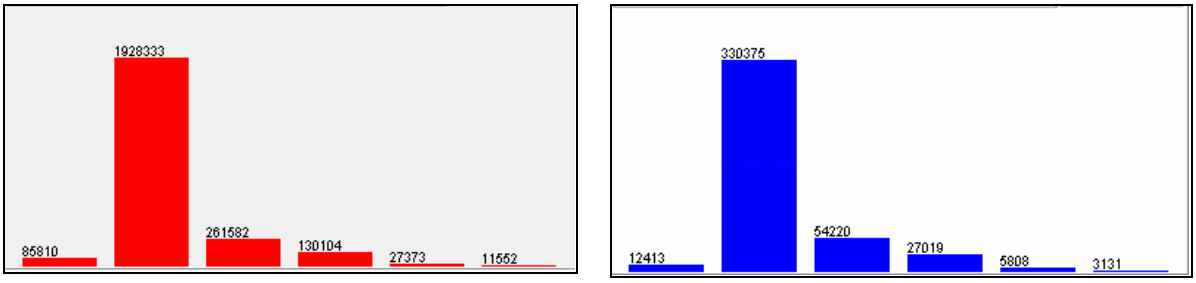
Tabela 3. Desempenho no ENEM 2014 por Região

Região	Participantes	Ciências humanas	Ciências da natureza	Linguagens e códigos	Matemática	Redação
Centro-Oeste	8,4%	542,6	480,7	503,3	467,3	437,6
Nordeste	33,7%	533,9	471	495,9	456,1	434,9
Norte	10,9%	529,9	464,8	487,1	442,7	417,5
Sudeste	34,9%	561,2	495,8	523,7	496,5	486,9
Sul	11,9%	557,7	491,2	517,8	487,8	468,9
Média Brasil	100%	546,5	482,2	507,9	473,5	455,4

Além disso, aproximadamente 82% dos inscritos oriundos do NE (2.014.143 pessoas) pertencem a classe E, ou seja, declararam possuir renda até 2 salários mínimos, dentre os quais 85.810 participantes afirmaram não possuir qualquer renda.

No cenário de Pernambuco, cerca de 80% (342.788) dos inscritos (432.966 pessoas) declararam pertencer a classe E, dentre os quais 12.413 pessoas afirmaram não possuir qualquer fonte de renda.

Figura 18. QSE_RENDA_FAMILIA – NE (2014) e QSE_RENDA_FAMILIA – PE (2014)



Em uma análise mais detalhada, focando apenas no cenário de Pernambuco e observando apenas notas não nulas de Matemática (304.955), nota-se que aproximadamente 70%(215.062) das notas estão na faixa entre 300 e 500 pontos, a partir do que é possível observar que:

[16] http://www.correiobraziliense.com.br/app/noticia/eu-estudante/especial_enem/2015/01/13/especial-enem-interna,466144/inep-revela-media-de-notas-dos-alunos-no-enem-2014.shtml

- 30% (90.999) de todos participantes obtiveram notas entre 300 e 400 pontos, o que é abaixo da nota média nacional (473,5) e até mesmo da média regional (442,7).
    - 88% (80.301) são pertencentes a famílias de classe E, isto representa aproximadamente 23,4% de todo o total de inscritos que declararam ser pertencentes a famílias de classe E;
    - 46% (42.485) declararam terem cursado todo ele ou a maior parte do ensino médio em escolas públicas;
    - 41% (38.003) são oriundos de famílias de classe E e declararam terem cursado o ensino em escolas públicas ou a maior parte dele.
- 40% (124.063) de todos participantes obtiveram notas entre 400 e 500 pontos.
    - 82% (102.521) são pertencentes a famílias de classe E, isto representa aproximadamente 30% de todos os inscritos que declararam renda de até dois salários mínimos;
    - 45% (55.869) declararam terem cursado todo o ensino médio ou a maior parte dele em escola pública;
    - 38% (47.989) são pertencentes a famílias de classe E e declararam terem cursado todo o ensino médio ou a maior parte dele em escola pública.

Assim, é possível constatar que aproximadamente 53% (182.822) de todos os inscritos que declararam renda familiar de até 2 salários mínimos obtiveram desempenho de no máximo 500 pontos na prova de Matemática. Considerando que houve 215.062 notas entre 300 e 500 pontos, 85% das notas mais baixas de Matemática são de inscritos oriundos de famílias de classe E.

