Pós-Graduação em Ciência da Computação

ROGÉRIO LUIZ CARDOSO SILVA FILHO

MODELO DE ANÁLISE E PREDIÇÃO DO DESEMPENHO DOS ALUNOS DOS INSTITUTOS FEDERAIS DE EDUCAÇÃO USANDO O ENEM COMO INDICADOR DE QUALIDADE ESCOLAR.

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE

2017

Rogério Luiz Cardoso Silva Filho

Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar.

Este trabalho foi apresentado à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

ORIENTADOR: **Prof. Dr. Paulo Jorge Leitão Adeodato**

RECIFE

2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S586m Silva Filho, Rogério Luiz Cardoso
Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar / Rogério Luiz Cardoso Silva Filho. – 2017.
93 f.:il., fig., tab.

Orientador: Paulo Jorge Leitão Adeodato.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2017.
Inclui referências e apêndices.

1. Mineração de dados. 2. Desempenho escolar. I. Adeodato, Paulo Jorge Leitão (orientador). II. Título.

006.312 CDD (23. ed.) UFPE- MEI 2017-236

Rogério Luiz Cardoso Silva Filho

Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre Profissional em 18 de agosto de 2017.

Aprovado em: ___/___/_____.

BANCA EXAMINADORA

_____
Prof. Alex Sandro Gomes
Centro de Informática / UFPE

_____
Profª. Patrícia Smith Cavalcante
Centro de Educação / UFPE

_____
Prof. Paulo Jorge Leitão Adeodato
Centro de Informática / UFPE
(Orientador)

Dedico este trabalho aos meus pais pelo amor e por não terem medido esforços para a minha formação educacional. Dedico também a minha avó Joana e minha tia Enedina (In Memoriam) por todo amor emanado.

Agradecimentos

Agradeço a Deus, pelo dom da vida, saúde e perseverança. A minha noiva Maria Alice pelo amor, apoio e revisões sempre pontuais. Ao meu orientador, Paulo Adeodato, pelos ensinamentos que sempre foram repassados de maneira clara e direta. Tenho certeza que serão muito úteis na continuidade da minha jornada.

Levo agradecimentos também aos colegas de classe que vieram de todas as partes do Brasil e participaram de inesquecíveis momentos durante essa formação. Pelo apoio profissional, agradeço ao Instituto Federal do Norte de Minas Gerais e a todos os colegas que de alguma maneira auxiliaram na realização desta pesquisa, em especial aos da DGTI.

"*O que a vida quer da gente é coragem*"

Guimarães Rosa

Resumo

O Ensino Médio brasileiro vem, ao longo dos anos, passando por constantes debates acerca dos seus problemas de acesso e permanência, qualidade do ensino e até mesmo da sua identidade. O crescimento da oferta da educação profissional integrada ao ensino médio protagonizada pelos Institutos Federais (IFs), criados em 2008, vem trazendo resultados interessantes diante dos grandes investimentos do Governo Federal. Dessa forma, novos mecanismos que subsidiem gestores no processo de tomada de decisão e na avaliação do binômio "oferta-qualidade" dessas instituições tornam-se cada vez mais necessários. Esta dissertação, considerando o papel avaliativo do Exame Nacional do Ensino Médio (ENEM), apresenta uma solução de mineração de dados em um processo de *Knowledge Discovery in Databases* (KDD) para predição e estimação do desempenho dos alunos do Ensino Médio dos IFs. Para a extração do conhecimento, foi utilizado o método baseado em etapas *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) aliado às ideias do *framework Domain-Driven Data Mining* (D^3M), visando à produção de resultados mais amigáveis aos especialistas do domínio. As bases de dados do ENEM e as do Censo escolar foram integradas para a formação de um *data-mart* apresentado no grão aluno. Após a interpretação e modelagem do problema, os dados foram preparados para diferentes técnicas de Inteligência Artificial; inserindo, modificando, preenchendo e excluindo variáveis através de informações de contexto. A etapa de transformação contou ainda com um procedimento supervisionado de redução de dimensionalidade que considerou a taxa de valores ausentes, variância e a correlação entre as variáveis independentes. Na construção dos modelos, a técnica de regressão logística produziu índices de propensão de sucesso dos alunos e atingiu resultados superiores a 0,84 e 0,51 para as métricas AUC_ROC e KS2_MAX, respectivamente. Para a extração do conhecimento em linguagem natural, árvores de decisão construíram condições sequenciais e regras foram geradas por meio de indução baseada em escores. Essas técnicas foram avaliadas quanto às métricas: confiança, suporte e *lift.* Ao final, concluiu-se que a abordagem apresentada (*Domain-Driven Data Mining*) teve um ótimo resultado na modelagem e na validação de políticas públicas.

Palavras-chave: KDD. Data-Mining. Regressão Logística. Árvore de Decisão. Desempenho Escolar. Institutos Federais de Educação.

Abstract

Throughout the years, the Brazilian Secondary School has gone through constant debates about its problems of access and permanence, quality of teaching and even of its identity. The growth in the offer of vocational education integrated to secondary schools, starred by the Federal Institutes (IFs), created in 2008, has brought interesting results in view of the large investments of the Federal Government. Thus, new mechanisms that subsidize managers in the decision-making process and in the evaluation of the "supply-quality" binomial of these institutions become increasingly necessary. This dissertation, considering the evaluative role of Secondary School Student Test (ENEM), presents a data mining solution in a Knowledge Discovery in Databases (KDD) process for predicting and estimating the performance of secondary school students of IFs. For the extraction of knowledge, the CRoss-Industry Standard Process for Data Mining (CRISP-DM) method was used associated with the ideas of the Domain-Driven Data Mining (D^3M) framework, in order to produce friendly results to domain experts. The ENEM and official school census databases were integrated into data-mart presented in student grain. After the interpretation and modeling of the problem, the data were prepared for different techniques of Artificial Intelligence; inserting, modifying, populating, and deleting variables through context information. The transformation stage also had a supervised procedure of dimensionality reduction that considered the rate of missing values, variance and the correlation between the independent variables. In the construction of the models, the logistic regression technique produced a propensity score for success of students and had your results higher than 0.84 and 0.51 for the metrics AUC_ROC and KS2_MAX, respectively. For the extraction of knowledge in natural language, decision trees constructed sequential conditions and rules were generated through induction based on scores. These techniques were evaluated for the metrics: confidence, support and lift. In the end, it was concluded that the approach presented (Domain-Driven Data Mining) had an excellent result in the modeling and the validation of public policies.

Keywords: KDD. Data-Mining. Logistic Regresion. Decision Trees. School Performance. Federal Institutes of Education.

Lista de Figuras

Lista de Tabelas

Lista de Equações

Lista de Abreviaturas

AKD – *Actionable Knowledge Discovery*

AUC_ROC – *Area under curve ROC*

D³M – *Domain-Driven Data Mining*

DDID-PD - *Domain-Driven In-Depth Pattern Discover*

ENEM – Exame Nacional do Ensino Medio

IF – Instituto Federal de Educação, Ciência e Tecnologia

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KS2 – Kolmogorov Sminorv (classificação binária)

MDE – Mineração de Dados Educacionais

MEC – Ministério da Educação

OLAP – *Online Analytical Processing*

PNE – Plano Nacional de Ensino

RFEPCT – Rede Federal de Educação Profissional, Científica e Tecnológica

UFPE – Universidade Federal de Pernambuco

UFRJ – Universidade Federal do Rio de Janeiro

Sumário

1 INTRODUÇÃO

No sistema educacional brasileiro é no ensino médio onde persistem os debates mais controversos, seja pelos problemas do acesso e da permanência, seja pela qualidade da educação oferecida, ou, ainda pela discussão sobre a sua identidade. Apesar de a premissa da obrigatoriedade do ensino médio brasileiro ter sido colocada recentemente, através da emenda constitucional N° 59/2009, que amplia o ensino básico para a faixa dos 6 aos 17 anos, o ensino médio brasileiro já vinha se expandindo de maneira significativa desde a década de 90, visando não somente as aspirações das camadas populares por mais escolarização, mas também a necessidade de tornar o país mais competitivo no cenário internacional (KRAWCZYK, 2009).

O Plano Nacional de Ensino (PNE) 2014-2024, lei n 13.005/2014, institui métricas, metas, estratégias e diretrizes para a educação brasileira para o período de 10 anos. O Plano estabeleceu em uma das suas metas, sem sucesso, universalizar, até 2016, o atendimento escolar para toda a população de quinze a dezessete anos e elevar, até 2024, a taxa líquida de matrículas no ensino médio para 85% (oitenta e cinco por cento) do seu total. A expansão das matrículas gratuitas de Ensino Médio integrado à Educação Profissional é uma das ações estratégicas vinculadas a essa meta. O plano estabelece ainda oferecer, no mínimo, 25% (vinte e cinco por cento) das matrículas de educação de jovens e adultos, nos ensinos fundamental e médio, na forma integrada à educação profissional (BRASIL, 2014).

A oferta do ensino profissional integrado ao ensino médio, objetivando além da formação técnica o acesso ao ensino superior, minimiza a dualidade histórica na discussão do ensino profissional no Brasil, que até então era direcionado aos filhos dos trabalhadores, enquanto o ensino voltado para o ingresso nas universidades, era ofertado para a elite (HELENA; CASTRO, 2003; MARTINS; PAULA, 2012).

Em 2008, a Lei n° 11.892/2008 criou os Institutos Federais de Educação, Ciência e Tecnologia (IFs), e os classificou, através do seu art. 2° "[...], como instituições de educação superior, básica e profissional, pluricurriculares e *multicampi,* especializados na oferta de educação profissional e tecnológica nas diferentes modalidades de ensino [...]" (BRASIL, 2008). A Lei estabelece ainda que 50% (cinquenta por cento) das vagas ofertadas deveriam ser para a educação tecnológica de nível médio, priorizando as ofertas de cursos integrados para os concluintes do ensino fundamental. Essa oferta de ensino constitui-se da articulação entre o ensino médio e formação profissional quando ofertada em um único currículo e em uma única matrícula. Esse tipo

de oferta havia sido extinta através do decreto n° 2.208/1997 (BRASIL, 1997) e só voltou a ser regulamentada em 2004, com o decreto n° 5.154/2004 (BRASIL, 2004). Isso demonstra uma contínua discussão acerca dos princípios e identidade da educação profissional ao longo dos anos, bem como o protagonismo dos Institutos Federais na potencialização deste tipo de oferta, destacada por (PACHECO, 2011) como uma revolução na educação profissional e tecnológica do Brasil.

De modo geral, diversas são as decisões que vêm sendo tomadas ao longo dos anos acerca do ensino médio brasileiro, cuja tônica envolve desde novos investimentos a mudanças das diretrizes curriculares (KUENZER, 2000; BERNARDIM; SILVA, 2014). Nesse sentido, novos mecanismos para avaliar todo esse processo vêm se mostrando cada vez mais necessários, de maneira que possam subsidiar os gestores, educadores e especialistas no processo de tomada de decisão (ARAÚJO; LUZIO, 2005).

1.1 Motivação e Justificativa

No Brasil, o Exame Nacional do Ensino Médio (ENEM) é utilizado para avaliar o desempenho do estudante ao final da educação básica. A partir de 2009, o exame passou a ser um dos principais mecanismos de seleção para ingresso no ensino superior (INEP, 2011). A partir dessa mudança, a prova é dividida em quatro grandes áreas, constrói-se uma Matriz de Referência mais complexa e modifica-se a metodologia de análise de desempenho – utilizando-se da Teoria da Resposta ao Item. Tal metodologia possibilita a comparação longitudinal dos resultados em diferentes anos, bem como o monitoramento do ensino médio no país (GONÇALVES JR; BARROSO, 2013).

Com todas essas mudanças o ENEM passa a ser um relevante objeto de estudo, pois além dos dados de conhecimento técnico do aluno, ele captura também dados socioeconômico-culturais, informações que possibilitam ao governo federal definir e validar políticas públicas para a educação nacional. Nos estudos acadêmicos, há ainda pesquisas acerca da função do ENEM como instrumento de avaliação do ensino médio e de inserção nas universidades brasileiras (VIANNA, 2003), além da sua influência nos currículos da educação básica (SOUSA, 2003).

Muito utilizado no enriquecimento das análises produzidas com a base de dados do ENEM (VIGGIANO; MATTOS, 2013; ALMEIDA FILHO, 2014), o Censo Escolar coleta dados anualmente sobre estabelecimentos de ensino, turmas, alunos, profissionais escolares em sala de aula, movimento e rendimento escolar. Essas informações são utilizadas para traçar um panorama

nacional da educação básica e, assim como o ENEM, servem de referência para a formulação de políticas públicas e programas na área da educação.

No ENEM 2014, o desempenho dos alunos do terceiro ano do ensino médio das escolas federais foi superior ao de outras redes de ensino, inclusive, das escolas particulares., conforme ilustrado no gráfico da Figura 1. As médias da rede federal foram maiores em todas as áreas de conhecimento, a saber: Linguagens, códigos e suas tecnologias (LC), Matemática e suas tecnologias (MT), Ciência da Natureza e suas tecnologias (CN), Ciências Humanas e suas Tecnologias (CH), além da redação. Analisando o *ranking* do Enem 2014 por Escola, 79 das 100 melhores médias das escolas públicas são federais. O *ranking* é divulgado pelo INEP com o intuito de auxiliar pais, professores, diretores de escolas e gestores educacionais nas reflexões sobre o aprendizado dos estudantes no Ensino Médio. Para isso, o INEP utiliza o desempenho dos alunos matriculados na 3° série do ensino médio, apresentando as proficiências médias por unidade escolar para cada uma das áreas de conhecimento e para a redação (INEP, 2011).

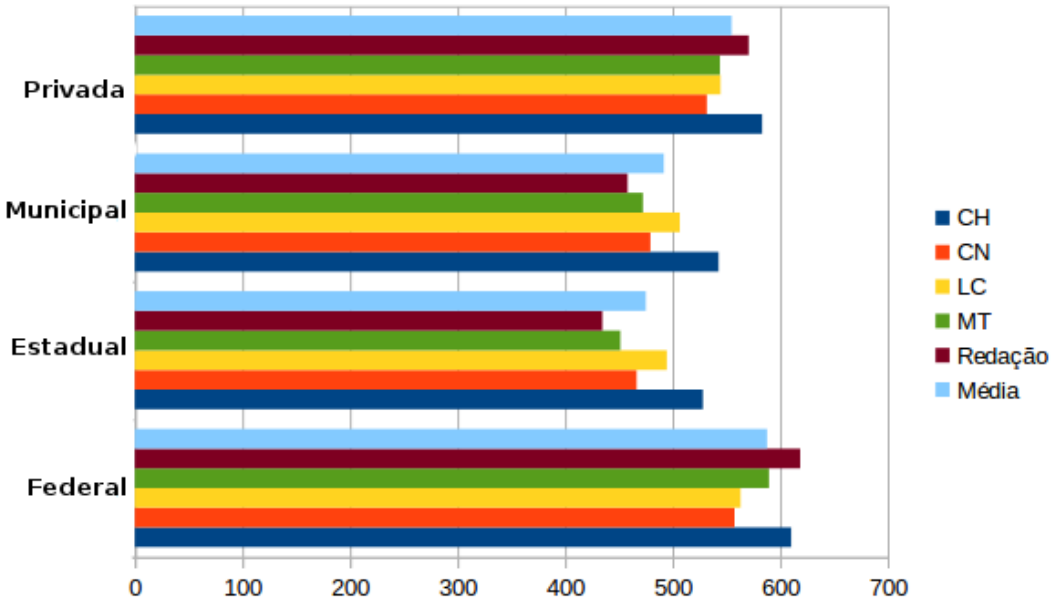

Figura 1: Gráfico: desempenho escolas federais - ENEM

A Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT), ligada ao Ministério da Educação (MEC), é atualmente composta pelos Centros Federais de Educação Tecnológica (CEFETs), Escolas Técnicas Vinculadas às Universidades Federais, Universidade Tecnológica, Colégio Pedro II e os IFs. Através dos IFs, a RFEPCT, gráfico ilustrado na Figura 2, está vivenciando a maior expansão da sua história. A rede passou de 140 *campi* em 2003, para 644

instalados em mais de 568 cidades em todos os estados brasileiros (BRASIL, 2016). Em 2013, o número de matrículas já chegou a um total de quase um milhão de alunos (MEC, 2013).

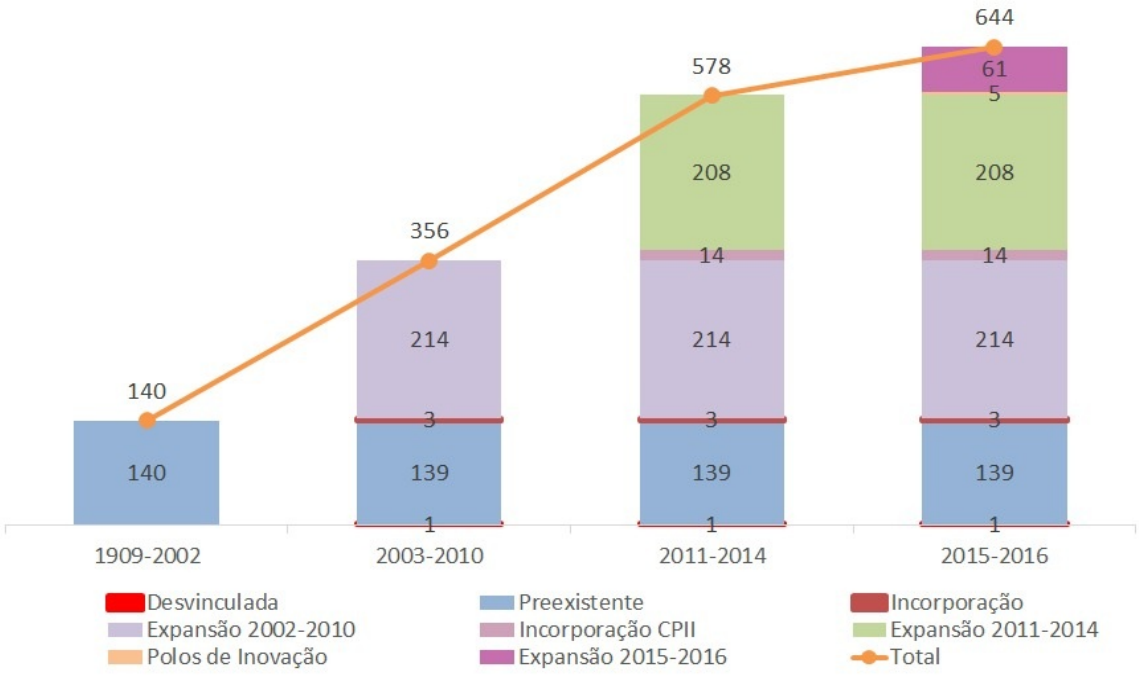

Figura 2: Gráfico - Expansão da RFEPCT (*BRASIL, 2016)*

Apesar de os IFs possuírem uma mesma lei de criação e uma semelhante estrutura organizacional, esses órgãos são autarquias, e como tais, possuem autonomia administrativa e pedagógica. Neste cenário, é evidente a heterogeneidade acerca do direcionamento dos investimentos no que se refere a capacitação de professores, grade curricular, estruturas físicas, dentre outros aspectos que, segundo (TRAVITSKI, 2013) podem interferir no desempenho dos alunos. (ALMEIDA FILHO, 2014) traz ainda a seguinte reflexão: "até que ponto existe a avaliação da gestão dos IFs no que diz respeito ao binômio oferta-qualidade do ensino, haja vista a quantidade de vagas criadas na última década?"

Emergente na tecnologia da informação, a Mineração de Dados (MD) é uma das áreas mais ativas na última década, impulsionando um grande volume de trabalhos acadêmicos e da própria indústria nos mais diversos domínios de aplicação (CAO, 2009).