4.2. Análise de desempenho utilizando o algoritmo *J48*

A fim de analisar o desempenho dos inscritos nas quatro áreas de conhecimento, tanto no cenário regional quanto no cenário local, considerando o gênero como atributo classificador, foi utilizado o algoritmo *J48*, mantidos os parâmetros padrão da ferramenta *Weka*. Os atributos considerados na análise foram `NOTA_MT`, `NOTA_CN`, `NOTA_CH`, `NOTA_LC` e `SEXO` (atributo classificador). A fim garantir maior precisão dos algoritmos os dados foram utilizados da seguinte forma: *(i)* Abrangendo-se apenas 60% dos dados para criação do modelo (*Percentual split*), *(ii)* Utilizou-se a opção *Use training set* e *(iii)* Utilizou-se a opção *Cross validation* com 10 *folds* (que recombina os dados *n* vezes para validar o modelo).

Inicialmente, no cenário estadual, os modelos gerados, sem variação significativa de instâncias classificadas corretamente, indicaram que estudantes do sexo Feminino de costumam ter notas inferiores ou iguais a *NotaMedia*, ou seja, concentrando-se nas *NotaMin*, *EntreNotaMinMedia* e Nota *Media*.

Figura 19. Árvore de decisão correlacionando sexo e desempenho no exame

```
J48 pruned tree
------------------

NOTA_MT = NotaMin: Feminino (117.84/42.84)
NOTA_MT = EntreNotaMinMedia: Feminino (268972.61/95113.16)
NOTA_MT = NotaMedia: Feminino (168.95/66.84)
NOTA_MT = EntreNotaMediaMax
|   NOTA_CN = NotaMin: Feminino (0.0)
|   NOTA_CN = EntreNotaMinMedia
|   |   NOTA_LC = NotaMin: Feminino (3.13/1.4)
|   |   NOTA_LC = EntreNotaMinMedia
|   |   |   NOTA_CH = NotaMin: Feminino (0.0)
|   |   |   NOTA_CH = EntreNotaMinMedia: Feminino (21065.46/9496.13)
|   |   |   NOTA_CH = NotaMedia: Feminino (18.41/7.99)
|   |   |   NOTA_CH = EntreNotaMediaMax: Masculino (8448.47/3919.28)
|   |   |   NOTA_CH = NotaMax: Feminino (0.0)
|   |   NOTA_LC = NotaMedia: Feminino (39.63/18.01)
|   |   NOTA_LC = EntreNotaMediaMax: Feminino (29996.6/12675.38)
|   |   NOTA_LC = NotaMax: Feminino (0.0)
|   NOTA_CN = NotaMedia: Masculino (74.55/37.12)
|   NOTA_CN = EntreNotaMediaMax
|   |   NOTA_LC = NotaMin: Masculino (0.0)
|   |   NOTA_LC = EntreNotaMinMedia: Masculino (18176.6/7577.69)
|   |   NOTA_LC = NotaMedia
|   |   |   NOTA_CH = NotaMin: Feminino (0.0)
|   |   |   NOTA_CH = EntreNotaMinMedia: Feminino (8.45/2.97)
|   |   |   NOTA_CH = NotaMedia: Masculino (0.0/0.0)
|   |   |   NOTA_CH = EntreNotaMediaMax: Masculino (30.54/14.14)
|   |   |   NOTA_CH = NotaMax: Feminino (0.0)
|   |   NOTA_LC = EntreNotaMediaMax: Feminino (85841.91/42079.01)
|   |   NOTA_LC = NotaMax: Masculino (0.0)
|   NOTA_CN = NotaMax: Feminino (0.0)
NOTA_MT = NotaMax: Masculino (2.84/0.51)

Number of Leaves  :     25

Size of the tree :      31
```

Os mesmos algoritmos, com os mesmos atributos e configurações foram aplicados a base de dados do ano de 2013 e obteve-se resultados muito similares, asseverando as descobertas encontradas.

4.3. Análise de redações com notas zero

Uma outra análise realizada abrangeu as redações com notas zero a fim de encontrar possíveis padrões. Esta análise, em especial, abrangeu o cenário nacional, contemplando um total 328.440 instâncias, excluindo-se os registros com notas zero na redação por não comparecimento à prova.