A MD é uma maneira de analisar um conjunto de dados buscando identificar regras, padrões e desvios. (J.HAN, J.PEI, M.KAMBER, 2012) definem a MD como um processo interdisciplinar de extração de informação e conhecimento a partir de grandes volumes de dados. (DEOGUN et al., 1997) ressaltam que essa interdisciplinaridade envolve, principalmente, as áreas de Estatística,

Banco de Dados (BD) e Inteligência Artificial (IA), que juntas, podem facilmente formular e resolver problemas de predição e diagnóstico.

(ADEODATO, 2016) utilizou as bases de dados disponibilizadas pelo INEP (ENEM e Censo Escolar) para a extração de informações sistemáticas capazes de auxiliar na construção de soluções de apoio à tomada de decisão, utilizando-se de técnicas de inteligência artificial, estatística e banco de dados. Esse estudo consiste na instanciação de métodos de MD ao domínio da educação, chamada por alguns autores (ROMERO; VENTURA, 2013) (MOHAMAD; TASIR, 2013) de Mineração de Dados Educacionais (MDE).

A Mineração de Dados é uma etapa do processo de Descoberta de Conhecimento (*Knowledge Discovery in Databases – KDD)* que consiste na aplicação de técnicas de análise de dados e algoritmos de descoberta em busca de padrões ou modelos (FAYYAD et al., 1996). Apesar de ser a MD a principal fase do processo de KDD, outras etapas devem ser percorridas para que o conhecimento extraído seja potencialmente útil e compreensível aos especialistas de domínio.

Para (KIANG, 2003), sistemas de classificação, utilizando MD, possuem um importante papel no desenvolvimento de soluções que visam a dar suporte ao processo de tomada de decisão.

Portanto, o presente trabalho versa sobre a aplicação de métodos e técnicas de mineração de dados por meio de um processo de descoberta de conhecimento nas bases de dados abertas do Censo Escolar e ENEM, provenientes do ano de 2014. Considera-se como principal ideia desta pesquisa, a concepção de um modelo capaz de predizer o desempenho dos alunos dos Institutos Federais, identificando os principais pontos para o sucesso ou insucesso desses indivíduos, de maneira que auxiliem gestores no processo de tomada de decisão.

1.2 Objetivos

Essa pesquisa visa a desenvolver um modelo, baseado na mineração de dados, que possa avaliar e prever o desempenho dos alunos que se encontram no último ano do ensino médio dos Institutos Federais de Educação.

1.3 Objetivos específicos

- Estimar as chances de alunos dos Institutos Federais de Educação terem sucesso no ENEM
- Justificar o porquê da estimativa de sucesso

- Identificar os principais fatores que influenciam para o desempenho dos alunos nos Institutos Federais de Educação.
- Gerar um *data-mart* para serem feitas consultas OLAP e modelos de classificação para suporte à decisão sobre os dados de 2014.
- Gerar uma base de regras e uma estrutura de conhecimento para permitir a interpretação dos aspectos mais relevantes da qualidade dos alunos dos IFs em 2014.
- Criar um fluxo de processamento automatizado para se poder replicar o modelo para outros anos a partir das novas edições do ENEM e do Censo Escolar.

1.4 Estrutura do trabalho

Este trabalho apresenta, no capítulo 2, a fundamentação teórica, onde são discutidas abordagens já realizadas acerca da avaliação da educação tecnológica. Além disso, são analisadas pesquisas da área de MD, destacando a sua aplicação aos problemas de educação.

No Capítulo 3, são discutidas a metodologia, técnicas e ferramentas utilizadas no desenvolvimento do trabalho.

No Capítulo 4, são apresentadas as etapas iniciais do CRISP-DM com a aplicação de estatística descritiva para o entendimento dos dados, além das fases de seleção, tratamento e, transformações realizadas sobre o conjunto de dados disponibilizados para esta pesquisa.

Os resultados obtidos pela extração do conhecimento são apresentados no Capítulo 5, assim como as discussões acerca das avaliações e interpretação dos modelos.

Finalmente, no último capítulo dessa dissertação, o Capítulo 6, são apresentadas as conclusões, as limitações e as ideias para trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Devido ao recente crescimento da Rede Federal de Educação Profissional e Tecnológica, muitos estudos de reflexão acerca da criação dos Institutos Federais, de suas características e das suas fundamentações político-pedagógicas são encontrados na literatura. Porém, para estas instituições, ainda são escassas pesquisas com foco avaliativo. Os trabalhos existentes, com tal ênfase, a maioria, é apresentado por meio de abordagens qualitativas. Para os poucos encontrados com viés quantitativo, as pesquisas se limitam em explorar fontes de dados utilizando técnicas de estatística e identificando pontos de interesse na avaliação do desempenho das instituições/alunos. Todavia, o principal objetivo deste estudo, vai além, e apresentará uma solução de mineração de dados capaz de prever o sucesso/insucesso dos alunos dos IFs.

No tocante aos estudos que se utilizam das bases públicas da educação brasileira para aplicação de técnicas de mineração de dados em busca de conhecimento e padrões, não se encontrou na literatura nenhuma abordagem com foco na avaliação da educação básica técnica e tecnológica. Dessa forma, cabe analisar também como trabalhos relacionados a esta pesquisa, estudos que utilizam de técnicas de inteligência artificial em diferentes ambientes educacionais, destacando as suas diversas abordagens e objetivos.

Dessa forma, este capítulo discorrerá sobre os estudos nacionais de extração de conhecimento a partir das bases de dados públicas da educação brasileira, estudos específicos do Ensino Básico Técnico e Tecnológico (EBTT), além de estudos nacionais e internacionais de MD genéricos e aplicados à educação (MDE).

2.1 Estudos a partir das bases de dados públicas da educação brasileira

Contribuindo com o caráter diagnóstico do ENEM, (VIGGIANO; MATTOS, 2013) investiga e compara o desempenho dos estudantes de cada uma das regiões geográficas do Brasil. O estudo utiliza dos microdados disponibilizados pelo INEP para o ano de 2010 e classifica as regiões do país em 3 grupos de desempenho: superior (Sul e Sudeste), médio (Centro-Oeste); e inferior (Norte e Nordeste). Foram analisados dados de 847.621 estudantes em diferentes contextos, utilizando-se do método de estatística descritiva. Gráficos e tabelas foram os principais artefatos

utilizados para a apresentação das informações. Fazendo uso das mesmas ferramentas, (PIRES, 2015) dividiu os alunos que realizaram o ENEM no ano de 2012 no estado de São Paulo em dois grandes grupos e depois em outros dois subgrupos, utilizou dos dados socioeconômicos – escolaridade dos pais e renda mensal familiar – como critério de seleção. Logo, por meio da análise descritiva, explorou diversas variáveis que pudessem vislumbrar as reais condições presentes, passadas e futuras dos participantes. O estudo sugeriu que o crescimento da renda potencializa as diferenças entre os grupos, e identificou também que todas as regiões do país apresentam baixo desempenho, inferiores a 58%.

Apresentando técnicas mais avançadas de exploração das bases de dados do MEC, (AMENDOEIRA et al., 2013) expuseram o potencial e as possibilidades do uso da visão multidimensional e de técnicas de *datawarehouse* (DW) em um processo de KDD, propiciando agilidade e facilidade na obtenção de indicadores de qualidade. O estudo apresentou e discutiu alguns indicadores relacionados ao ensino da Língua Portuguesa, extraídos pelas técnicas apresentadas.

Indo um pouco mais além, (FONSECA; NAMEN, 2016) aplicou inferências estatísticas e técnicas de MD, com o intuito de identificar fatores que relacionam o perfil dos professores com a proficiência obtida pelos seus alunos na Prova Brasil – Sistema de Avaliação da Educação Básica (SAEB). Os professores foram separados em grupos para a avaliação da influência positiva e negativa no desempenho dos seus alunos. Os autores dividiram os professores que lecionam Matemática em dois grupos: “Até 65%” e “Maior que 65%”. No primeiro, estavam os que apresentaram até 65% dos seus alunos com o desempenho acima da média, já no segundo, os que apresentaram mais de 65% dos alunos. De maneira análoga, para análise da influência negativa, os professores foram separados em duas classes, uma com “Até 35%” dos alunos com desempenho acima da média e outra com mais de 35% dos alunos. Apesar do cuidado em se manter uma certa simetria de quantidade de indivíduos na divisão dos grupos, e acreditando na capacidade do critério adotado para o levantamento de alguns indicadores de desempenho, o autor reconhece as possíveis limitações em suas escolhas.

Com os grupos formados, por meio do algoritmo *Naive bayes,* calculou-se a probabilidade que cada atributo implicou na classificação dos professores nas classes citadas. Em uma de suas conclusões, o trabalho identificou que a estabilidade do vínculo profissional do corpo docente é relevante para o desempenho positivo dos alunos. Além disso, avaliou que a desvalorização salarial da profissão do educador; os altos índices de absenteísmo dos alunos e a crença do professor de que poucos alunos entrarão na universidade tendem a influenciar negativamente o desempenho dos estudantes.

Apresentando uma proposta de aplicação de múltiplos modelos de regressão, (GUERRA et al., 2014)) realizaram um estudo capaz de prever a demanda potencial por vagas de ensino superior público nos municípios brasileiros. A pesquisa utilizou os dados do Censo IBGE 2010, Censo do Ensino Superior 2012, os Microdados do ENEM 2014 e aplicou a seguinte tarefa de predição via regressão: $((X_{0,}y0),(X_{1,}y1),...,(X_n,y_n))$, para as quais X_i é o vetor de variáveis independentes para o i-ésimo município e y_i é o número de alunos matriculados no ensino superior deste município. Considerando a função $f:X \rightarrow y$, que aproxima a função real e desconhecida f , foi calculado, para cada município, o resíduo de regressão $\hat{e}=\hat{y}-y$, isto é, a diferença entre o valor predito pelo modelo e o valor real de y . Diferença quer os autores consideram ser a demanda potencial de vagas no ensino público superior para cada município. Os algoritmos experimentados com resultados satisfatórios foram o modo linear, árvore de regressão e *random forest*. Os autores consideraram o valor de 0,74 para a métrica R^2 como um resultado de boa qualidade quando comparado às outras pesquisas da área de Economia da Educação. Todavia, as inferências produzidas são questionáveis por não conseguir tratar na geração dos modelos as particularidades de várias cidades e regiões. Para uma análise de demanda, deve-se considerar, por exemplo, a cooperação entre cidades inseridas em regiões metropolitanas. De toda forma, o trabalho é importante em ressaltar as possibilidades de contribuição na exploração de dados públicos por métodos de aprendizagem de máquina de modo que auxilie no aprimoramento da gestão pública no Brasil. Outra importante contribuição é a utilização de modelos de regressão para aplicações que a informação interessada está contida no erro da predição, e não somente na predição em si.

2.2 EBTT – Institutos Federais

Por meio das bases de dados do ENEM, Censo e IBGE, (ALMEIDA FILHO, 2014) analisou como a ampliação da oferta da educação básica nos IFs da região Nordeste vem afetando a qualidade do ensino no período de vigência do PNE 2001-2010. Embora o estudo explore as bases de dados públicas da educação, inclusive com a proposição de indicadores de gestão escolar, o trabalho é de cunho descritivo. Sua contribuição se deve principalmente na reflexão acerca de distorções entre instituições pertencentes a uma mesma Rede. Para tal, o autor estabeleceu um quadro comparativo com os dados que refletem várias dimensões, como infraestrutura, corpo docente, didático-pedagógica e desempenho no ENEM de 52 campi do Nordeste. O estudo também fez um resgaste histórico dos IFs no Brasil e apresentou pontos da conjuntura analisada para

melhoria da prestação do serviço educacional brasileiro, em especial, as escolas da rede federal da região nordeste.

Através de uma abordagem hermenêutica, (DORNELES, 2011) investigou indicadores que pudessem revelar e contribuir para o desenvolvimento da educação profissional de qualidade. O estudo teve como ponto de partida indicadores educacionais existentes e utilizou uma entrevista qualitativa. A pesquisa teve a participação de 2 docentes, 2 egressos e os reitores dos Institutos Federais do estado de Goiás, além de 5 diretores de campi localizados em Roraima, Tocantins, Rio Grande do Sul e São Paulo. Por meio de técnicas de análise de conteúdo, selecionou-se os principais temas abordados pelos entrevistados que pudessem construir, junto ao referencial teórico e a legislação, os indicadores específicos da educação profissional nos Institutos Federais. Também de maneira qualitativa, (LIMA et al., 2013) estudaram os resultados obtidos pelas escolas federais do estado do Espírito Santo em relação aos das escolas estaduais no ENEM de 2012. Os autores identificaram um maior desempenho das escolas federais e justificaram essa vantagem, por meio de uma análise documental, devido às ofertas nas escolas estaduais serem, quase sempre, na modalidade subsequente (não concomitante ao ensino médio) e associada a eixos tecnológicos de menor complexidade.

Através de análise descritiva e *scores* de propensão – utilizando o método do vizinho mais próximo - (JOSÉ; ARAÚJO, 2014) caracterizou a Rede de Educação Profissional e Tecnológica (EPT) analisando o desempenho e inserção produtiva dos alunos. De posse das bases de dados do Censo de 2007 a 2012 e da base de dados do ENEM do ano de 2009, o dividiu-se os indivíduos em dois grupos, aqueles que realizaram EPT e os que não realizaram. O autor destaca, na análise descritiva dos dados, o crescimento de 248% das matrículas no ensino integrado de 2012 em relação a 2007 e afirma ser devido ao aumento do investimento por parte do Governo Federal. Há também uma correlação positiva entre cursar EPT e possuir melhor desempenho escolar, além de maior inserção produtiva. Com isso, o estudo identifica que a EPT pode ser mais efetiva no desenvolvimento das habilidades cognitivas e socioemocionais.

2.3 Mineração de Dados

A MD quando aplicada a um contexto educacional, recentemente foi chamada de Mineração de dados Educacionais. Nos últimos anos, principalmente a partir de 2008, quando foi criada a primeira conferência internacional exclusiva para MDE: EDM2008 em Montreal, Canadá,

a quantidade de trabalhos acerca dessa área evoluiu de forma exponencial, juntamente aos periódicos, livros e demais eventos (SIEMENS; BAKER, 2012).

Existe uma série de trabalhos de revisões do estado da arte de MDE (BAKER; YACEF, 2009; ROMERO; VENTURA, 2010, 2013; PEÑA-AYALA, 2014). O trabalho mais citado do gênero, segundo o *Google Scholar*, (BAKER; YACEF, 2009), utilizou de obras produzidas no período de 1995 a 2005 destacando principalmente o surgimento da nova subárea, a diversidade de fontes de dados existentes e algumas expectativas de resultados quando da aplicação de MDE. Além desse, destaca-se também uma análise dos principais trabalhos de pesquisa publicados no Brasil na área de MDE (RODRIGUES et al., 2014) O artigo resgata trabalhos que foram produzidos desde o ano de 2006, e por meio da classificação dos artigos, os resultados dão conta de dimensões interessantes da pesquisa da área no país.

A predição da performance dos alunos é uma das aplicações de MDE mais antigas e também mais utilizadas. Os estudos, em sua maioria, visam a estimativa do desempenho dos alunos, bem como o entendimento da relação do valor estimado com aspectos contextuais e características dos próprios estudantes. Este não é um problema simples, vários são os fatores que podem influenciar no desempenho dos mesmos, como informações demográficas, culturais, sociais, familiares, socioeconômicas, psicológicas, histórico curricular, interações durante o curso etc. (ARAQUE et al., 2009).

Dentre as pesquisas na área de MDE, destacam-se, em número de publicações, as relativas aos ambientes de ensino a distância, devido à possibilidade de os resultados advindos da mineração de dados suprirem a ausência de interação e supervisão física direta com os estudantes. Outro fator corroborante é a valiosa coleção de dados disponíveis nos Ambientes Virtuais de Aprendizagem (AVA).

Aplicando KDD sob os dados demográficos dos alunos da faculdade de Economia e Ciências Sociais da Universidade de Mugla na Turquia no ano de 1995, (GURULER et al., 2010), exploram os fatores que impactam no sucesso dos estudantes. Os autores utilizam regras de classificação por árvore de decisão a fim de encontrar qual dado demográfico é mais influente no desempenho do estudante. A base de dados utilizada foi uma visão de 111 variáveis independentes oriundas de 13 tabelas pré-selecionadas, onde a média dos alunos foi utilizada como indicador de sucesso. No intuito de envolver todas as funcionalidades dos softwares *SQL Server e Analysis Services* acopladas à base de dados do sistema educacional, foi desenvolvida uma solução de estudo chamada *Mugla University Student Knowledge Discovery Unit Program* – MUSKUP. O software centraliza várias tarefas inerentes ao processo de KDD de maneira harmônica e transparente. O autor as utilizou para construir dois modelos, o primeiro com alunos com nota maior ou igual à

média (2,0) e outro com nota maior ou igual a 3.0; e os validou através de curvas *Lift.* Os valores encontrados foram de 1,74 para o primeiro modelo e 1,36 para o segundo, comprovando a capacidade de predição dos dois modelos, sendo que, o segundo, teve um menor desempenho devido ao menor número de casos positivos.

Utilizando os dados dos fóruns das redes sociais dos ambientes de aprendizagem da educação a distância, (ROMERO et al., 2013), desenvolveram um modelo de performance do estudante. O experimento contou, inicialmente, com dois conjuntos de dados. O primeiro com as mensagens dos fóruns relativos à metade da disciplina de Fundamentos de Ciências da Computação de 114 estudantes do curso de Ciências da Computação, e o segundo com as mensagens já do fim do curso. Esses dois conjuntos foram subdivididos em mais dois, em que os atributos de maior relevância foram separados por meio de um processo de seleção. O processo de seleção considerou os atributos que foram selecionados por pelo menos 5 dos 10 algoritmos de seleção de variáveis aplicados. Por fim, estes 4 subconjuntos foram novamente divididos em dois, considerando as mensagens que não tinham relação com o assunto da disciplina, formando então 8 subconjuntos de dados. Os dados foram categorizados em 3 categorias, a saber: Quantitativos (número de mensagens lidas/postadas, tempo gasto, etc), Qualitativos (avaliação das postagens) e Sociais (ponderação quanto a proporção de respostas dadas e de respostas recebidas). Como situação final dos alunos, foram considerados de forma binária os valores “aprovado” ou “reprovado”.

Para fase de mineração, vários algoritmos de classificação tradicional e *clustering* foram testados a fim de responder às seguintes questões: a) Quais técnicas de DM são melhores para prever performance de estudantes a partir da participação em fóruns? Apesar de algoritmos de classificação supervisionados já serem amplamente utilizados para estas tarefas, algoritmos não-supervisionados, como o *cluster*, poderiam trazer resultados interessantes. b) Quais atributos são melhores para a predição? c) Quais mensagens são melhores? d) É possível uma predição precoce?

Após os experimentos por meio do método 10 *cross fold-validation,* para a abordagem de classificação tradicional, técnicas chamadas de “caixa-preta”, conforme esperado, obtiveram, de maneira geral, maior desempenho, embora não significantes ao ponto de descartar as abordagens de "caixa branca", que possuem uma melhor interpretabilidade. Já na abordagem via *clustering*, somente um algoritmo obteve desempenho semelhante, se comparado à classificação tradicional. Para estes modelos, que também oferecem uma boa interpretabilidade, foi construído um modelo adicional de regras de associação para cada *cluster*, a fim de se identificar as regras mais representativas. Essa medida visou dar ainda mais interpretabilidade através de regras SE-ENTÃO às informações dos centroides*,* além de construir regras mais específicas para cada *cluster.*

Para entender quais atributos são melhores preditores, o autor mediu o comportamento dos algoritmos em pares por meio do teste estatístico, *Wilconxon signed rank test* (REY; NEUHÄUSER, 2011). O teste permitiu identificar que os subconjunto de dados com variáveis selecionadas, quase sempre obtiveram melhores métricas. Outro fator colocado pelos autores que corrobora com o uso da seleção de atributos é o aumento da compreensibilidade do modelo. O mesmo teste estatístico foi utilizado para identificar que o subconjunto de mensagens relacionadas ao curso obteve melhores desempenhos do que quando todas as mensagens disponíveis eram utilizadas. Para analisar a possibilidade de predição precoce, ainda com o teste *Wilconxon,* os autores identificaram que em nenhum dos experimentos o grupo de mensagens colhidas no meio do curso obteve métricas melhores. Contudo, comparando os índices de acurácia (70 e 80% para os dados do meio do curso, e 80 e 90% para os dados do fim do curso) aos de outros trabalhos correlatos, o estudo afirma a viabilidade de seus modelos para ambos os períodos. Porém, vale destacar que nenhum teste estatístico foi utilizado pelos autores para a comparação dos resultados desses trabalhos.

A compreensibilidade dos modelos gerados a partir da Mineração de Dados é uma preocupação constante em MDE. Dessa forma, com o objetivo de ressaltar a dificuldade na construção de um bom modelo de predição, Xing e colegas, em 2015, propôs um modelo utilizando os dados de um ambiente colaborativo e virtual de resolução de problemas matemáticos da ferramenta Geogebra (VMTwG). Foi proposto um modelo com boa qualidade de predição e ao mesmo tempo de fácil interpretação, contextualização e implementação. Os autores utilizaram técnicas de Programação Genética, abordagens de análise de aprendizado e teorias de contexto (XING et al., 2015).

No Brasil, destacam-se os trabalhos de (BAKER et al., 2011), em que foi apresentada uma revisão das pesquisas realizadas na área de MDE. O estudo enfatiza nos principais métodos e aplicações que vêm influenciando, de maneira satisfatória, a pesquisa e a prática da educação em vários países, além de discutir as condições que viabilizam este campo de pesquisa no Brasil. Dentre os trabalhos mais citados, estão os que envolvem o contexto de educação a distância, como (KAMPFF, 2009), que implementou em um ambiente virtual de aprendizagem uma arquitetura de geração de alertas para auxílio de professores na interação e no acompanhamento dos alunos. Os alertas são gerados a partir da classificação dos estudantes, considerando suas características e comportamento dentro do ambiente virtual.

Para o processo de MD foram considerados dados de alunos de turmas passadas para a extração das regras, e de alunos acompanhados para a geração dos alertas. O experimento permitiu comprovar que as intervenções realizadas pelo professor, a partir dos alertas, direcionadas a grupos que compartilham necessidades específicas, contribuíram para a melhoria dos índices de

desempenho dos mesmos. Também com uma abordagem preditiva, (MANHÃES, 2015) conseguiu com uma acurácia de 75% a 80%, identificar o risco de evasão dos alunos de graduação do curso de Engenharia Civil da Universidade Federal do Rio de Janeiro – UFRJ. O experimento, através de uma grande quantidade de testes, utilizou 10 algoritmos em três soluções diferentes de estratificação de base de dados trazidas pela ferramenta *WEKA*[1], *10 folds cross-validation, Train/Test Percentage Split (data randomized) e Supplied test set.*

Na Universidade Federal de Pernambuco (UFPE), (BARROS; ADEODATO, 2012), apresentaram uma avaliação sistemática do problema de retenção e evasão universitária na mesma universidade. O trabalho utiliza os dados acadêmicos e socioeconômicos dos alunos de vários cursos no período de 1998 a 2008 e, seguindo o processo de extração de conhecimento *Cross-Industry Standart Process for Data Mining* (CRISP – DM), desenvolve uma solução de mineração de dados para identificar e estimar o risco de evasão ou retenção ainda no início do curso. Foram utilizadas técnicas de regressão logística e redes neurais artificiais que produziram resultados com alto desempenho, segundo as métricas KS2_MAX e AUC_ROC, além de indução de regras para a construção de um modelo em linguagem natural. O estudo também apresenta uma análise do ponto de vista de custos das perdas versus investimentos na prevenção da evasão escolar com o modelo preditivo apresentado.

Quanto aos trabalhos que se assemelham à pesquisa proposta no que tange à fonte de dados, destaca-se (ADEODATO, 2016). Com uma abordagem também preditiva, utiliza as bases de dados disponibilizadas pelo INEP (ENEM e Censo Escolar) para produzir, através de regressão logística, um classificador capaz de gerar uma pontuação de propensão ao sucesso das escolas privadas brasileiras, além de identificar e quantificar os principais fatores que influenciam nesses resultados. A metodologia utilizada para a extração do conhecimento foi a CRISP – DM, visando à futura implantação de um Sistema de Suporte à Decisão para operação e navegação em tempo real. Através de árvores de decisão e de indução de regras, os autores geraram modelos mais compreensíveis, a fim de explicitar como o especialista humano decidiria de forma sequencial utilizando-se de regras. A qualidade da pontuação de propensão ao sucesso das escolas foi validada pelas métricas AUC_ROC e Max_KS2, obtendo os valores 0,897 e 0,632, respectivamente.

Pesquisas na área de educação discutem sobre a utilização do ENEM como indicador de performance escolar. (TRAVITSKI, 2013) faz um amplo estudo sobre as questões sociopolíticas da escolarização com métodos quantitativos, além de investigar os efeitos da utilização do ENEM para avaliação escolar. No estudo é destacado que o ENEM não é um indicador adequado de qualidade escolar para fins de responsabilização, sendo necessário a multiplicidade de indicadores, visto a

1 – Sítio do Weka na Ineternet disponível em: http://www.cs.waikato.ac.nz/~ml/weka/

pluralidade de fontes de informações que refletem no desempenho dos estudantes. Porém, tendo em vista a complexidade do problema da qualidade escolar, a sua importância e a falta de um mecanismo mais completo que avalie as escolas de maneira mais justa e multidisciplinar, o autor diz ser preciso tomar o *ranking* do ENEM como ponto de partida. E o coloca como um importante instrumento no que se refere a definição de políticas públicas com vistas à transparência de informação e aprimoramento da qualidade escolar.

3 METODOLOGIA, TÉCNICAS E FERRAMENTAS UTILIZADAS

Neste capítulo são abordadas as fases do processo de descoberta de conhecimento, incluindo as etapas de mineração de dados, concepção e avaliação do modelo preditivo.

3.1 Introdução

Na área de Mineração de Dados, várias são as técnicas que podem ser utilizadas para extração de conhecimentos em grandes bases de dados. Selecionar e utilizar as melhores técnicas para o domínio da pesquisa de maneira adequada é crucial para a obtenção de resultados satisfatórios.

Para guiar todo o processo, a utilização de uma metodologia faz-se necessária, pois sistematiza o projeto em fases, guiando a implementação das etapas necessárias para a extração do conhecimento (KDD), além de destacar os principais objetivos e preocupações de cada uma delas.

Este trabalho utilizou a metodologia CRISP-DM, que envolve todas as etapas destacas por (FAYYAD et al., 1996), que são: seleção, pré-processamento, transformação, mineração de dados, interpretação e avaliação. Contudo, outra metodologia, também inserida durante o processo de KDD, a *Data Drive Domain* (D3M), possibilitou embutir conhecimento de domínio durante o processo de descoberta de conhecimento, por meio de revisão de literatura e da experiência do autor no âmbito dos Institutos Federais.

3.2 CRISP-DM

O modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) é um processo hierárquico e iterativo proposto no ano 2000 por um consórcio de três grandes empresas de *data mining* e *data warehouse* que permite o controle de todo o ciclo de vida de um projeto de mineração de dados. O modelo CRISP-DM, Figura 3, é composto por fases, cada qual composta por tarefas (CHAPMAN et al., 2000).

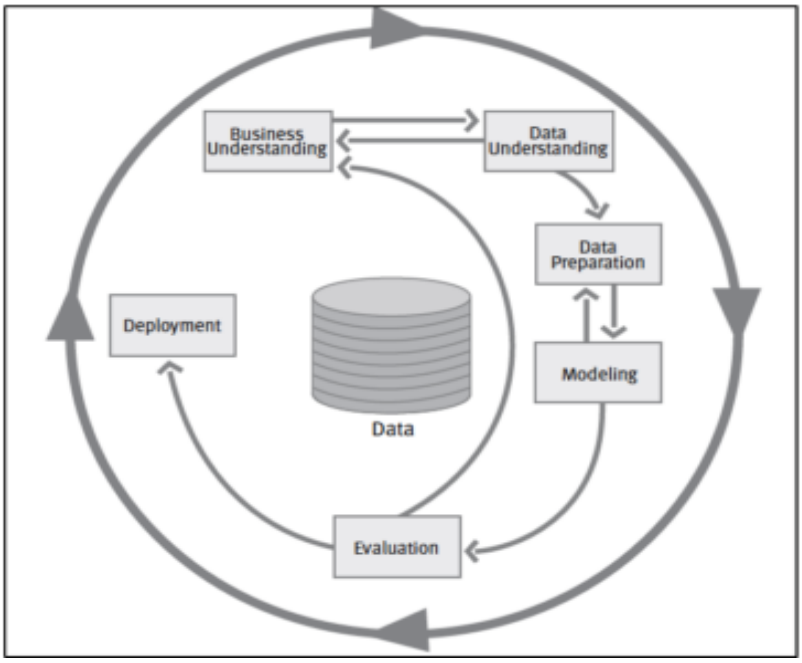

Figura 3: Fases do modelo CRISP-DM

As 6 fases do modelo são cíclicas e interativas:

1. Entendimento do negócio (*Business understanding)* – Fase inicial do processo, em que os objetivos e requisitos do projeto a partir de sua perspectiva de negócio são convertidos em um problema de mineração de dados.

2. Entendimento dos dados (*Data understanding*) – Análise dos dados iniciais buscando um conhecimento maior dos mesmos. Nesta fase podem ser identificados problemas na qualidade dos dados e subconjuntos com anomalias que podem ser utilizados para formular hipóteses sobre informações escondidas.

3. Preparação dos dados (*Data preparation*) – A fase de preparação dos dados consiste nas ações necessárias para se construir o conjunto final de dados, chamado de *input*. É nesta fase que ocorre todas as seleções de variáveis, criação de novas variáveis, transformações e limpezas dos dados.

4. Modelagem (*Modeling*) – Fase em que os modelos inteligentes são selecionados e aplicados para a visão de dados final. Nesta fase os parâmetros dos modelos são

ajustados para a obtenção de melhores resultados. Nessa fase podem ocorrer retornos para a fase de preparação dos dados.

5. Avaliação (*Evaluation*) – Nesta fase, os modelos construídos devem ser revisados e avaliados quanto ao atendimento dos objetivos do negócio. Ao final desta fase, deve chegar a conclusão se deve ou não utilizar os resultados da mineração de dados no negócio.

6. Implantação (Deployment) – Nesta fase, que ocorre nos casos onde a criação do modelo não é o fim do projeto, o conhecimento obtido precisa ser organizado e apresentado de uma forma que o usuário possa usar.

Ressalta-se que a fase 6, fase de implantação, não será aplicada, pois foge do escopo desse projeto, que se limita apenas à construção de um modelo preditivo, que possa identificar o sucesso e ou insucesso dos alunos dos institutos federais de educação no exame nacional do Ensino Médio. No entanto, a construção desse modelo visa a dar condições para uma fácil implantação do conhecimento extraído.

3.3 Domain Drive Data Mining – D³M

A inferência de conhecimento por especialistas de domínio durante o processo de extração de conhecimento, permitindo a entrega de resultados amigáveis à sociedade, chamado por (CAO, 2009) por *Actionable Knowledge Discovery* (AKD), vem se tornando cada vez mais necessária e recorrente. Apesar dos grandes desafios, a AKD é vista como uma alternativa às metodologias voltadas aos dados, que não se concentram nas reais necessidades do domínio da aplicação. Para Cao, os pesquisadores estão muitas vezes interessados em otimizar técnicas e padrões, e não em resolver problemas reais. Fato que contribui consideravelmente para a diferença existente entre os pensamentos acadêmicos e a expectativa do domínio.

(WANG; WANG, 2009), ressaltam ainda que, apesar de inúmeras pesquisas presentes na literatura conceberem técnicas eficientes de mineração de dados, além de novos métodos e algoritmos, elas muitas vezes focam a atenção apenas nos problemas técnicos e ignoram problemas básicos, como: o que é MD? Qual o produto de um processo de MD? O que é feito durante um processo de MD? Qual a relação entre o especialista do domínio e o minerador de dados? Nesse

sentido, a fim de nortear a metodologia dirigida aos dados, (CAO et al., 2005) propuseram uma abordagem prática, por meio do *framework Domain-Driven In-Depth Pattern Discover*-DDID-PD, que envolve praticamente as mesmas fases do conhecido CRISP-DM, porém com três grandes diferenças, a saber: i) os resultados e a modelagem estão envolvidas na natureza cíclica do modelo, ii) as fases comuns ao CRISP-DM são enriquecidas com a interação com especialistas do domínio, iii) as diferenças nas fases do ciclo do modelo são responsáveis por alcançar os objetivos do mundo real.

3.4 Mineração de Dados

Com o crescimento nas últimas décadas dos sistemas de informação e das suas volumosas bases de dados que são armazenadas em hardware cada vez mais baratos, a Mineração de Dados é considerada uma das tecnologias mais promissoras da atualidade. Para (J.HAN, J.PEI, M.KAMBER, 2012) MD pode ser vista como um resultado natural da evolução da tecnologia da informação.

As tarefas de MD num processo de extração de conhecimento podem ser divididas em abordagens supervisionadas e não-supervisionadas. As abordagens supervisionadas referem-se às tarefas em que se é definido um foco, por meio de uma das variáveis, a qual espera-se que o processo de mineração explique a relação entre ele e as demais variáveis independentes. Já numa abordagem não-supervisionada, busca encontrar relações e padrões a partir do cruzamento entre todas as variáveis do conjunto de dados disponíveis.

Para a abordagem supervisionada, modelo empregado neste trabalho, várias são as tarefas inerentes ao processo de Mineração de Dados, que por sua vez utilizam-se de técnicas que especificam métodos que auxiliam a extração do conhecimento desejado. Segundo (MCCUE, 2014), para um mesmo problema, várias técnicas devem ser testadas e combinadas, visando a obtenção de uma melhor solução

3.5 Técnicas de Classificação

Um dos objetivos mais comuns na MD, a Classificação visa identificar a qual classe um determinado registro pertence. Para uma abordagem de aprendizado supervisionado, a classificação recebe uma base de dados rotulada (classificada) e aprende a classificá-la para os mesmos rótulos, novos registros. As técnicas de classificação também podem ser não-supervisionadas, nesse caso, o

modelo preditivo deve encontrar padrões e classificar os registros de acordo com medidas, que podem ser de similaridade ou de dissimilaridade.

As técnicas de classificação são componentes importantes para sistemas de suporte a decisão. Muitos problemas de tomada de decisão podem ser facilmente formulados para um problema de classificação, o que faz com que uma variedade de métodos estatísticos e heurísticas da literatura de Inteligência Artificial venham ser utilizadas para esses problemas. Por meio da literatura, percebe-se que o comportamento utilizado na escolha das melhores técnicas pra um determinado problema, tem sido justamente a utilização de várias abordagens. (KIANG, 2003), recomenda que para a construção de um bom sistema de classificação é fundamental a utilização de diferentes algoritmos ou combinação de diferentes métodos.

Para aplicações com o objetivo de dar suporte a sistemas de tomada de decisão, um fator importante nos modelos de predição de desempenho, é a interpretabilidade (HUYSMANS et al., 2011). Propriedade também destacada por (XING et al., 2015) que colocam a importância desses modelos serem do tipo "caixa-branca", para que sejam facilmente interpretados por humanos. Os autores ainda citam o inapropriado uso de alguns tradicionais modelos de predição concebidos por meio de técnicas como máquinas de suporte a vetor e redes neurais, devido à necessidade de conhecimento avançado de computação para seu entendimento, validação e refinamento. Não diferente, (ROMERO; VENTURA, 2013) enfatizam a importância da interpretação dos resultados da mineração de dados em ambientes educacionais, e que, embora algumas abordagens "caixa-preta" possam ter um melhor desempenho preditivo, tornam-se pouco úteis no empoderamento de pessoas.

3.6 Árvores de Decisão

A Árvore de Decisão é um fluxo em uma estrutura de árvores, em que cada nó representa um teste de um dos atributos, cada ramo representa o resultado desse teste e cada folha possui uma classe. A partir do campo escolhido como dado de saída, que será exibido nas folhas da árvore, o campo mais significativo é colocado como nó raiz (topo) e, partindo daí, de forma recursiva, novos campos são colocados como nós, conectando a raiz às folhas com os próximos campos mais significativos.

A técnica de árvore de decisão, devido à sua boa visualização, é comumente utilizada para tarefas de classificação, pois sua construção é intuitiva e de fácil entendimento por humanos. Outros fatores que corroboram para ampla utilização dessa técnica, além de possuir um bom desempenho, é

a possibilidade de trabalhar com dados multidimensionais e não requerer conhecimento prévio de domínio (J.HAN, J.PEI, M.KAMBER, 2012).

Para melhor compreensão, a Tabela 1 exibe um exemplo criado por meio de um conjunto de dados fictícios, que representam um subconjunto de dados simplificados de determinados alunos no ENEM.

Tabela 1: Exemplo de um subconjunto da base de dados do ENEM

Aluno	Sexo	Idade	Escolaridade da Mãe	Desempenho ENEM
Aluno 2	M	16	Superior	Satisfatório
Aluno 3	F	16	Superior	Satisfatório
Aluno 4	F	16	Fundamental	Insatisfatório
Aluno 5	F	15	Superior	Satisfatório
Aluno 6	M	16	Médio	Satisfatório
Aluno 7	M	18	Médio	Satisfatório
Aluno 8	M	18	Superior	Satisfatório
Aluno 9	M	19	Médio	Satisfatório
Aluno 10	M	20	Fundamental	Insatisfatório
Aluno 11	F	22	Superior	Satisfatório
Aluno 12	F	22	Médio	Insatisfatório
Aluno 13	F	15	Superior	Satisfatório
Aluno 14	F	15	Superior	Insatisfatório
Aluno 15	M	22	Fundamental	Insatisfatório
Aluno 16	F	15	Fundamental	Insatisfatório
Aluno 17	F	22	Superior	Satisfatório
Aluno 18	M	15	Fundamental	Satisfatório
Aluno 19	M	16	Superior	Satisfatório
Aluno 20	M	20	Superior	Satisfatório

Para gerar a árvore, a partir dos dados da Tabela 1, utilizou-se o algoritmo *Decision Tree Learner*, implementado na ferramenta de Mineração de Dados KNIME[2], similar ao C4.5 (QUINLAN, 1993). O algoritmo exige que o alvo (classe) seja um atributo nominal, já as outras variáveis independentes, também podem ser numéricas. As decisões de divisão são tomadas por meio de duas medidas de qualidade, Índice de Gini e Taxa de Ganho.

2 Sítio do KNIME na internet disponível em: https://www.knime.org/

Figura 4: Árvore de Decisão - Tabela 1

O algoritmo escolhido analisa os atributos de maneira individual em relação à classe dominante, "Satisfatório", demonstrado as condições que determinam a classificação dos alunos nos atributos mais significativos.

3.7 Regras de Classificação

Regras é uma boa maneira de representar informação e conhecimento. Um algoritmo de classificação por regras tem por objetivo encontrar relacionamentos entre os atributos e uma classe, a fim de que a regra encontrada possa predizer as classes de um novo registro de dados. Regras são fáceis de serem lidas e são consideradas como um dos paradigmas mais cotidianos para a aprendizagem de conhecimento interpretável.

As regras de classificação são do tipo SE-ENTÃO e são expressadas da seguinte forma:

SE condição ENTÂO conclusão

A condição consiste em uma série de testes e a conclusão se dá à classe ou classes que se aplicam aos casos cobertos pelas regras com uma probabilidade *p (WITTEN et al., 2011)*.

Essa técnica é uma popular alternativa aos modelos de árvores de decisão. É relativamente fácil traduzir uma árvore por meio de Regras. Uma regra é gerada para cada folha e a condição da regra consiste na condição necessária para cada nó desde a raiz, sendo a conclusão a classe atribuída para a folha.