Observou-se ainda que as notas zeradas agruparam-se em sua maioria, em três grupos principais: *Fuga ao tema* (216.239), *Em branco* (75.538) e *Cópia* (12.809). As redações classificadas como *"Fuga ao tema"* indicavam que as redações tiveram profundos desvios em suas temáticas, as redações classificadas como *"Cópia"* indicavam que os alunos inseriram em seus trabalhos cópias de trechos de textos motivadores ou de frases de efeito.

Com este objetivo foi utilizado o algoritmo *J48* considerando-se os seguintes atributos: `STATUS_REDACAO` (atributo classificador), `QSE_RENDA_FAMILIA`, `QSE_ESCOLARIDADE_PAI`, `QSE_ESCOLARIDADE_MAE`. No intuito de garantir maior precisão dos algoritmos os dados foram utilizados da seguinte forma: *(i)* Abrangendo-se apenas 60% dos dados para criação do modelo (*Percentual split*), *(ii)* Utilizou-se a opção *Use training set* e *(iii)* Utilizou-se a opção *Cross validation* com 10 *folds* (que recombina os dados *n* vezes para validar o modelo).

O atributo `STATUS_REDACAO` possui os seguintes valores possíveis: Em branco, anulada, Não atende, Texto insuficiente, Atende, Fere os Direitos Humanos, Cópia, Inserção de trecho incoerente com o contexto. Os resultados obtidos concentraram-se em duas variáveis:

- Inscritos oriundos das classes A, B e C apresentaram tendência de entregar as redações em branco;

- Inscritos oriundos das classes D e E possuem tendência de entregar redações com Fuga ao tema.

Adicionalmente a escolaridade dos pais não demonstrou influenciar o rendimento na prova de redação.

Em um segundo momento o mesmo conjunto de dados foi submetido ao algoritmo *Apriori*, mas não foram encontradas regras que atendessem o suporte mínimo de 10% e grau de confiança de, pelo menos, 80%.

Capítulo 5

Considerações Finais

Ao longo deste trabalho, buscou-se identificar regras de associação e modelos de classificação a partir dos dados dos inscritos no ENEM, com ênfase para os dados advindos do questionário socioeconômico. Buscou-se encontrar possíveis relações, principalmente, entre o local de residência do estudante, a renda familiar, a escola em que os participantes cursaram ensino fundamental e médio, e se estes aspectos possuíam alguma relação com o respectivo desempenho nas provas do ENEM.

As análises de desempenho utilizando o algoritmo *Apriori* endossam uma forte relação entre a renda da família e o desempenho dos estudantes no exame, sobretudo, os provenientes de escolas públicas.

As análises de desempenho envolvendo questões de gênero utilizando o algoritmo *J48*, indicaram a tendência de que inscritos do sexo feminino tenham desempenho variando da nota mínima a nota média.

As análises, considerando apenas notas zeradas, indicaram que inscritos oriundos das classes A, B e C apresentaram tendência de entregar as redações em branco, ao passo que inscritos oriundos das classes D e E possuem tendência de entregar redações com Fuga ao tema. Adicionalmente a escolaridade dos pais não demonstrou influenciar o rendimento na prova de redação.

Os resultados obtidos, alguns deles considerados óbvios, podem ser resultantes de fatores como: base de dados reduzida, escolha inadequada do algoritmo a ser utilizado, escolha inadequada de atributos, dentre outros. A fim de obter resultados relevantes, deve-se considerar grandes bases de dados, a escolha adequada dos algoritmos e atributos, dentre outros. Por exemplo, caso a base utilizada abrangesse todo o Brasil talvez os resultados fossem distintos dos encontrados.

No entanto, a capacidade de processamento dos computadores utilizados foi um dos fatores mais impactantes (e limitantes) ao longo de todo este trabalho, ora na etapa de pré-processamento dos dados, formatação e principalmente na etapa de mineração dos dados. Neste direcionamento, foram tomadas diversas medidas a fim de contornar as limitações encontradas sem comprometer o desenvolvimento do trabalho, tais como a redução do escopo inicial do trabalho e da base de dados.

A configuração das principais máquinas disponíveis (*notebook* pessoal e *desktops* dos laboratórios da universidade) era de 8 *gigabytes* de memória RAM e processadores quatro núcleos com 2.9 Ghz frequência, excetuando por um *desktop* pessoal de 12 *gigabytes* de memória RAM. A memória sobre o qual o *Weka* opera, por exemplo, foi alterada para valores entre 5 e 6 *gigabytes*, a fim de maximizar a capacidade de execução dos algoritmos.