A partir da árvore, Figura 3, é possível converter o modelo criado para regras de decisão, Quadro 1. O Conjunto de regras também foi gerado pelo algoritmo *Decision Tree Learner.*

$Idade$ <= 16.0 AND $Esc. da Mãe$ = 'Superior' => 'Satisfatório'
$Idade$ > 16.0 AND $Esc. da Mãe$ = 'Superior' => 'Satisfatório'
$Esc. da Mãe$ = 'Fundamental' AND TRUE => 'Insatisfatório'
$Esc. da Mãe$ = 'Médio' AND TRUE => 'Satisfatório'

Quadro 1: Exemplo de regras de classificação

Uma diferença entre a geração de regras por algoritmos próprios de regras de classificação e regras geradas por algoritmos de árvores de decisão é a restrição de que toda regra obtida a partir de uma árvore tenha o atributo raiz em sua condição. Outro importante fator é que a ordem em que as regras são apresentadas estabelece uma lista sequencial de decisão, com prioridade maior de predição de classe para a primeira regra. Quando um registro é classificado nenhuma outra regra posterior poderá ser aplicada sobre ele (KAMPFF, 2009). No Quadro 1, quatro regras independentes foram geradas. Caso um novo registro se enquadre na primeira regra, ele será classificado com o desempenho "Satisfatório", caso não se encaixe, as condições seguintes serão testadas. Se não atender a nenhuma das condições, estabelecidas pelo modelo, será classificado como "Insatisfatório".

Para avaliação dos modelos criados pelas regras de classificação, métricas podem ser inferidas, como cobertura e precisão, também chamado de risco e *lift.* A cobertura é o percentual de tuplas ou instâncias cobertas pela regra em relação ao total, e a precisão é o percentual das tuplas cobertas pela instância que a regra classificou corretamente (J.HAN, J.PEI, M.KAMBER, 2012)

3.8 Regressão Logística

A regressão logística, em sua forma tradicional, consiste de um modelo que relaciona um conjunto de p variáveis independentes $X_{1,}X_{2,\ldots},X_p$ a uma variável dependente e dicotômica

Y . Considerando a variável Y assumindo um valor 0 ou 1, o modelo permitira, por exemplo, a estimação direta da probabilidade de ocorrência de um evento para $(Y=1)$ (HOSMER; LEMESHOW, 1989)

$$P(Y=1)=\frac{\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)}$$

Equação 1: Função de probabilidade

E, consequentemente,

$$P(y=0)=1-p(y=1)=\frac{1}{1+\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)}$$

Onde β são os parâmetros do modelo, e a etapa de aprendizagem ou treinamento, basicamente se dá pela estimação dos coeficientes, geralmente pelo método da máxima verossimilhança (MONTGOMERY et al., 2009)

A transformação que está por trás do modelo logístico, a chamada função *logit*, denotada por $g(x)$ é uma função linear nos parâmetros β , contínua e que pode variar de $-\infty$ a $+\infty$:

$$logit(x)=g(x)=\ln\left[\frac{P(Y=1)}{1-p(y=1)}\right]=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p$$

Equação 2: Função de resposta *logit*

Essa técnica estatística é possível através de uma alteração em sua função de resposta, que permite o tratamento de variáveis dicotômicas ao invés das quantitativas, como nos modelos lineares (J.HAN, J.PEI, M.KAMBER, 2012). Esse comportamento pode ser observado nos gráficos presentes na figura 4.

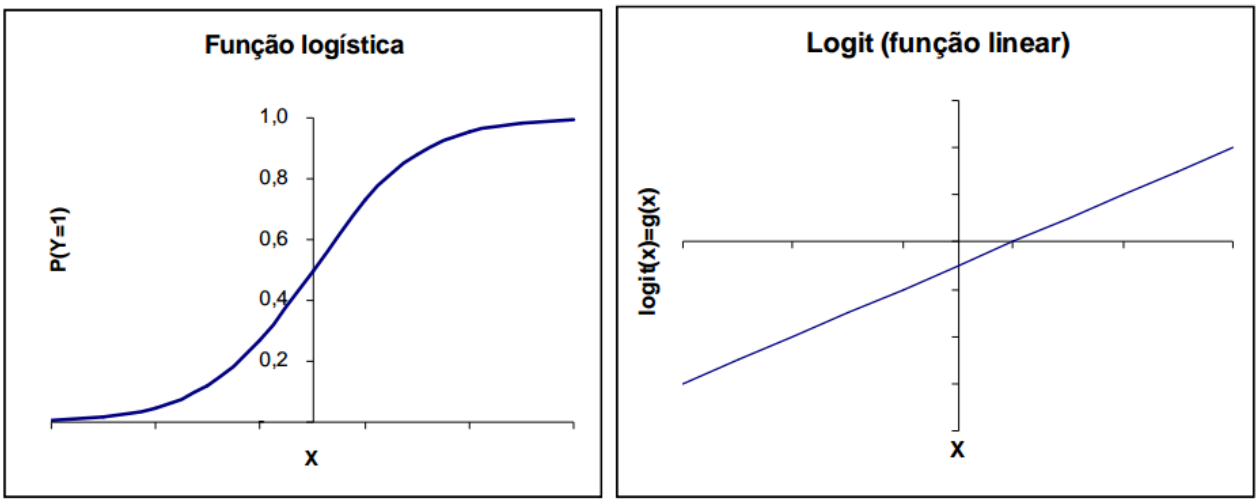

Figura 5: Função logística e a relação logit (BITTENCOURT, 2003)

3.9 Avaliação de Modelos

Para avaliar os desempenhos das regressões logísticas treinadas neste trabalho, foram utilizadas, além das medidas da matriz de confusão, as métricas KS2 (Komolgorov-Sminorv para classificação binária) e AUC_ROC (Área sob a curva ROC).

A matriz de confusão, técnica mais simples de avaliação de modelo, é a sumarização dos termos gerados pelo sistema de classificação entre os dados atuais e a classificação predita. Dadas duas classes, pode-se falar em tuplas positivas e tuplas negativas. Verdadeiros positivos se referem à tuplas positivas classificadas corretamente pelo sistema classificador, enquanto que os verdadeiros negativos são as tuplas negativas marcadas corretamente. Falsos positivos são as tuplas negativas marcadas erroneamente e por fim, falsos negativos as tuplas negativas marcadas também de maneira equivocada pelo classificador (J.HAN, J.PEI, M.KAMBER, 2012), conforme Tabela 2.

Tabela 2: Matriz de Confusão

	Positivo atual	Negativo atual
Predição positiva	*TP*	*FP*
Predição negativa	*FN*	*TF*

Partindo da tabela de confusão, várias medias de avaliação do modelo podem ser extraídas, conforme tabela da Figura 6 (em inglês).

Measure	Formula
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Figura 6: Tabela com medidas geradas a partir de uma matriz de confusão

A curva *Receiver Operation Characteristics* (ROC) é uma técnica para visualização, organização e seleção de classificadores baseados em sua performance. Segundo (FAWCETT, 2006) curva ROC foi longamente utilizada na teoria de detecção de sinal para descrever a relação entre as taxas de acerto e os alarmes falsos dos classificadores. Na literatura, encontra-se uma ampla utilização da técnica em sistemas médicos de diagnósticos. Um dos pioneiros na utilização dessa técnica na área de aprendizagem de máquina foi (SPACKMAN, 1989) que demonstrou os valores das curvas ROC na avaliação e comparação de alguns algoritmos. Nos últimos anos o uso da técnica na área de mineração de dados vem crescendo em detrimento do uso da medida de acurácia, considerada pobre enquanto medida de performance (FAWCETT, 2006). Para (PROVOST et al., 1998) o uso da medida de acurácia na comparação de modelos é ineficiente. O autor ainda destaca a impossibilidade de se fazer fortes conclusões partindo de uma simples métrica.

Os gráficos ROC são bidimensionais. A taxa dos verdadeiros positivos (razão entre os positivos corretamente classificados e o total de positivos) é plotada no eixo Y e a taxa de falsos positivos (razão dos negativos incorretamente classificados e o total de negativos) no eixo X. No plano, um ponto tem uma melhor performance se está mais ao nordeste (taxa alta de verdadeiro

positivo e baixa taxa de falso positivo), demonstrando uma relação entre os benefícios (verdadeiros positivos) e os custos (falsos negativos) (FAWCETT, 2006).

Para comparar classificadores é necessário sintetizar a curva ROC para um simples medida escalar que possa representar a performance dos classificadores. A maneira mais comum é calcular a área sobre a curva, em inglês, *Area under Curver* (AUC). A métrica possui o valor 1 como máxima e melhor medida, no entanto, não se deve ter valores menores que 0,5 pra classificadores reais, pois essa é a área da linha diagonal com início no ponto mínimo (0,0) e fim no ponto máximo (1,1) do plano.

Um indicador no domínio contínuo do escore é o teste Kolmogorov-Smirnov (KS). Originalmente criado pra determinar se duas amostras possuem a mesma distribuição, o teste é uma importante medida de separação. Segundo (SIEGEL, 1975) o teste envolve especificar a distribuição de frequência acumulada que ocorreria e compará-la com a distribuição de frequência acumulada observada, viabilizando assim, a utilização do método como medida de dissimilaridade para problemas de classificação (KS2) (ADEODATO, 2016). Para ambos os propósitos a métrica usual é calculada através da máxima distância entre as curvas acumuladas dos escores de cada amostra. Todavia, a dissimilaridade do modelo é avaliada em um único ponto operacional.

Uma boa métrica para avaliar a representatividade de instâncias de uma mesma classe, as medidas de confiança e *lift,* a última também chamada de risco, são muito utilizadas para entender a relevância dos modelos preditivos baseados em regras. O *lift* é calculado com a frequência relativa de representantes de uma classe de uma regra pela frequência relativa de representantes da mesma classe na população. Ou seja, é o percentual das tuplas cobertas pela instância que a regra classificou corretamente (J.HAN, J.PEI, M.KAMBER, 2012).

3.10 Ferramentas Utilizadas

Segundo o influente site da área de mineração de dados, *Kdnuggets*[3]*,* em 2013, das 5 melhores ferramentas utilizadas em projetos reais de mineração de dados, apenas uma é comercial. O domínio de software *opensource* para este fim, decorre da maturidade e disponibilidade de um grande número de implementações de algoritmos de aprendizagem de máquinas nessas ferramentas (JOVIĆ et al.). Os autores ainda enfatizam a inexistência de uma "melhor ferramenta", além de expor as principais vantagens e fragilidades de vários softwares. Por fim, colocam os softwares *R,*

3 – Sítio sobre *Bussiness Analytics, Big Datam Data Mining* e *Data Science*

Weka e *KNIME* dentro de um conjunto de ferramentas capaz de ser utilizado na maioria das tarefas de DM.

Para a realização desta pesquisa foram utilizadas somente ferramentas licenciadas sob licença de software livre. O Sistema de gerenciamento de banco de dados (SGBD) *PosgreeSQL* foi utilizado na fase de concepção do *data-mart*, além de servir para a fase inicial da exploração dos dados, fase que também utilizou-se do *Calc,* ferramenta de escritório baseada em planilha presente no pacote *LibreOffice*. Nesta fase também foram exploradas algumas funcionalidades da plataforma de análises *KNIME*. Para esta última, destaca-se a estruturação por meio de fluxos de todo o pré-processamento e transformação do conjunto de dados. Para a mineração, utilizou-se das ferramentas *R* para a técnica de regressão logística e *Weka* para geração das árvores e indução de regras.

Todo o processo de KDD e as abordagens utilizadas em cada etapa está ilustrado na Figura 7, servindo para entendimento geral do fluxo adotado nesta pesquisa.

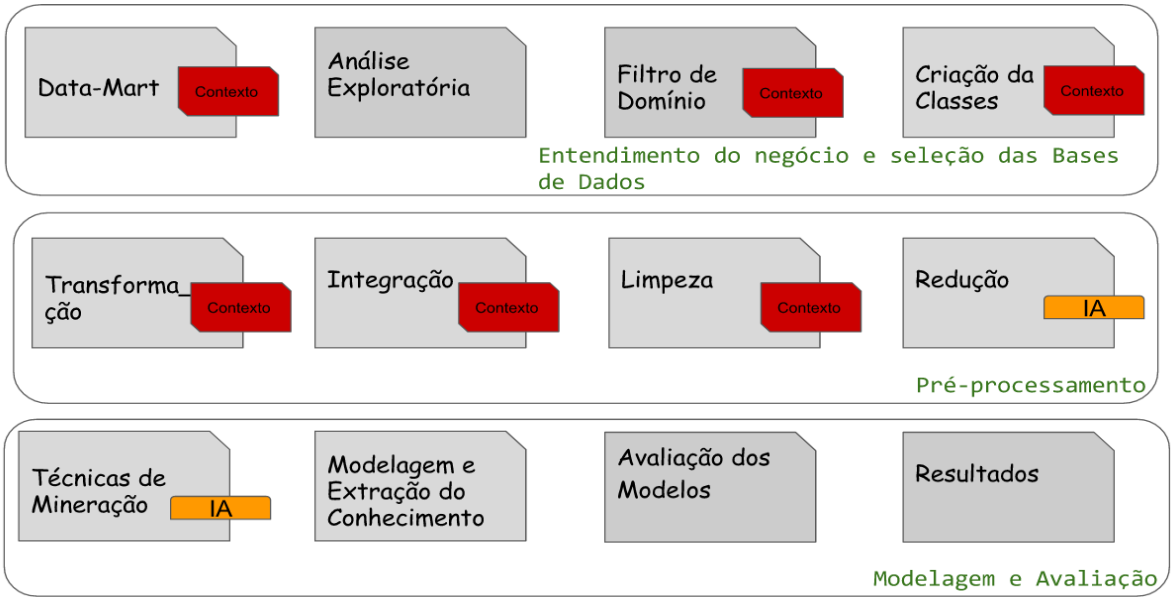

Figura 7: Processo KDD desenvolvido na dissertação

4 SELEÇÃO, ANÁLISE E TRANSFORMAÇÃO DOS DADOS

Este capítulo descreve as fases iniciais do processo de KDD. São evidenciados todos os procedimentos de extração, redução, limpeza e transformações realizadas para a formação dos conjuntos finais de dados para modelagem, denominadas *inputs*..

4.1 Bases de Dados

Os dados escolhidos para o estudo proposto nesta dissertação foram os últimos disponíveis para o ano de 2014 para os Microdados do ENEM, Censo Escolar e Enem por Escola, todos disponíveis no portal do INEP.

A base de dados do ENEM é formada por variáveis relativas aos alunos, à escola e à prova em si. As provas estão estruturadas em 4 áreas de conhecimento, mais uma prova de redação. As provas contêm, cada uma, 45 questões de múltipla escolha, englobando os componentes curriculares descritos na Tabela 3.

Tabela 3: Componentes curriculares – ENEM

Área de Conhecimento	**Componentes Curriculares**
Linguagens, Códigos e suas tecnologias (LC)	Língua Portuguesa, Literatura, Língua Estrangeira (Inglês ou Espanhol), Artes, Educação Física e Tecnologias da Informação e Comunicação
Matemática e suas tecnologias (MT)	Matemática
Ciências Humanas e suas tecnologias (CH)	História, Geografia, Filosofia e Sociologia
Ciências da Natureza e suas tecnologias (CN)	Química, Física e Biologia

Os Microdados do ENEM são disponibilizados acompanhados da sua documentação em vários arquivos. Os dados em si estão em um único arquivo no formato CSV (*Comma-Separated Values*) e trazem informações sobre a avaliação, de controle dos inscritos e das escolas. Para o ano

de 2014, foram disponibilizados 6 pastas, a saber: Dados, Dicionário, Inputs, Leia-me, Planilhas e Provas/Gabritos, totalizando um total de pouco mais de 6 GB de dados.

O Censo Escolar, também disponibilizado no formato *CSV*, é o principal instrumento de coleta de informações da educação básica. A sua base de dados é dividida e traz informações das Escolas, Turmas, Docentes e Matrículas em diferentes tabelas. Além dos dados, também é disponibilizado um dicionário das variáveis e alguns filtros que auxiliam os pesquisadores na exploração e utilização dos Microdados.

Junto aos microdados do ENEM, o INEP tem divulgado, nos últimos anos, os resultados e informações contextuais dos estabelecimentos de ensino. Essas informações estão relacionadas às suas proficiências médias em cada uma das áreas de conhecimento do ENEM, com informações geradas a partir do Censo Escolar e ENEM. Alguns exemplos são: o porte da escola (gerado a partir da quantidade de alunos declarados no censo escolar em turmas do ensino médio), a taxa de participação (gerado a partir da diferença entre a quantidade de alunos em turmas do ensino médio e a quantidade de inscritos no ENEM) e Média dos Top 30 (gerada a partir da média dos trinta melhores alunos da escola).

Além dessas informações, passíveis de serem geradas apenas com os dados disponibilizados pelo INEP, o sistema traz também alguns indicadores mais elaborados, que puderam ser formulados somente após a junção da base do ENEM e Censo Escolar no grão aluno. Essa junção ainda não é possível para os pesquisadores que possuem acesso apenas aos dados abertos do INEP, pois não existe uma chave de ligação entre as bases no grão indicado. Para isso, o INEP realizou uma busca exata dos dados pessoais dos alunos informados no Censo Escolar na base de dados dos inscritos no ENEM 2014, aplicando um processo de consistência por meio de *scripts* de verificação de fonética (INEP, 2015a)

Os indicadores elaborados são: i) Indicador de Adequação da Formação Docente (IFD), ii) Indicador de Permanência na Escola (IPE) e iii) Indicador de Nível Socioeconômico (Inse). O IFD apresenta uma classificação dos docentes em exercício na educação básica, considerando sua formação acadêmica e a(s) disciplina(s) que leciona. Os docentes são classificados em 4 grupos, de acordo com os requisitos legais de formação para cada disciplina em que atua e em cada instituição. O índice é então calculado por meio da porcentagem de ocorrências de professores que pertencem ao grupo que possuem formação ideal para lecionar as disciplinas.

O IPE é um indicador que se remete aos estudos do "efeito-escola", que procura estimar o impacto da escola sobre o desempenho dos alunos. O indicador parte da premissa de que os ganhos dos alunos quando potencializados pela escola estão diretamente ligados ao tempo em que o mesmo esteve exposto aos processos de ensino e de aprendizagem na respectiva instituição escolar. Através

de censos passados, são identificadas as instituições em que os alunos cursaram os anos anteriores ao ensino médio, e de maneira ponderada, este tempo é relacionado com o seu desempenho (INEP, 2015b).

Já o Inse, utilizou-se de três bases de dados da educação, que são: a Avaliação Nacional da Educação Básica (ANEB), a Avaliação Nacional do Rendimento Escolar (Anresc) e a base do ENEM. Foram considerados apenas os alunos, que ao preencher o questionário contextual, assinalaram pelo menos 5 questões sobre: posse de bens no domicílio, contratação de serviços domiciliares, renda familiar e escolaridade dos pais. Utilizando uma abordagem probabilística para preenchimento de repostas faltantes, calculou a medida do nível socioeconômico para cada aluno através de uma escala contínua, e, a partir da análise de *cluster* (K-means), foram classificados sete grupos da seguinte maneira: Muito Baixo, Baixo, Médio Baixo, Médio, Médio Alto, Alto e Muito Alto (INEP, 2015c)

Assim, segundo o INEP, o Inse, juntamente ao IPE e ao IFD,

> "contribuem para melhor contextualização dos resultados dos estabelecimentos que oferecem o Ensino Médio, ao destacar a importância de se observar os fatores extra e intraescolares no momento de sua análise, qualificando tanto o diagnóstico e a avaliação dos eventuais problemas encontrados, quanto a proposição e a implementação de medidas administrativas e pedagógicas elaboradas para solucioná-los."