Em trabalhos futuros, espera-se integrar várias bases de dados, como a do Censo Escolar, ENEM e do Censo de Educação Superior, a fim de identificar padrões e modelos que suportem a criação de estratégias para um melhor acolhimento do estudante recém ingresso dentro do cenário universitário.

Espera-se ainda, em trabalhos futuros, atuar com programação paralela e métodos de amostragem, uma vez que no trabalho atual foi utilizada programação sequencial.

Assim, espera-se contribuir, através de trabalhos futuros que serão embasados no trabalho atual, na oferta de subsídios para a criação e/ou fortalecimento de iniciativas que visem melhorias na diminuição da evasão e retenção dos estudantes universitários, avaliação e diagnóstico do corpo discente, bem como auxiliar, por exemplo, os gestores públicos na criação/consolidação de ações afirmativas dentro da universidade, que vão desde o estabelecimento de cotas, programas de nivelamento a programas de bolsas de estudo e de apoio acadêmico.

Referências

AGRAWAL, Rakesh et al. Fast Discovery of Association Rules. Advances in knowledge discovery and data mining, v. 12, n. 1, p. 307-328, 1996.

BAKER, Ryan S.J.d.; YACEF, Kalina. The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining, v. 1, n. 1, p. 3-17, 2009.

BAKER, Ryan. S. J. d. Data mining for education. International encyclopedia of education, v. 7, p. 112-118, 2011.

BERNARDINI, Flavia; COSTA, Jefferson; ARTIGAS, Danilo. Proposta de uma técnica de Mineração em Grafos para identificação de gargalos em currículos de graduação. In: Anais do Simpósio Brasileiro de Informática na Educação. 2015. p. 1062.

BRANDÃO, Maria de Fátima Ramos; DOS SANTOS RAMOS, Carlos Renato; TRÓCCOLI, Bartholomeu T. Análise de agrupamento de escolas e Núcleos de Tecnologia Educacional: mineração na base de dados de avaliação do Programa Nacional de Informática na Educação. In: Anais do Simpósio Brasileiro de Informática na Educação. 2003. p. 366-374.

BRILHADORI, Melina; LAURETTO, Marcelo S. Estudo comparativo entre algoritmos de árvores de classificação e máquinas de vetores suporte, baseados em ensembles de classificadores. In: Anais do Simpósio Brasileiro de Sistemas de Informação. 2013. p. 97-108.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Goiânia: Universidade Federal de Goiás, 2009.

FALCI JÚNIOR, Geraldo Ramos. Metodologias de mineração de dados para ambientes online. 2010. Dissertação de Mestrado – Faculdade de Engenharia Elétrica e de Computação. Universidade Estadual de Campinas. Campinas, SP.

FAYYAD, Usama M. et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: KDD. 1996. p. 82-88.

FERREIRA, Gisele. Investigação acerca dos fatores determinantes para a conclusão do Ensino Fundamental utilizando Mineração de Dados Educacionais no Censo Escolar da Educação Básica do INEP 2014. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2015. p. 1034.

FONSECA, Stella Oggioni da. Utilização de modelos de classificação para mineração de dados relacionados à aprendizagem de matemática e ao perfil de professores do ensino fundamental. 2014. Dissertação de Mestrado – Centro de Ciência e Tecnologia – Instituto Politécnico. Universidade do Estado do Rio de Janeiro. Nova Friburgo, RJ.

GOTTARDO, Ernani. Estimativa de desempenho acadêmico de estudantes em um AVA utilizando técnicas de mineração de dados. 2012. Dissertação de Mestrado. Universidade Tecnológica Federal do Paraná. Curitiba, PR.

GOUVEIA, Roberta Macêdo Marques. Mineração de Dados em Data Warehouse para sistema de abastecimento de água. 2009. Dissertação de Mestrado. Universidade Federal da Paraíba, João Pessoa, PB.

KAMPFF, Adriana Justin Cerveira. Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. 2009. Tese de Doutorado - Centro de Estudos Interdisciplinares em Novas Tecnologias da Educação. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS.