4.2 Integração das Bases de Dados

Devido à já mencionada inexistência de uma variável-chave que interliga as bases do ENEM e Censo Escolar no grão aluno, optou-se por relacionar o atributo identificador da escola, presente nas duas bases, a fim de normalizar o conjunto de dados em conformidade aos objetivos da pesquisa. Dessa forma, todas as variáveis das escolas, presentes na base do Censo (tabela Escolas), passaram a estar relacionadas aos alunos da base do ENEM que estudam/estudaram naquela escola. O processo de transformação de granularidade pode ser observado mais adiante ainda neste capítulo na Figura 19.

4.3 Filtro do Domínio

Após a composição do *data-mart,* foi necessário aplicar alguns filtros na base de dados, a fim de selecionar apenas registros de interesse da pesquisa. Nesse sentido, os milhões de linhas iniciais foram reduzidas para apenas 22.183 (vinte e dois mil cento e oitenta e três). De maneira concomitantemente, os seguintes pontos como filtro de domínio foram considerados: ser aluno dos Institutos Federais de Educação, Ciência e Tecnologia; estar cursando o ensino médio regular com conclusão em 2014 excluindo o ensino médio não seriado; ter realizado as quatro provas objetivas e a prova de redação obtendo proficiências superiores a zero em todas as provas objetivas [4].

4.4 Análise Exploratória dos Dados

Com o domínio de pesquisa definido, é necessário, antes de aplicar qualquer técnica de estatística mais avançada ou iniciar a fase de processamento, um entendimento dos dados através da análise exploratória. O principal objetivo dessa análise é conhecer as características da base de dados, apresentando-as através de tabelas e gráficos apropriados. Nesta fase também são detectados valores discrepantes (*Ouliers*), registros incompletos e até mesmo, valores faltantes.

O mapa do Brasil, Figura 8, ilustra as instituições que tiveram alunos presentes nas bases de dados da pesquisa após o filtro de domínio. Percebe-se que há a presença de instituições em todos os estados brasileiros, porém é notória uma maior concentração nas regiões sudeste e nordeste. A Figura 9 exibe a distribuição de frequência dos dados por região do país. Por meio do gráfico de bolhas da Figura 10, percebe-se ainda a formação de dois grupos quanto ao número de inscritos, sendo grupo 1: Sudeste e Nordeste e grupo 2: Sul, Centro-oeste e Norte.

4Conforme Portaria INEP n° 116, de 22 de junho de 2015. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/legislacao/2015/portaria_n_267_19062015.pdf>

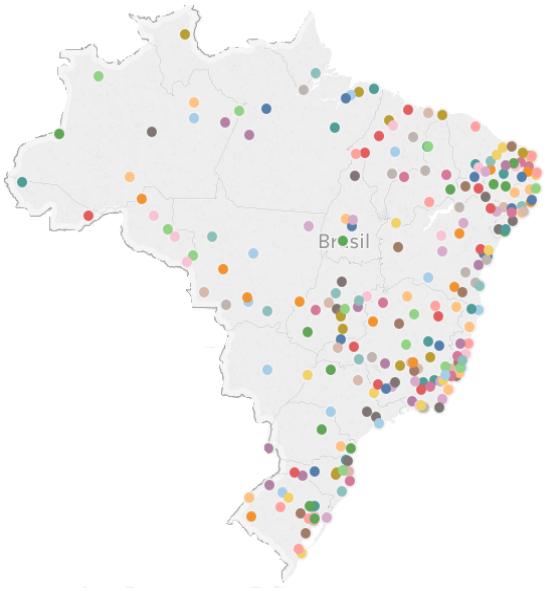

Figura 8: IFs no Brasil

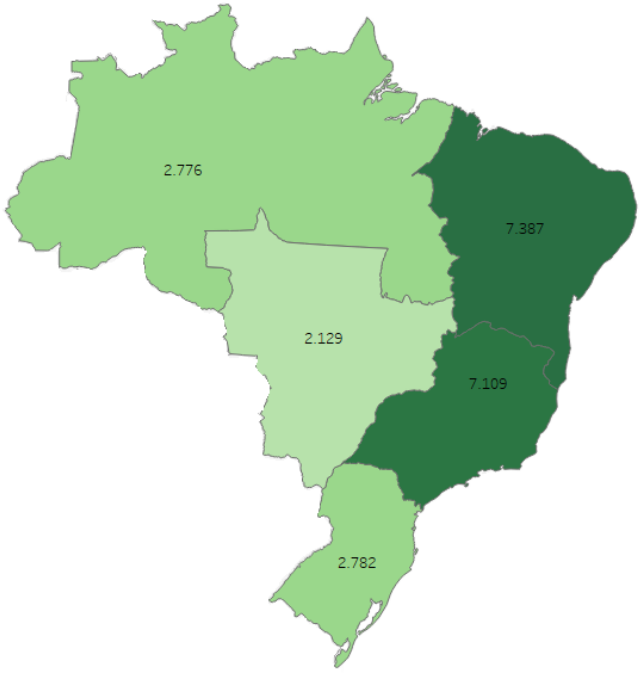

Figura 9 - Mapa: Inscritos por região do Brasil

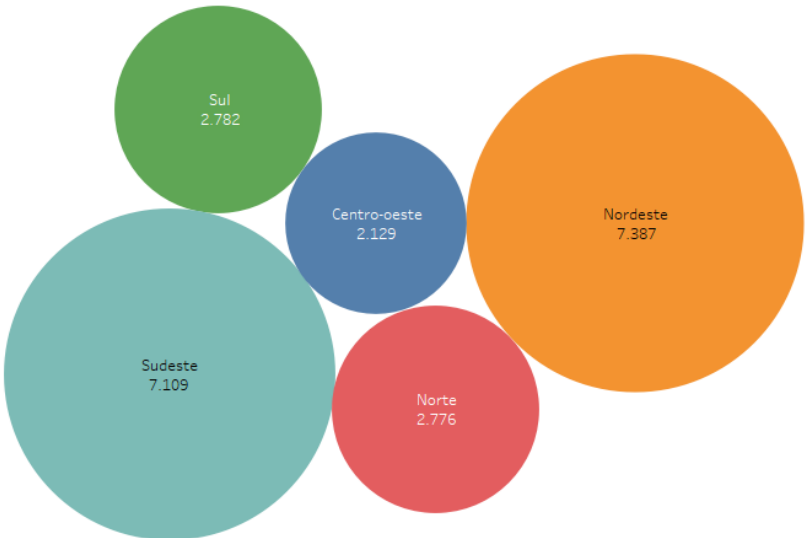

Figura 10 - Gráfico: inscritos por regiões do país

Para que os dados sejam compreendidos e comparados entre estados e regiões, a partir do desempenho dos alunos[5], foi utilizado o gráfico mapa de árvore, Figura 11 e 12. É possível notar que a região sudeste, além de possuir o maior número de inscritos, também possui o melhor desempenho, estando todos os estados dessa região entre os 5 melhores do país. A região norte foi a que demonstrou os piores resultados, não tendo nenhum dos seus estados presentes entre os 10 melhores.

5 – Calculado através da média do somatório de médias das provas objetivas e redação, semelhante ao ENEM por Escola 2014. Disponível em: download.inep.gov.br/educacao_basica/enem/nota_tecnica/2014/nota_explicativa_enem_2014_por_escola.pdf

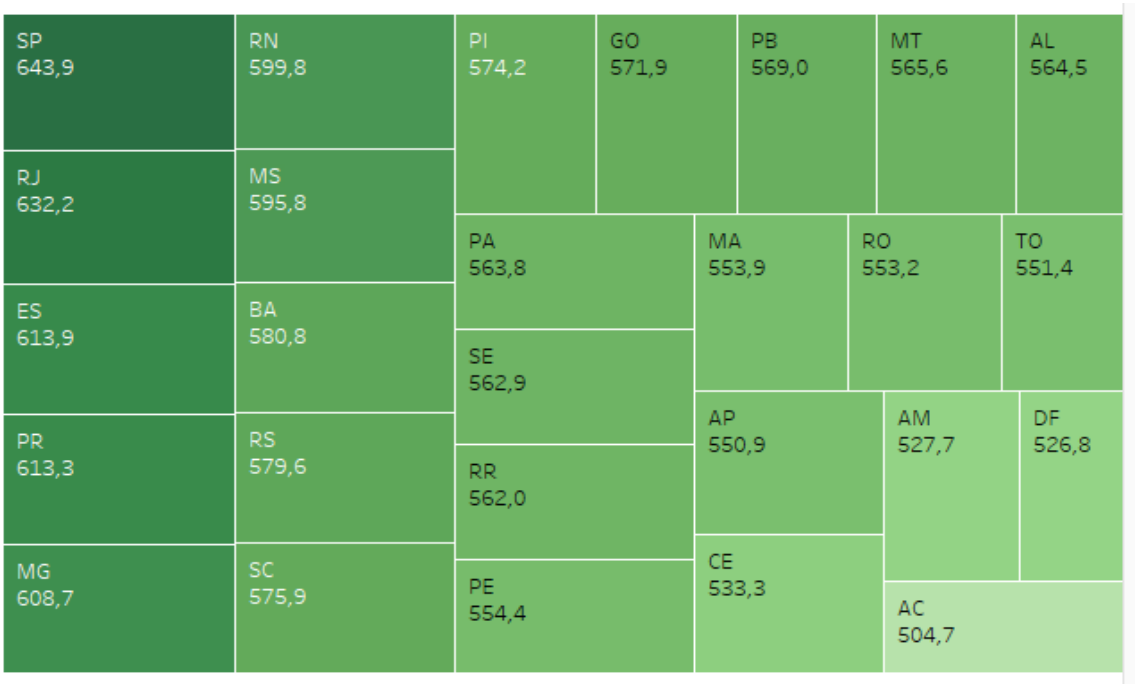

Figura 11 - Gráfico: desempenho por estados federativos

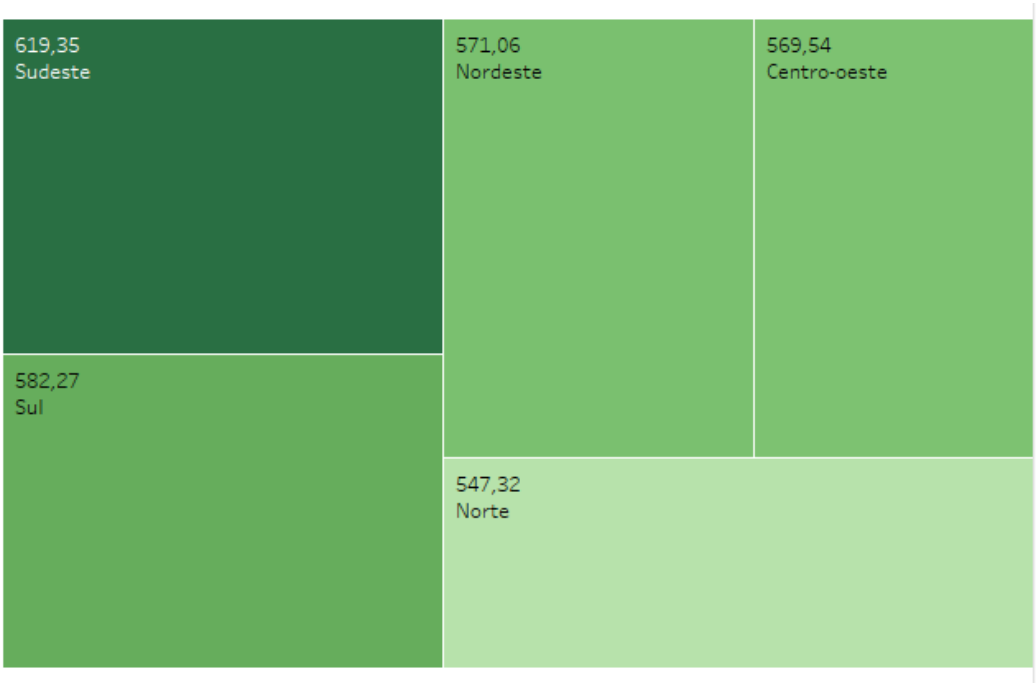

Figura 12 - Gráfico: desempenho por regiões

A fim de entender a correlação entre o desempenho nas proficiências para cada uma das áreas de conhecimento (CH, LC, MT e CN) e a redação, um gráfico de dispersão foi construído,

Figura 13. As variáveis possuem alta correlação para todos os casos, com destaque para as proficiências CH e LC. É possível perceber uma quantidade homogênea de indivíduos que tiveram nota zero na redação, porém com diferentes resultados para as 4 áreas.

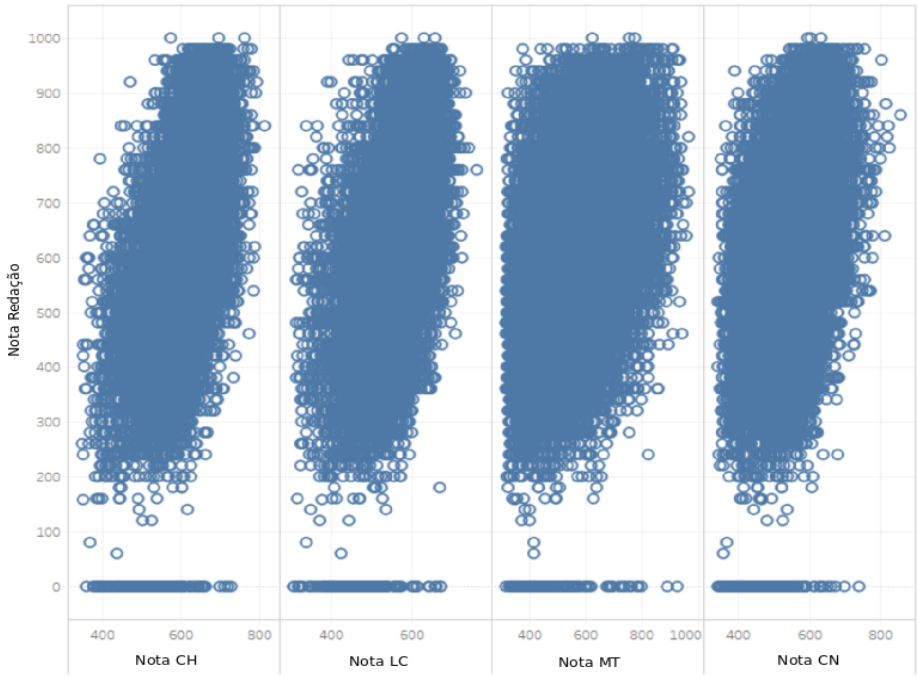

Figura 13: Gráfico - correlação entre as proficiências e redação

Em uma análise de distribuição de frequência de gêneros, a Figura 14 demonstra um equilíbrio quando analisa-se os dados para todo o país.

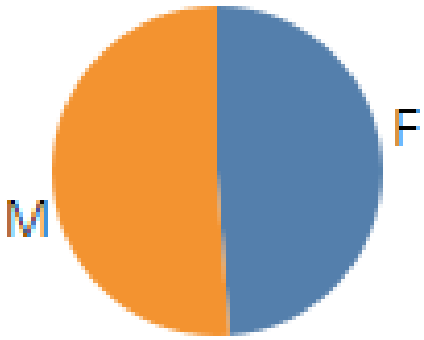

Figura 14: Gráfico - distribuição de gênero

Quanto à raça declarada, a Figura 14 exibe uma disparidade de frequência entre as regiões do país. A região sul se destaca pela grande maioria de brancos e a região nordeste com uma maioria declarada parda. Todos os estados possuem uma quantidade pequena de negros. A região norte foi a que apresentou uma maior quantidade de indivíduos da raça indígena.

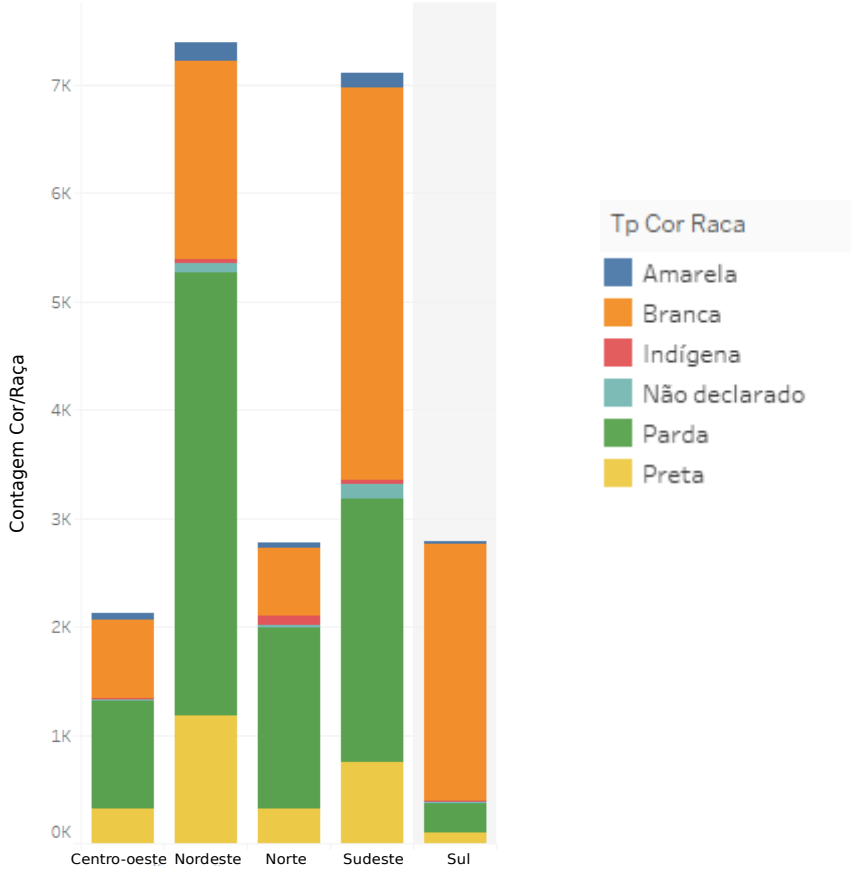

Figura 15: Gráfico - raça por região

Outros gráficos importantes de serem plotados são os que representam características já conhecidas do domínio, que tem como objetivo identificar valores discrepantes (*outliers*). A Figura 16 ilustra a relação da idade dos participantes e sua média final no exame, em que é possível perceber a existência de valores que fogem do padrão de alunos que estejam cursando o Ensino Médio[6]. A Figura 17, explora ainda mais essa diferença.

6 – Conforme lei n° 12.796 de 4 de abril de 2013 encontrada em: <http://www.planalto.gov.br/CCIVIL_03/_Ato2011-2014/2013/Lei/L12796.htm>

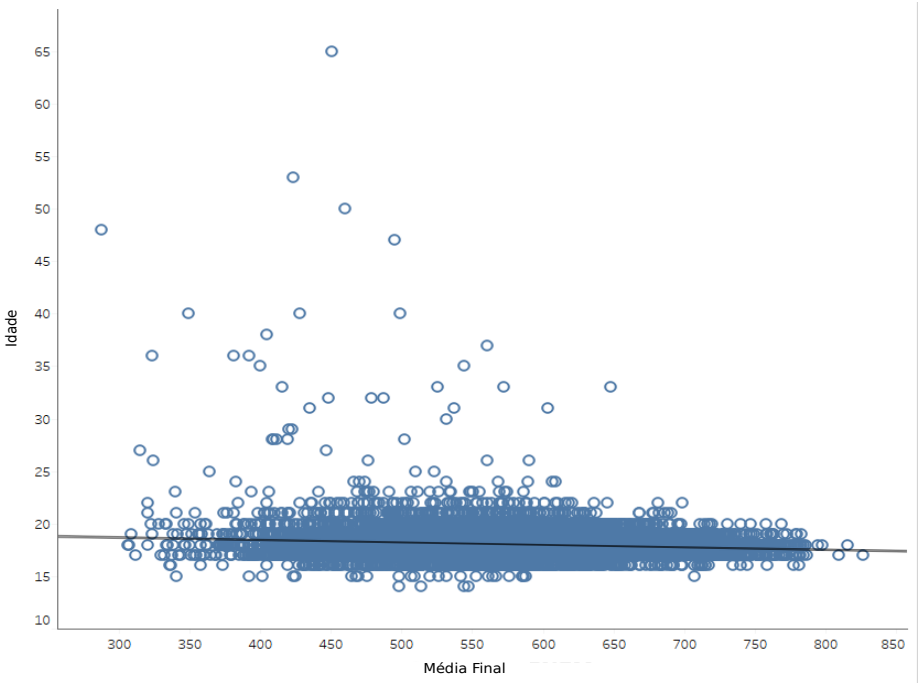

Figura 16: Gráfico - relação idade e média final

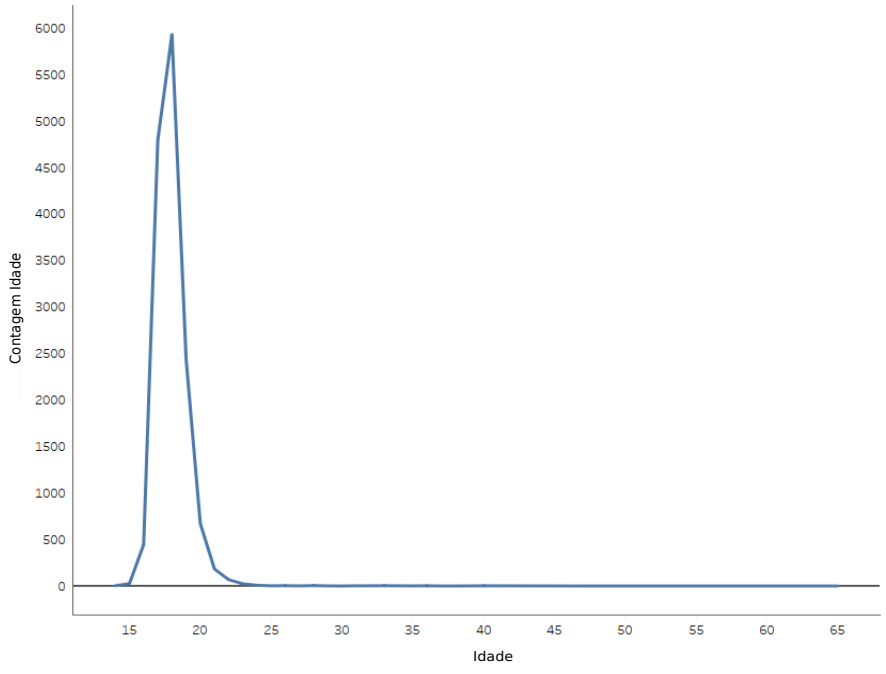

Figura 17: Gráfico - frequência da idade dos participantes

O desempenho dos estudantes, definido nesse trabalho como a média dos resultados obtidos nas quatro provas objetivas e redação, pode ter sua frequência ilustrada por meio de um gráfico histograma, Figura 18, e sua distribuição estatística por meio de um gráfico *box-plot*, Figura 19.

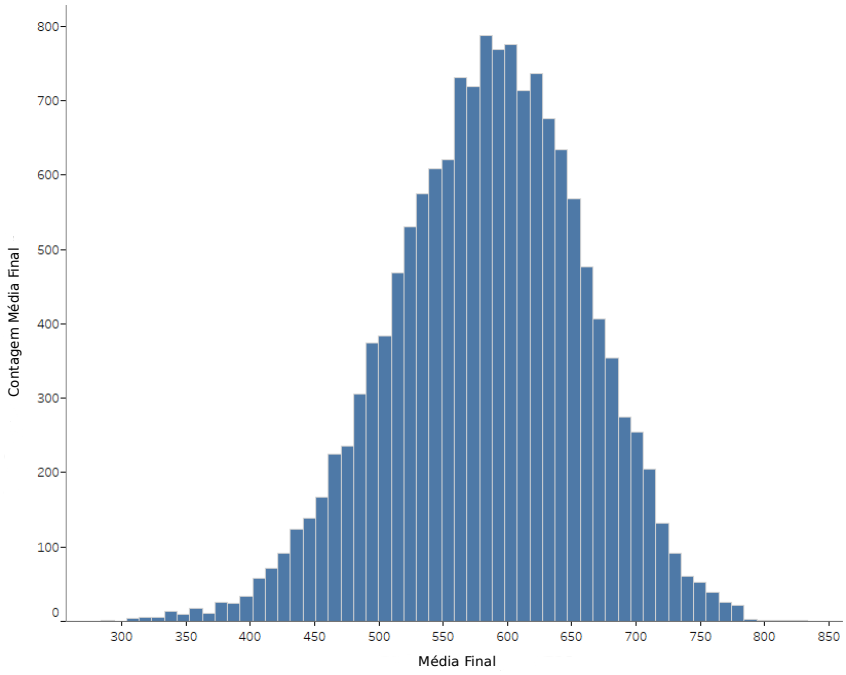

Figura 18: Gráfico - histograma média final ENEM

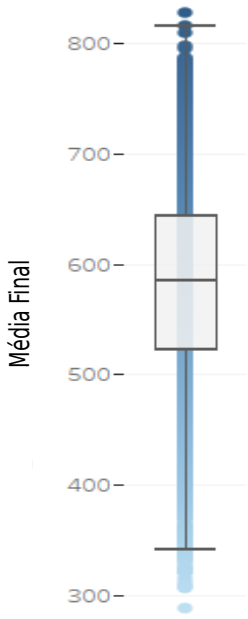

Figura 19: Diagrama de Caixas - média final

A fim de entender a correlação entre as variáveis independentes, calculou-se por meio da utilização do coeficiente de *Pearson* e Teste qui-quadrado, para as variáveis numéricas e nominais, respectivamente, os coeficientes de correlação. Os valores variam de 1(forte correlação positiva) e -1 (forte correlação negativa). Devido ao grande número de variáveis, a matriz ilustrada na Figura 20 com a representação da correlação em cores, foi reduzida para apresentação.

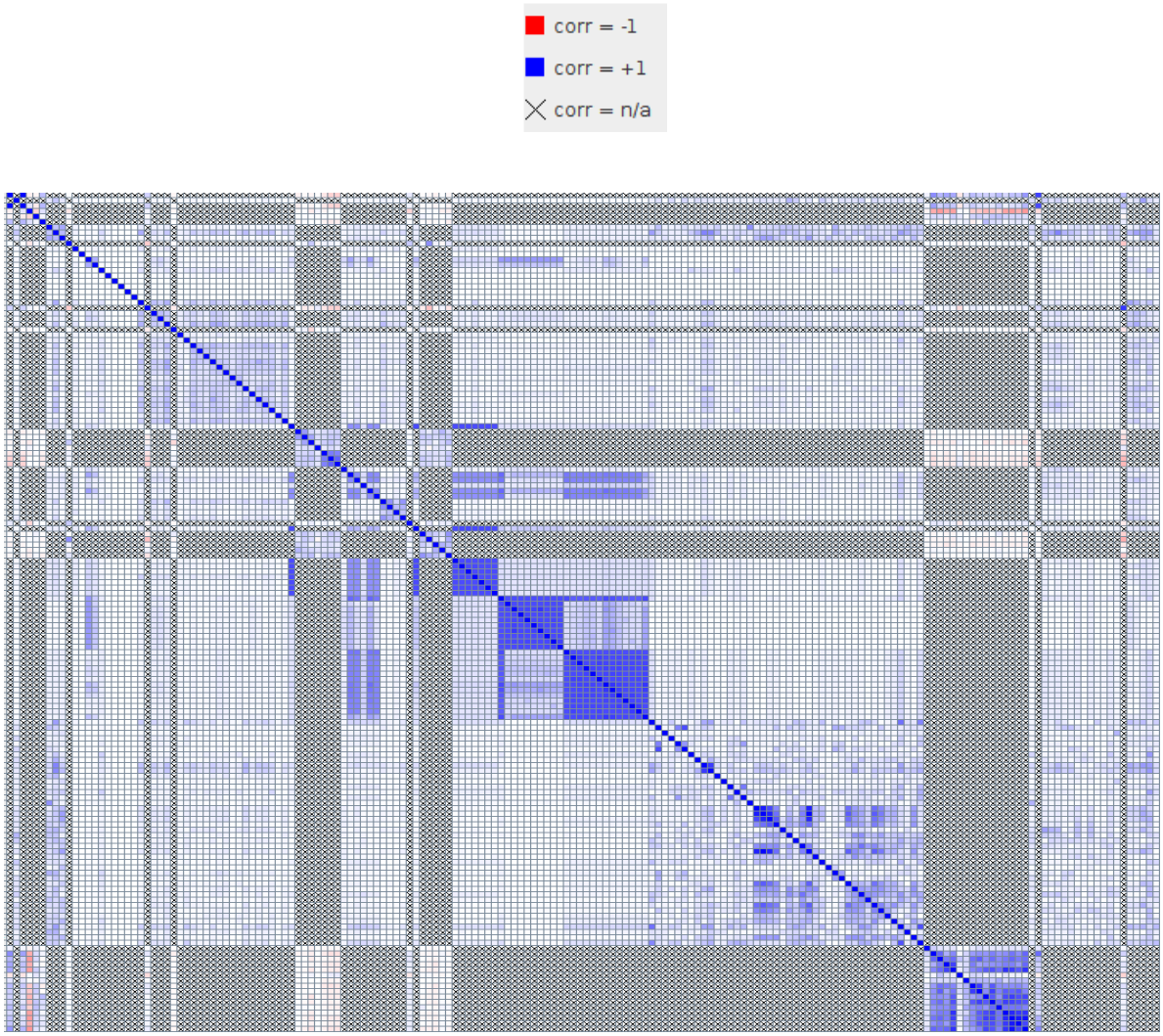

Figura 20: Matriz de correlação entre variáveis independentes

4.5 Variáveis criadas

Considerando a titulação dos professores como um importante fator no desempenho dos alunos (GOLDHABER; BREWER, 1996; CLEARINGHOUSE, 2000), relação também defendida por (TRAVITSKI, 2013) que a coloca de maneira proporcional e direta, procurou-se inserir o valor dessas informações ao conjunto de dados.

Partindo das informações presentes na base do Censo (tabela Docentes), pôde-se extrair o grau de titulação de todos os professores para cada escola. Porém, como agregar essas informações para o conjunto de dados preservando a sua essência conforme colocada pelos especialistas de domínio? Nesse sentido, considerando o estudo de caso que trata da transformação de granularidade em banco de dados relacionais, por meio de D³M,(ADEODATO, 2016), optou-se em ponderar, arbitrariamente, a titulação de cada professor, resguardando que cada docente fosse considerado apenas uma vez, no seu maior grau de titulação. Isto é, o peso (w) da titulação do professor i na escola j foi definida por:

$$w_{ij} = \begin{bmatrix} 1, \textit{se o professor i possui graduação e leciona na escola j} \\ 2, \textit{se o professor i possui especialização e leciona na escola j} \\ 3, \textit{se o professor i possui mestrado e leciona na escola j} \\ 4, \textit{se o professor i possui doutorado e estuda na escola j} \end{bmatrix}$$

Logo,

$$ITD_j = \frac{\sum_{i=1}^{i=n} X_{ij} w_{ij}}{N_j \times 4} \quad , \qquad se\ w_{ij} = 1$$

Equação 3: Fórmula do índice de titulação docente

Onde:

$ITD_j = Indicador\ de\ titulação\ na\ escola\ j$

$\sum_i X_{ij} = Total\ de\ docentes\ com\ apenas\ graduação\ na\ escola\ j$

$N_j = Total\ de\ docentes\ na\ escola\ j$

Exemplo de cálculo: suponha que uma escola tenha 50 professores e que, destes, 7 são doutores, 20 possuem apenas mestrado, 15 apenas especialização e 8 apenas graduação. O índice de titulação dessa escola será 0,63. Vale ressaltar que, um professor pode lecionar em mais de uma escola e que a sua titulação é considerada em todas as escolas em que atua.

Posteriormente, esses indicadores foram transformados para o grão aluno, conforme ilustra a Figura 21. Além disso, a partir de colunas identificadoras, alguns campos como: “Escola_Capital” (se a escola está situada em uma capital federal ou não) “Escola_Interior” (se a escola está situada no interior ou não), “Região_Escola” (região brasileira da escola) e “Estuda_Fora” (alunos que

residem em cidades diferente das escolas em que estudam) também foram gerados com o intuito de aumentar o número de informações e permitindo um maior entendimento semântico dos dados. Por fim, após a junção de todas as bases e a criação das novas variáveis, formou-se uma grande base de dados no grão aluno, compondo um novo *data-mart* composto de 316 variáveis.

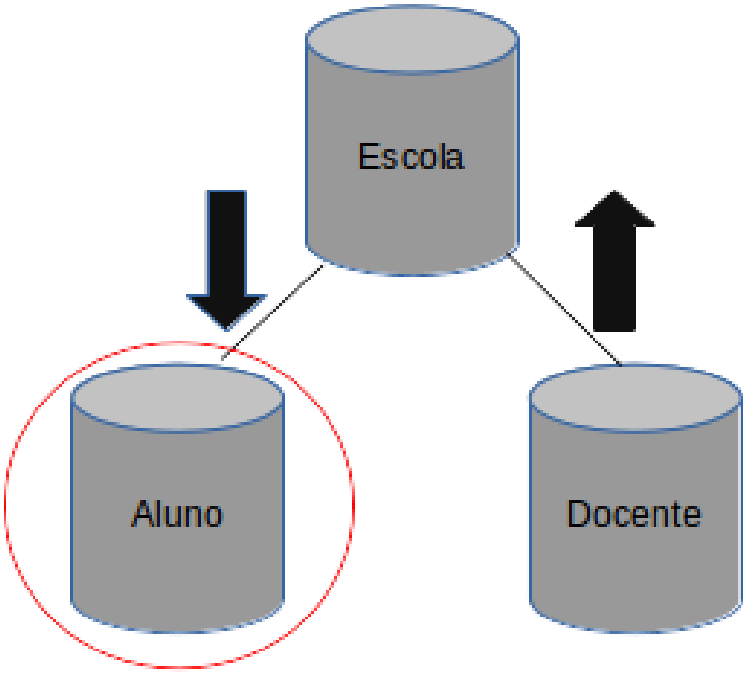

Figura 21: Ciclo da transformação de granularidade

4.6 Criação das Classes (Alvo)

Para a criação dos modelos preditivos é necessário classificar os alunos em função do seu desempenho no ENEM. Sendo assim, utilizou-se da média final dos alunos em todas as 4 provas objetivas e redação. Em seguida, a fim de classificá-los de maneira binária em alunos com bom desempenho e ruim, foi utilizada a separação por quartis, considerando o quartil superior como limiar de binarização (ADEODATO, 2016). Assim, alunos com bom desempenho são aqueles cuja a média esteja no quartil superior da média dos alunos.

4.7 Estatística Descritiva

Utilizada na fase inicial, a estatística descritiva é utilizada para descrever e resumir os dados, priorizando a menor perda de informações possível. Tipos de medidas de frequências, de tendências

centrais, de variância e de nível de taxa de nível de preenchimento foram coletadas de todas as variáveis. Algumas medidas que fazem referência ao alvo, como ganho de informação, também foram relacionadas. Uma tabela com toda a estatística descritiva dos dados está apresentada no Apêndice B e C deste trabalho.

4.8 Redução dos Dados

Muitos fatores afetam o sucesso de uma aplicação de mineração de dados. A qualidade dos dados é um desses. (HALL; HOLMES, 2003) destacam que a irrelevância e a redundância de informações, bem como a existência de ruídos e discrepâncias, podem ser fatores dificultadores no processo de descoberta de conhecimento.

Dessa forma, a redução da dimensionalidade além de ser útil para diminuir o tempo de execução dos algoritmos de mineração, também auxiliam no desempenho final de algoritmos de classificação, fato destacado por (J.HAN, J.PEI, M.KAMBER, 2012). Os autores ressaltam ainda a necessidade de se manter a integridade dos dados originais, além da possibilidade de obtenção de resultados mais compreensíveis.

Muitos trabalhos foram encontrados na literatura acerca da redução da dimensionalidade para aplicações de DM. Essa redução pode ser feita de duas maneiras, diminuindo o número de instâncias (KALEGELE et al., 2012) ou diminuindo o número de atributos que descrevem essas instâncias (KOHAVI; JOHN, 1997).

A redução de atributos de um conjunto de dados possui duas abordagens; em inglês, *Wrapper model (KOHAVI; JOHN, 1997) e filter model (SUBRATA, DAS, 1971)*. A abordagem *Wrapper model* requer um determinado algoritmo de aprendizagem e utiliza a sua performance para seleção do melhor conjunto de variáveis. Já na abordagem *filter model*, nenhum algoritmo de aprendizagem é utilizado e a redução de dimensionalidade acontece baseada em características dos próprios dados

Utilizando do conhecimento de domínio e da base de dados, optou-se, primeiramente, em retirar todas as variáveis irrelevantes à classe alvo, além das variáveis a posteriori e identificadores. Em seguida, foram excluídas ainda as que possuíssem 100% de frequência, ou seja, com variância próxima a 0. Variáveis assim tornam-se desprezíveis aos algoritmos de classificação.

Considerando a análise exploratória do capítulo anterior, pôde-se perceber que a presença de alta correlação entre algumas variáveis independentes e a baixa variância poderiam ser melhor tratadas a fim de se obter um resultado mais profícuo do processo de mineração. Além dessas

propriedades, a presença de valores ausentes, também passou a ser um problema, ao passo que, apesar de poucas colunas terem os chamados, em inglês, *missing values,* algumas dessas continham um número elevado.

Nesse sentido, buscando entender em pesquisas semelhantes a abordagem frente a esses três fatores, percebeu-se que, apesar de serem frequentemente tratados, não está claro na literatura, devido às particularidades de cada domínio, um limite máximo tolerável para nenhum desses problemas.

Diante disso, ocorreram as seguintes questões para o conjunto de dados: i) Qual o melhor grau limite de correlação entre as variáveis independentes? ii) Qual a porcentagem máxima de *missing values* aceitável para o modelo? E iii) qual a grau de variância mínimo?

Buscando respostas, resolveu-se sistematizar um processo supervisionado baseado em (SILIPO et al., 2014). Foram escolhidas 3 técnicas de classificação, *Nayve bayes*, Redes Neurais e Árvores de decisão, a fim de encontrar os melhores valores de corte para serem usados nos filtros para todas as questões colocadas anteriormente.

Um subconjunto de dados (34%) foi separado para a análise de redução de dimensionalidade. Variáveis já consideradas importantes foram separadas e não participaram do processo, que foi esquematizado pra cada uma dos três fatores da seguinte forma:

- **Valores ausentes*:*** através de um processo iterativo, com valores de corte v iniciando de 0 a 100, com *step* igual a 1, todas as técnicas de classificação eram executadas sob os dados de todas as colunas com até $v\%$ de valores ausentes. Para cada iteração, o valor da melhor métrica AUC_ROC, dentre as três técnicas executadas, era armazenado. Ao final do processo iterativo, o valor v, que tivesse relacionado à melhor métrica, era escolhido como taxa máxima aceitável de valores ausentes para todas as variáveis independentes.

- **Baixa variância**: através de um processo iterativo, com valores v iniciando de 0,001 a 0,132 (variância máxima da base) com *step* de 0,01, todas as técnicas de classificação eram executadas sob os todos os dados das colunas numéricas com variância mínima de $v\%$. Para cada iteração, o valor da melhor métrica AUC_ROC, dentre as três técnicas executadas, era armazenado. Ao final do processo iterativo, o valor v, que tivesse relacionado à melhor métrica, era escolhido como valor mínimo de variância para todas as variáveis independentes. Vale observar que este método é aplicado apenas sob variáveis numéricas, que por sua vez, foram normalizadas para tornar a variância independente do domínio da variável.