KIPPES, Alexander. Mineração de dados e análise do comportamento do consumidor: estudo de caso em um website de ensino a distância. 2010. Dissertação de Mestrado. Universidade Federal de Lavras. Lavras, MG.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, Patricia Mariotto. Análise dos Algoritmos de Mineração J48 e Apriori Aplicados na Detecção de Indicadores da Qualidade de Vida e Saúde. Revista Interdisciplinar de Ensino, Pesquisa e Extensão, v. 1, n. 1, 2014.

LUAN, Jing. Data Mining Applications in Higher Education. 2007.

MACEDO, Dayana Carla; MATOS, Simone Nasser. Extração de conhecimento através da mineração de dados. Revista de Engenharia e Tecnologia, v. 2, n. 2, p. Páginas 22-30, 2010.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas Inteligentes-Fundamentos e Aplicações, v. 1, p. 1, 2003.

MORO, L. F. de S. Caracterização de alunos em ambientes de ensino online: estendendo o uso da DAMICORE para minerar dados educacionais. 2015. Dissertação de Mestrado - Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo. São Carlos, SP.

NAMEN, Anderson Amendoeira; SOARES, ACS. Mineração de dados relacionados ao aprendizado de Língua Portuguesa: um estudo exploratório. Encontro de Modelagem Computacional, v. 14, p. 295-304, 2011.

QUEIROGA, Emanuel; CECHINEL, Cristian; ARAÚJO, Ricardo. Um Estudo do Uso de Contagem de Interações Semanais para Predição Precoce de Evasão em Educação a Distância. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2015. p. 1074.

REIS, Eduardo Squario; ANGELONI, Maria Terezinha; SERRA, Fernando Ribeiro. Business Intelligence como Tecnologia de Suporte a Definição de Estratégias para a Melhoria da Qualidade de Ensino. Informação & Sociedade: Estudos, v. 20, n. 3, 2010.

RODRIGUES, Rodrigo Lins et al. A literatura brasileira sobre mineração de dados educacionais. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2014.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, v. 40, n. 6, p. 601-618, 2010.

SANTOS, Fabricia Damando et al. Análise de Evidências do Estado de Ânimo Desanimado de Alunos de um AVEA: uma Proposta a partir da Aplicação de Regras de Associação. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2015. p. 1054.

SANTOS, Fabricia Damando; BERCHT, Magda; WIVES, Leandro. Classificação de alunos desanimados em um AVEA: uma proposta a partir da mineração de dados educacionais. In: Anais do Simpósio Brasileiro de Informática na Educação. 2015. p. 1052.

SILVA, Jéssica; NUNES, Isabel. Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. In: Anais do Simpósio Brasileiro de Informática na Educação. 2015. p. 1112.

SILVA, Leandro A. et al. Prática de Mineração de Dados no Exame Nacional do Ensino Médio. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2014.

SILVA, Leandro Augusto et al. Mineração de Dados em publicações de Fóruns de Discussões do Moodle como geração de Indicadores para aprimoramento da Gestão Educacional. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2015. p. 1084.

SILVA, Ricardo et al. Mineração de dados educacionais na análise das interações dos alunos em um Ambiente Virtual de Aprendizagem. In: Anais do Simpósio Brasileiro de Informática na Educação. 2015. p. 1197.

USDE. United States - Department of Education, Office of Educational Technology, Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief, Washington, D.C., 2012.

VASCONCELOS, Lívia Maria Rocha de; CARVALHO, Cedric Luiz de. Aplicação de Regras de Associação para Mineração de Dados na Web. Brasil, Universidade Federal do Rio Grande do Sul, p. 11-14, 2004.

WEST, Darrell M. Big data for education: Data mining, data analytics, and web dashboards. Governance Studies at Brookings, p. 1-10, 2012.