- **Alta correlação**: através de um processo iterativo, como nas outras duas, com valores v iniciando de 0,1 a 0,99 com step de 0,1, as técnicas de classificação eram executadas sob os todos os dados com correlação de até v% (para o par de variáveis com correlação fora do limite v, apenas uma era escolhida para permanecer). Assim como nas outras duas técnicas, o valor v, associado à melhor métrica AUC_ROC, era escolhido como valor máximo de correlação entre as variáveis independentes. Observa-se que, para o cálculo da correlação utilizou o coeficiente de *Pearson* para as variáveis categóricas e o valor Chi-quadrado para as variáveis numéricas. Nenhum coeficiente foi definido entre variáveis numéricas e categóricas.

Os gráficos presentes nas Figuras 22, 23, 24 demonstram a variação do coeficiente AUC_ROC para cada uma dos filtros durante o processo iterativo. Observa-se que ao alterar os valores dos limites e consequentemente o número de colunas envolvidas, o desempenho dos modelos gerados também são afetados.

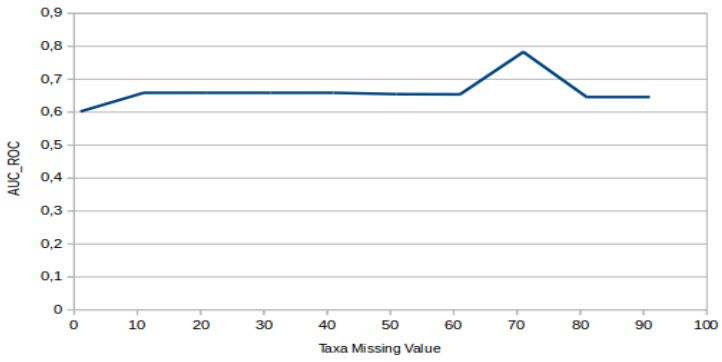

Figura 22: Gráfico - filtro para valores ausentes

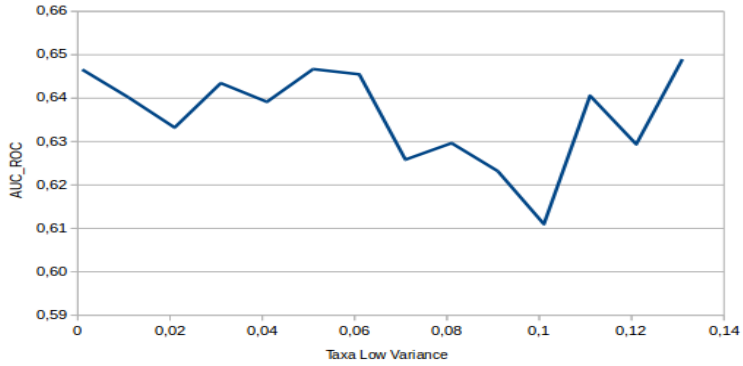

Figura 23: Gráfico - filtro para baixa variância

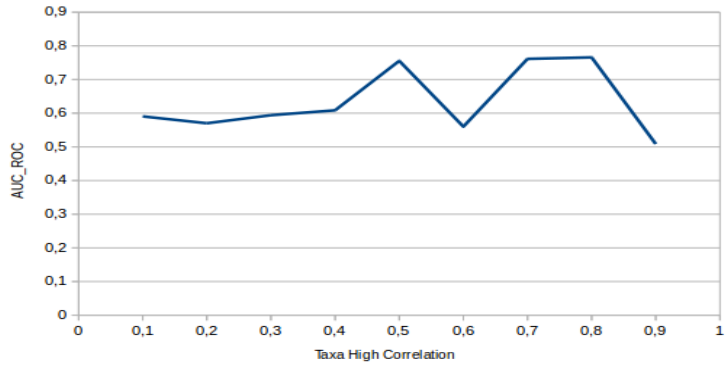

Figura 24: Gráfico - filtro para alta correlação

A Tabela 4 mostra um resumo do processo de redução de dimensionalidade. O filtro relativo aos valores ausentes obteve um melhor desempenho, além de ter causado uma maior redução nos dados. Já o de baixa variância, apesar de reduzir consideravelmente o volume de dados, não obteve um bom desempenho. As variáveis afetadas por esse filtro são sensíveis à qualidade do modelo, conforme mostra o gráfico da Figura 24. Para encontrar uma combinação que pudesse manter o melhor custo-benefício entre redução e desempenho, os melhores limites de cada filtro foram testados de maneira sequencial. Os filtros de valores ausentes e alta correlação, Tabela 4, alcançaram os melhores resultados. O Apêndice D exibe todas as colunas retiradas pelos filtros adotados.

Tabela 4: Resultados redução dimensionalidade

Filtro de Redução	Redução	AUC_ROC	Melhor Limite
Nenhuma	0%	0,64	-
Missing Value	18%	0,78	71
Low Variance	15%	0,65	0,071
High Correlation	5,3%	0,77	0,8
Missing Value + High Correlatiom	19%	0,78	
Missin Value + Low Filter	31,5%	0,62	
Low Filter + High Correlation	18%	0,52	
Missing Value + High Correlatiom+ Low Filter	35%	0,59	

Percebeu-se também, através da exploração das variáveis, que 6 (seis) atributos do questionário socioeconômico do ENEM inerentes ao exercício de emprego, possuíam praticamente a mesma taxa de valores ausentes (~80%). Observando que essas ausências eram respectivas a quantidade de pessoas que responderam nunca ter trabalhado em outra questão do questionário, as que ainda não haviam sido retiradas pelo processo supervisionado, foram desconsideradas.

É importante observar que técnicas que reduzem a capacidade interpretativa dos dados, como PCA, não foram consideradas. Pois fogem ao objetivo do trabalho.

4.9 Transformação dos dados

Durante o processo de transformação, os dados foram modificados ou consolidados de uma maneira que se tornassem mais apropriados para o processo de mineração (J.HAN, J.PEI, M.KAMBER, 2012). Nesse sentido e, considerando que o conjunto de dados ideal para o processo de mineração depende da técnica a qual irá ser submetido, foram aplicadas diferentes transformações pra cada um dos classificadores..

Para aplicação da técnica de indução de regras, as variáveis numéricas foram inseridas em quatro grupos por meio da discretização por frequência. Devido a distribuição da variável “idade” e a presença de *outltiers*, conforme ilustrado nas Figuras 15 e 16, a variável teve seus valores separados em apenas três grupos.

Outra transformação realizada foi a normalização das variáveis numéricas para a aplicação da técnica de regressão. O método utilizado foi o *Min-max normalization*[7], disponível no KNIME,

7Método de transformação linear implementado na ferramenta Knime. <https://www.knime.org/files/nodedetails/_manipulation_column_column_transform_Normalizer.html>

que transforma de maneira linear os valores das variáreis, colocando-os entre 0 e 1 através da seguinte fórmula:

$$\frac{(Varíável - MÍNIMO(Variável))}{(MÁXIMO(Variável) - MÍNIMO(Variável))}$$

Equação 4: Fórmula de normalização

Para todas as técnicas, os valores ausentes das variáveis numéricas foram preenchidos com o valor médio. Para as variáveis categóricas, foram separadas as que possuíam elevada taxa de valores ausentes, e através de uma análise de domínio, os valores foram preenchidos. Um bom exemplo é a "q031"(Você deixou de estudar o ensino médio?) em que foi preenchida a resposta NÃO. Pois, além de ter sido a opção escolhida por 94% dos que responderam, 74% dos alunos do conjunto de dados possuem idade regular para o curso do ensino médio (até 17 anos). Para que a resposta seja SIM, esperava-se uma idade mais avançada, devido à necessidade de abandono e posterior regresso do aluno ao ensino médio.

As variáveis que possuíam uma elevada taxa de ocorrência da moda, tiveram este valor preenchido para os valores ausentes. Em alguns casos, o valor "AUSENTE" foi imputado.

Para todas as técnicas, atributos numéricos, com exceção dos índices gerados pelo MEC na base ENEM POR ESCOLA, *outliers* foram identificados pela dispersão de três vezes o desvio padrão, os quais foram substituídos pelo valor extremo.

Todo o processamento dos dados feito antes da submissão às técnicas de mineração de dados foi estruturado em um fluxo, a fim de minimizar possíveis problemas quando da aplicação em novos conjuntos de dados. Uma versão simplificada do fluxo pode ser observada na Figura 25.

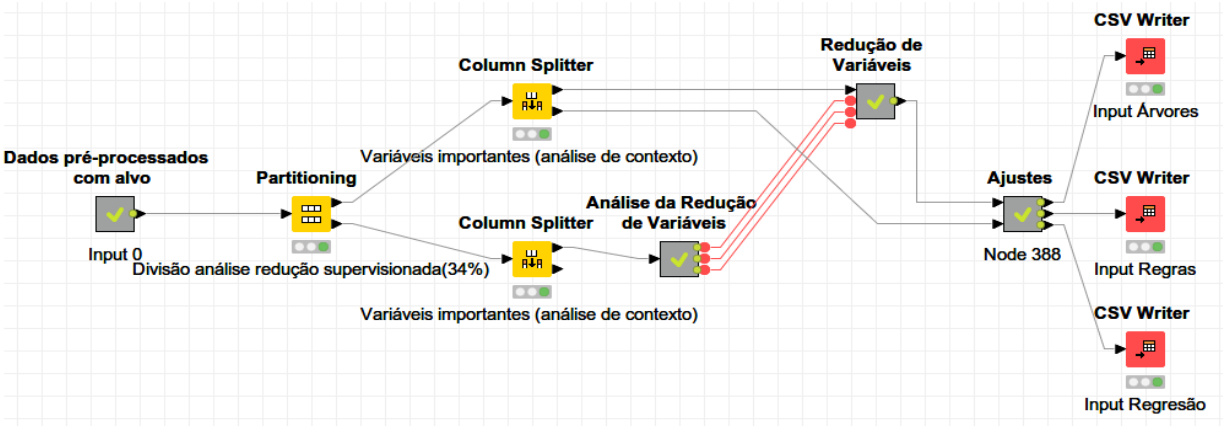

Figura 25: Fluxo simplificado do processamento dos dados

5 EXTRAÇÃO DO CONHECIMENTO E INTERPRETAÇÃO DOS RESULTADOS

Seguindo a fase da metodologia CRISP-DM, este capítulo exibe os resultados das técnicas de modelagem que foram propostas no capítulo 3. Por se tratar de uma fase iterativa, é comum a repetição para ajustes dos parâmetros, a fim de obterem melhores resultados (FAYYAD et al., 1996). Nesta seção também é apresentada a análise dos resultados da pesquisa, sendo este capítulo o mais importante desse trabalho.

5.1 Árvore de Decisão

As árvores de decisão foram construídas por meio do algoritmo J48, que é uma versão otimizada do C4.5 (QUINLAN, 1993) implementada no pacote de software *Weka*. Justifica-se essa escolha devido ao algoritmo ser o mais utilizado e conhecido na sua categoria, sendo amplamente incorporado em ferramentas de mineração de dados (KIANG, 2003).

Utilizando o método "dividir para conquistar", o C4.5 é chamado com três parâmetros: "*D",* partição de dados que inicialmente é compreendida por todos os atributos do conjunto de treino com suas respectivas classes; "Lista_Atributos", valores que descrevem os atributos da partição "D" e "Método_Seleção_Atributo", que especifica o procedimento heurístico usado para revelar o atributo que melhor define a amostra.

Dessa maneira, em cada nó da árvore, o algoritmo escolhe através do seu Método_Seleção_Atributo, o atributo que mais particiona o seu conjunto de amostra "D" em novos subconjuntos que tenderão a uma categoria ou outra. Esse procedimento pode ser implementado com base na medida de ganho de informação ou índice de Gini (J.HAN, J.PEI, M.KAMBER, 2012).

A estrutura final da Árvore de Decisão é formada por uma sequência hierárquica em forma de árvore invertida da raiz para as folhas. As folhas das árvores representam regras que explicam o conhecimento embutido nos dados de forma humanamente compreensível por regras "se-então" (SAFAVIAN; LANDGREBE, 1991).

Com o objetivo de construir uma árvore com bom desempenho e com um tamanho que facilite a análise dos especialistas do domínio, os 14.152 exemplos caracterizados por 112 atributos

foram submetidas ao processo de 10-*cross-validation,* sem a opção de poda. Além disso, o número mínimo de registros por folha foi o valor 708, que representa 5% do conjunto de dados.

A Figura 26 ilustra parte de um ramo da árvore, indicando a representatividade da regra (*i.e.* suporte), seguida da concentração de alunos pertencentes à classe alvo (primeiro quartil) (*i.e.* confiança), ambas expressas em porcentagem, e a razão entre a confiança da regra e a população (*i.e. lift*).

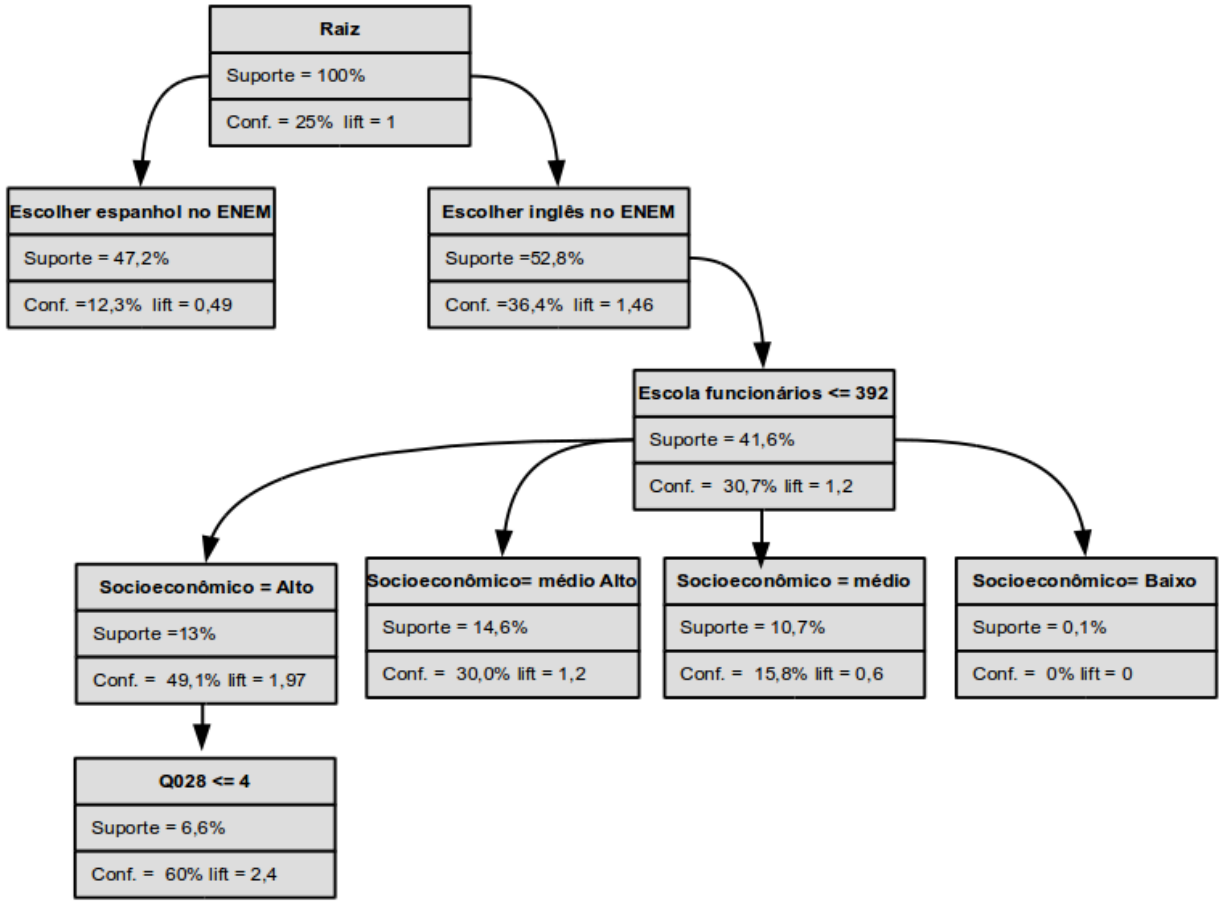

Figura 26: Ramificação da Árvore de Decisão

atributos mais importantes. Nota-se uma maior concentração de bons alunos entre os que optaram por língua inglesa na prova do ENEM. Aspectos que representam o tamanho da escola e fatores socioeconômicos também influenciam de maneira direta o desempenho dos estudantes.

5.2 Regras de Classificação

A indução de regras também gera regras do tipo se-então, porém, “pepitas do conhecimento” que passariam indetectáveis para as Árvores de Decisão, podem ser encontradas. A indução de regras não particiona o espaço de entrada e nem pondera sua força pelo seu suporte (ADEODATO, 2016).

Para a aplicação da técnica indução de regras, o input de dados construído, conforme discutido no capítulo 4, foi submetido ao algoritmo JRIP, uma versão otimizada do algoritmo IREP (COHEN, 1995) e PART, ambas sem a opção de poda. Por meio da técnica 10-cross-validation, com a restrição de número mínimo de instâncias por regras em 1% do conjunto de dados, os algoritmos JRIP e PART geraram 7 e 10 regras, respectivamente.

As três regras com maior e menor *lift* são exibidas na Tabela 5, em que pode-se observar, novamente, a influência direta de atributos econômico-financeiros. Essas características aparecem, sejam ligadas às condições do aluno, como ter estudado em escola particular durante todo o ensino fundamental, sejam ligadas à estrutura da escola. O índice de titulação docente aparece associado aos atributos socioeconômicos como condição para o sucesso do aluno. A regra com maior *lift* com 5 condições, reflete bem a concentração de bons alunos no grupo de estudantes com melhores condições financeiras e que estudam em escolas mais estruturadas. A regra explicita a baixa intenção do candidato na contemplação de uma bolsa do Prouni[8], a importância da titulação dos professores e da estrutura da escola, uma vez que são nas capitais que situam os IFs com maiores investimentos.

De maneira contrária, a regra com segundo menor *lift*, 0,68, explicita uma condição divergente ao observado pelos especialistas de domínio, merecendo ser maior investigada. A regra relaciona ao grupo de alunos com baixo desempenho àqueles que estudam em escolas que possuem salas de professores e alojamento estudantil. O alojamento estudantil é visto como um importante aspecto da assistência ao educando, auxiliando na permanência de alunos de baixa renda na escola. Já a sala de professores é um requisito importante para a ampliação do atendimento extra-classe dos alunos.

8 – É o programa do Ministério da Educação que concede bolsas de estudo integrais e parciais de 50% em instituições privadas de educação superior, em cursos de graduação e sequenciais de formação específica, a estudantes brasileiros sem diploma de nível superior. <http://siteprouni.mec.gov.br/>

Tabela 5: Maiores e menores *lift*

Regras	Algoritmo	Confiança	Suporte	*Lift*
Escolher Inglês, socioeconômico alto, professores com o índice de titulação de 0,3 a 0,5, estudar na capital e não desejar bolsa ProUni.	Jrip	88,3%	1,27%	3,53
Escola não ter sala de professor e socioeconômico do aluno alto	PART	83,3%	2,12%	3,33
Escolher inglês, estudar em capital, ter estudado o ensino fundamental somente em escola particular e escola reciclar lixo.	Jrip	83,2%	1,34%	3,33
Socioeconômico médio	PART	19,7%	1,60%	0,79
Escola possuir sala de professor e alojamento de alunos	PART	17,1%	1,78%	0,68
Escolher espanhol	PART	12,3%	5,81%	0,49

5.3 Regressão Logística

A Regressão Logística foi a técnica utilizada para a produção dos índices de propensão de sucesso dos alunos no ENEM. A técnica também é capaz de identificar e quantificar os principais atributos que influenciam na presença desses alunos no quartil superior de notas.

Para a aplicação da técnica, a amostra de dados foi dividido em dois subconjuntos, treinamento e testes. O subconjunto de treinamento, com 66% dos dados, foi submetido ao método *forward stepwise* para a extração do conhecimento. Amplamente utilizado em modelos de regressão logística, o método consiste em um processo que adiciona ou exclui variáveis a cada etapa, baseado em um critério que otimize o modelo, reduzindo a variância e evitando problemas de multicolinearidade (KUTNER et al., 2004). Para este trabalho o critério de informação Akaike (AIC)[9] foi utilizado.