Apêndice A – Atributos selecionados e respectivos valores nominais

Atributo	Categorias
IDADE	{Ate20Anos, Entre20e25Anos, Entre25e30Anos, Entre30e40Anos, AcimaDe40Anos}
TP_SEXO	{Masculino, Feminino}
ST_CONCLUSAO	{ConcluiuEnsinoMedio, NaoConcluiuEnsinoMedio, NaoConcluiuENaoEstaCursando}
IN_TP_ENSINO	{EnsinoRegular, EJA, EnsinoProfissionalizante, EnsinoEspecial}
TP_COR_RACA	{NaoDeclarado, Branca, Preta, Parda, Amarela, Indigena}
NOTA_CN / NOTA_CH/ NOTA_LC/ NOTA_MT/ NU_NOTA_REDACAO	{Ate200Pontos, Entre200e300Pontos, Entre300e400Pontos, Entre400e500Pontos, Entre500e600Pontos, Entre600e700Pontos, Entre700e800Pontos, Entre800e900Pontos, ApartirDe900Pontos}
QSE_ESCOLARIDADE_PAI / QSE_ESCOLARIDADE_MAE	{NaoEstudou, Ate5AnoFundamental, Ate9AnoFundamental, EnsinoMedioIncompleto, EnsinoMedioCompleto, SuperiorIncompleto, SuperiorCompleto, PosGraduacao, NaoSei}
QSE_RENDA_FAMILIA	{NenhumaRenda, De1Ate2, De2Ate4, De4Ate10, De10Ate20, AcimaDe20}
QSE_QTDE_PESSOAS_RESIDENCIA	{Ate2Pessoas, Entre3e5Pessoas, Entre6e10Pessoas, AcimaDe10Pessoas}
QSE_TIPO_RESIDENCIA	{PropriaQuitada, PropriaFinanciada, Alugada, Cedida, OutraSituacao}
QSE_LOCALIZACAO_RESIDENCIA	{ZonaRural, ZonaUrbana, ComunidadeIndigena, ComunidadeQuilombola}
QSE_QTDE_ANOS_ENS_FUND	{Menos8Anos, 8Anos, 9AnosOuMais, NaoConcluiu, NaoCursou}
QSE_PAROU_ESTUDOS_FUND/ QSE_PAROU_ESTUDOS_MEDIO	{Nao, 1Ano, 2Ou3Anos, 4AnosOuMais}
QSE_ESCOLA_FUND/ QSE_ESCOLA_MEDIO	{SomentePublica, MaiorPartePublica, SomenteParticular, MaiorParteParticular,

	SomenteIndigena, MaiorParteIndigena, SomenteQuilombola, MaiorParteQuilombola}
QSE_QTDE_ANOS_ENS_MEDIO	{Menos3Anos, 3Anos, 4AnosOuMais, NaoConcluiu, NaoCursou}
CAPITAL	{Interior, Capital/RM}

Anexo A – Recorte do Dicionário de Dados do ENEM 2013

A	B	C	D	E	F
DICIONÁRIO DAS VARIÁVEIS – ENEM 2013					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
DADOS DE INSCRIÇÃO					
NU_INSCRICAO	Número de inscrição[1]			12	Numérico
NU_ANO	Ano do Enem			4	Numérico
COD_MUNICIPIO_RESIDENCIA	Código do município de residência			7	Numérico
	1º dígito: Região				
	1º e 2º dígitos: UF				
	3º, 4º, 5º e 6º dígitos: Município				
	7º dígito: dígito verificador				
NO_MUNICIPIO_RESIDENCIA	Nome do município de residência			150	Alfanumérico
COD_UF_RESIDENCIA	Código da Unidade da Federação de residência				
UF_RESIDENCIA	Sigla da Unidade da Federação de residência			2	Alfanumérico
IN_ESTUDA_CLASSE_HOSPITALAR	Indicador de inscrição em Unidade Hospitalar	1	Sim	1	Numérico
		0	Não		
DADOS DA ESCOLA					
COD_ESCOLA	Código da Escola[1]			8	Numérico
COD_MUNICIPIO_ESC	Código do município da escola			7	Numérico
	1º dígito: Região				
	1º e 2º dígitos: UF				
	3º, 4º, 5º e 6º dígitos: Município				
	7º dígito: dígito verificador				
NO_MUNICIPIO_ESC	Nome do município da escola			150	Alfanumérico
COD_UF_ESC	Código da Unidade da Federação da escola				
UF_ESC	Sigla da Unidade da Federação da escola			2	Alfanumérico
ID_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)	1	Federal	1	Numérico
		2	Estadual		
		3	Municipal		
		4	Privada		
ID_LOCALIZACAO_ESC	Localização (Escola)	1	Urbana	1	Numérico
		2	Rural		
SIT_FUNC_ESC	Situação de funcionamento (Escola)	1	Em atividade	1	Numérico
		2	Paralisada		
		3	Extinta		

Anexo B – Recorte dos dados brutos do ENEM 2013

MICRODADOS_ENEM_2013_BSMIN.csv