Para as variáveis categóricas foram geradas variáveis *dummies,* utilizando-se da última categoria como referência. Dessa forma, para cada categoria gerou-se uma nova variável, que, por definição, assumiu os valores 1 ou 0.

Das 112 variáveis independentes iniciais, somadas a $K-1$ *dummies* para cada variável categórica de K níveis, foram incluídas apenas 57 variáveis no modelo de regressão, de acordo com o método *forward stepwise.,*

Os índices de propensão dos dez melhores atributos, que influenciam de maneira negativa e positiva o sucesso dos alunos, para o nível de significância de 5%, estão dispostos na Tabela 6.

9 – Métrica de qualidade apresentada em: (AKAIKE, 1976)

Percebe-se, assim como na árvore de decisão e no modelo de regras, a grande influência de aspectos socioeconômicos dos alunos e estruturais da escola.

Vale ressaltar que outros atributos aparecem somente na regressão, como o tempo que o aluno levou para cursar o Ensino Fundamental. O modelo mostra também uma vantagem dos alunos da região sudeste, bem como de quem presta o exame com o objetivo de ingressar na Educação Superior Pública. É importante destacar ainda, uma maior concentração de atributos do grão aluno entre os mais preditivos. No Apêndice E, estão dispostos todos os atributos com nível de significância menor que 5%.

Tabela 6: Atributos mais relevantes do modelo de Regressão Logística

Atributo	Beta	*p-value*	Natureza	Fonte
Tem como objetivo ingressar no ensino superior público	1,56	0,00	Numérica	Aluno/Enem
Escola Localizada em unidade de uso sustentável	1,27	0,01	Categórica	Escola/Censo
Escola situada na região Sudeste	1,14	0,00	Categórica	Escola/Criada
Índice de Permanência na Escola = de 20% a 40%	1,01	0,00	Categórica	Aluno/Enem
Esgoto sanitário inexistente	0,96	0,02	Binária	Escola/Censo
Índice Nível Socioeconômico = médio baixo	-1,25	0,00	Categórica	Aluno/Enem
Levou mais de 11 anos para concluir o Ensino Fundamental	-1,25	0,02	Categórica	Aluno/Enem
Quantas pessoas moram em sua casa?	-1,42	0,01	Numérica	Aluno/Enem
Cor/Raça Indígena	-1,54	0,04	Categórica	Aluno/Enem
Índice Nível Socioeconômico = Baixo	-2,60	0,00	Categórica	Aluno/Enem

Como a solução de KDD proposta ainda não possui um uso específico em um cenário de suporte a decisão, optou-se em avaliar a performance do modelo concentrando nas características do mesmo, sem considerar um único limiar de decisão (ADEODATO, 2016). Posto isso, para a avaliação do modelo treinado, foi executada a previsão do conjunto de testes com 4.812 registros, e, de posse dos respectivos escores de propensão, foi gerada a curva ROC para o cálculo da área sob a curva (AUC_ROC).

A métrica AUC_ROC é considerada uma das métricas de performance mais aceitas para avaliação de modelos de classificação binária (PROVOST; FAWCETT, 2001) e apresentou, neste trabalho, um valor de 0,84 para a curva ilustrada na Figura 27.

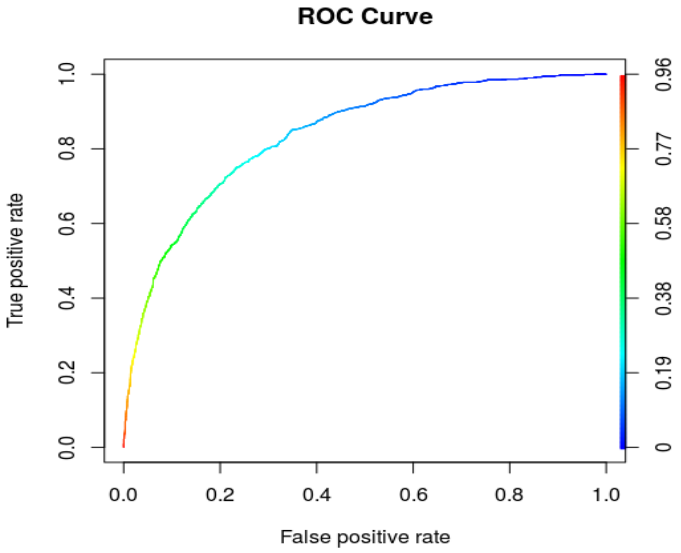

Figura 27: Curva ROC para o modelo de regressão logística

Outra métrica utilizada para avaliação do modelo foi a Max_KS2, que consiste no maior valor da curva obtida por meio da diferença entre as funções de distribuição acumuladas dos alvos 1 e 0 - Teste de Kolmogorov-Sminrov. A representação das curvas das funções acumuladas e do teste KS2, que atingiu a distância máxima no ponto 0,51, pode ser observada na Figura 28.

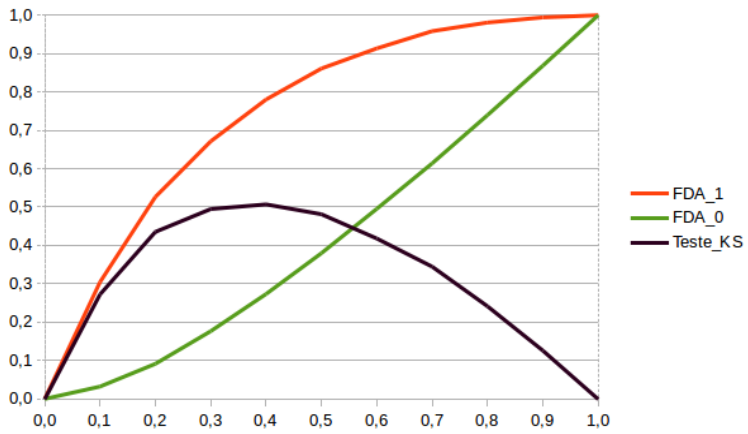

Figura 28: Curvas acumuladas e teste KS2

Gerando a matriz de confusão da aplicação do modelo treinado no subconjunto de testes para a técnica de regressão logística, verificou-se que a mesma classificou corretamente 39.25 casos e errou 887, obtendo uma acurácia de 0,82. Dentre os casos que o modelo errou, a maioria fora situações em que o alvo previsto foi 1, porém o alvo real era 0. A Tabela 7 exibe a matriz de confusão com maiores detalhes.

Tabela 7: Matriz de confusão regressão logística

	Positivo atual	Negativo atual
Predição positiva	**612**	297
Predição negativa	590	**3313**

6 CONCLUSÃO

Neste capítulo são apresentadas as conclusões dessa pesquisa, discorrendo acerca da essência e da natureza de todo o trabalho apresentado nos capítulos anteriores. Também são citadas as principais contribuições científicas, limitações do trabalho além de alguns pontos que extrapolam o escopo do que foi proposto, servindo como referência para a expansão da pesquisa em trabalhos futuros.

6.1 Resumo

Essa dissertação analisou e extraiu conhecimento das bases de dados do ENEM e do Censo Escolar de 2014, sobre a qualidade do ensino médio brasileiro integrado à educação profissional. A oferta do Ensino Profissional integrado ao Ensino Médio, por meio dos institutos federais, tem tido um grande crescimento nos últimos anos mediante o alto investimento do governo federal.

Os desempenhos dos alunos oriundos dessas instituições no ENEM são muitas vezes maiores que os de outras escolas, inclusive, escolas particulares. Considerando a grande importância do ENEM na atualidade e a variedade de informações encontradas em suas bases de dados foi sistematizado um modelo, baseado em mineração de dados, que pudesse auxiliar a construção de mecanismos que subsidiem os gestores, educadores e especialistas no processo de tomada de decisões estratégicas e validação de políticas públicas.

Este trabalho utilizou o conhecimento de especialistas do domínio publicados na literatura, e por meio da metodologia CRISP-DM, aliada à D³M, forneceu um modelo capaz de avaliar e prever o desempenho dos alunos que se encontram no último ano do Ensino Médio dos Institutos Federais.

Os alunos foram classificados em duas diferentes classes a partir da média das notas das competências do ENEM e da redação. O quartil superior foi utilizado como limiar de desempenho, possibilitando a aplicação de técnicas de classificação binária. A técnica de regressão logística gerou indicador de chances de sucesso/insucesso dos alunos no ENEM além de identificar fatores importantes que influenciam no futuro do estudante. Técnicas convencionais de avaliação de modelos binários, AUC_ROC e KS_MAX, foram utilizadas para medir o desempenho dos índices de propensão gerados pela técnica de regressão. A árvore de decisão e a indução de regras explicitaram em linguagem natural as condições que influenciam no desempenho dos alunos e teve as métricas confiança, suporte e *lift* avaliadas.

6.2 Contribuições

Considerando a abrangência e importância atual do ENEM e tomando-o como ponto de partida para avaliação do Ensino Médio brasileiro, considera-se a construção de um modelo preditivo, baseado em MD, como uma das principais contribuições dessa dissertação. O modelo foi capaz de estimar as chances de sucesso/insucesso dos alunos dos IFs no ENEM identificando os principais fatores que influenciam o desempenho desses estudantes.

O conhecimento extraído neste trabalho por meio de técnicas de mineração do tipo "caixa-branca" favorecem a construção de um sistema de suporte a decisão de boa qualidade e de fácil entendimento aos especialistas de domínio. Para cálculo dos escores de propensão para o bom desempenho foi utilizada a técnica de Regressão Logística. A indução de regras foi utilizada visando à obtenção de "pepitas do conhecimento" e explicação da propensão e árvores de decisão para explicar como seria a sequência decisória na visão de um especialista humano.

Como aspectos mais relevantes trazido pelos modelos, os fatores socioeconômicos apareceram como um dos principais influenciadores no desempenho dos estudantes. Ocorrendo de maneira direta, por meio de atributos presentes no grão aluno, ou ligados à estrutura da escola. Outros atributos, como formação do professor, opção pela língua inglesa, tempo de permanência na mesma escola e perspectivas dos alunos após o ingresso no ensino superior, também foram relevantes. É importante destacar, ainda, uma maior concentração de atributos do grão aluno entre os mais preditivos e a vantagem dos alunos da região sudeste.

Aspectos vistos pelos especialistas em educação como importantes, como moradia estudantil na escola e estrutura de salas para os professores, aparecem como fatores de insucesso e precisam ser melhor investigados.

Além disso, o trabalho demonstrou a utilização de técnicas de manipulação de dados que levaram a construção de um *data-mart* no grão aluno no âmbito dos Institutos Federais a partir das bases de dados abertas da educação brasileira do ano de 2014. Essa sistematização possibilita futuras consultas OLAP e reuso para construção de novos modelos de classificação. Todo o fluxo do processo, desde a integração das bases e pré-processamento foi sistematizado e poderá ser reaplicado a novos dados de outras versões do ENEM e Censo Escolar, minimizando assim o esforço técnico para que a solução proposta seja utilizada em ambientes de produção.

6.3 Limitações

Das limitações do trabalho, destacam-se as referentes aos dados, principalmente a inexistência de um atributo que interligasse as bases do ENEM às do Censo no grão aluno.Tal chave possibilitaria a exploração de dezenas de atributos do Censo Escolar presentes no grão aluno, agregando ao *data-mart* mais informações individuais dos estudantes.

Uma outra limitação é a grande quantidade de atributos presentes nas bases de dados. Apesar de o presente trabalho ter se preocupado em eliminar os atributos menos relevantes, ainda restaram 112 variáveis no conjunto de dados final. A grande quantidade de atributos aumenta a complexidade dos modelos e dificulta, inclusive, a aplicação técnicas de mineração de dados que exigem maior poder de processamento e que passariam a ser incompatíveis com o *hardware* disponível para a pesquisa.

6.4 Trabalhos Futuros

Com os objetivos propostos do trabalho alcançados, estabelecem-se algumas recomendações para trabalhos futuros, de modo que os objetivos possam ser refinados, estendidos, ou até modificados e as deficiências sejam sanadas.

Considerando a continuidade da dissertação, entende-se como principal sugestão de trabalho futuro, a implementação de um sistema de informação que incorpore o processo de construção dos modelos preditivos propostos, de maneira que facilite a utilização do conhecimento extraído pelos especialistas de domínio.

Considerando a contribuição dessa dissertação quanto a classificação dos alunos em bons e ruins e todo o escopo de KDD apresentado, entende-se também como trabalho futuro a replicação do conjunto ferramental apresentado em uma base de dados de um sistema acadêmico de uma específica instituição. O estudo poderia priorizar alunos com baixo potencial para fins de intervenção.

Outros desafios pertinentes também se apresentam da seguinte maneira:

- Implementar novas técnicas de mineração, a fim de comparação dos resultados, inclusive abordagens do tipo “caixa-preta”.
- Realizar um estudo com diversas técnicas de mineração a fim de identificar os dados mais relevantes do ENEM e Censo Escolar para comporem o *data-mart* de maneira a alcançar resultados mais satisfatórios.

- Repetir o estudo utilizando as bases de dados dos anos posteriores a 2014, para que possa ser medida a performance desses modelos com novos dados reais.

REFERÊNCIAS

ADEODATO, P. J. L. DATA MINING SOLUTION FOR ASSESSING BRAZILIAN SECONDARY SCHOOL QUALITY BASED ON ENEM AND CENSUS DATA. , p. 2658–2679, 2016.

AKAIKE, H. An information criterion (AIC). **Math Sci**, v. 14, n. 153, p. 5–9, 1976.

ALMEIDA FILHO, Á. C. DE. Modelo De Mensuração Do Desempenho Dos Institutos Federais : Uma Análise a Partir De Microdados Modelo De Mensuração Do Desempenho Dos Institutos Federais : Uma Análise a Partir De Microdados. , 2014.

AMENDOEIRA, A.; SONIA, N.; DE, X.; et al. RBEP ESTUDOS Indicadores de qualidade do ensino fundamental: o uso das tecnologias de mineração de dados e de visões multidimensionais para apoio à análise e definição de políticas públicas *. **Rev. bras. Estud. pedagog**, v. 94, n. 238, p. 677–700, 2013.

ARAQUE, F.; ROLDÁN, C.; SALGUERO, A. Factors influencing university drop out rates. **Computers and Education**, v. 53, n. 3, p. 563–574, 2009.

ARAÚJO, C. H.; LUZIO, N. Avaliação da educação básica: em busca da qualidade e eqüidade no Brasil. **Inep/MEC - Instituto Nacional de Estudos Educacionais Anísio Teixeira**, p. 71, 2005.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, p. 3, 2011. Disponível em: <http://br-ie.org/pub/index.php/rbie/article/view/1301%5Cnhttp://www.br-ie.org/pub/index.php/rbie/article/view/1301>. .

BAKER, R. S. J. D.; YACEF, K. The State of Educational Data Mining in 2009 : A Review and Future Visions. **Journal of Educational Data Mining**, v. 1, n. 1, p. 3–16, 2009.

BARROS, H. R.; ADEODATO, P. J. L. A Data Mining Approach for Preventing Undergraduate Students Retention. **WCCI IEEE World Congress on Computational Intelligence**, p. 10–15, 2012. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6252437>. .

BERNARDIM, M. L.; SILVA, M. R. DA. Políticas Curriculares para o Ensino Médio e para a Educação Profissional : propostas , controvérsias e disputas em face das proposições do Documento Referência da Conae 2014 Curricular Policies to High School and Professional Education : **Jornal De Políticas Educacionais**, v. 16, p. 23–35, 2014.

BITTENCOURT, H. R. Regressão logística politômica : revisão teórica e aplicações. **ACTASCIENTIAE**, v. 5, p. 77–86, 2003.

BRASIL. DECRETO Nº 2.208, DE 17 DE ABRIL DE 1997. , 1997. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/D2208.htm>. .

BRASIL. DECRETO Nº 5.154 DE 23 DE JULHO DE 2004. , 2004. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/decreto/D5154.htm>. .

BRASIL. LEI Nº 11.892, DE 29 DE DEZEMBRO DE 2008. , 2008. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm>. .

BRASIL. LEI Nº 13.005, DE 25 DE JUNHO DE 2014. , 2014. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm>. .

BRASIL. A Expansão da Rede Federal. , 2016. Disponível em: <http://portal.mec.gov.br/component/content/article?id=20015:rede>. .

CAO, L. Introduction to domain driven data mining. **Data Mining for Business Applications**, p. 3–10, 2009. Disponível em: <https://www-staff.it.uts.edu.au/~lbcao/publication/dmba-dddm.pdf>. .

CAO, L.; LIN, L.; CHENGQI, Z. Domain Driven in Depth Pattern Discovery: APractical Methodology. **Proceedings 4th Australasion Data Mining Conference AusDM05**, v. 6, p. 101–114, 2005. The University of Technology, Sydney. Disponível em: <http://hdl.handle.net/10453/1903>. Acesso em: 10/3/2017.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; et al. **CRISP-DM 1.0 Step-by-step data mining guide**. 2000.

CLEARINGHOUSE, E. Teacher Quality and Student Achievement: A Review of State Policy Evidence Linda Darling-Hammond Stanford University. **Quality**, v. 8, n. 1, p. 1–48, 2000.

COHEN, W. W. Fast Effective Rule Induction. In: A. Prieditis; S. Russell (Eds.); **Machine Learning Proceedings 1995**. p.115–123, 1995. San Francisco (CA): Morgan Kaufmann. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9781558603776500232>. .

DEOGUN, J. S.; RAGHAVAN, V. V; SARKAR, A.; SEVER, H. Data Mining: Trends in Research and Development. **Rough Sets and Data Mining: Analysis of Imprecise Data**. p.9–45, 1997. Boston, MA: Springer US. Disponível em: <http://dx.doi.org/10.1007/978-1-4613-1461-5_2>. .

DORNELES, R. P. Avaliação da educação profissional: um estudo sobre indicadores educacionais específicos. **Unb**, p. 139, 2011.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

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

FONSECA, S. O. DA; NAMEN, A. A. Mineração Em Bases De Dados Do Inep: Uma Análise Exploratória Para Nortear Melhorias No Sistema Educacional Brasileiro. **Educação em Revista**, v. 32, n. 1, p. 133–157, 2016. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982016000100133&lng=pt&nrm=iso&tlng=en>. Acesso em: 20/10/2016.

GOLDHABER, D. D.; BREWER, D. J. Evaluating the Effect of Teacher Degree Level on Educational Performance. , 1996.

GONÇALVES JR, W. P.; BARROSO, M. F. AS QUESTÕES DE FÍSICA E O DESEMPENHO DOS ESTUDANTES NO ENEM PHYSICS ITEMS AND STUDENT'S PERFORMANCE AT ENEM. , 2013.

GUERRA, P. C.; YUJI, R.; NAKAMURA, M.; HRUSCHKA, E. R. Estimativa de Demanda Potencial de Matrículas em Ensino Superior usando Dados Públicos e Múltiplos Modelos de Regressão. , v. 2, 2014. Disponível em: <http://www.producao.usp.br/handle/BDPI/48650>. .

GURULER, H.; ISTANBULLU, A.; KARAHASAN, M. A new student performance analysing system using knowledge discovery in higher educational databases. **Computers and Education**, 2010.

HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for data mining. **… and Data Engineering, IEEE Transactions …**, v. 15, n. 6, p. 1437–1447, 2003. Disponível em: <http://researchcommons.waikato.ac.nz/handle/10289/1026%5Cnhttp://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1245283%5Cnhttp://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1245283>. .

HELENA, M.; CASTRO, G. DE. A reforma do ensino médio e a implantação do Enem no Brasil 1. , 2003.

HOSMER, D. W.; LEMESHOW, S. Applied regression analysis. **New York, John Willey**, 1989.

HUYSMANS, J.; DEJAEGER, K.; MUES, C.; VANTHIENEN, J.; BAESENS, B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. **Decision Support Systems**, v. 51, n. 1, p. 141–154, 2011.

INEP. NOTA EXPLICATIVA ENEM 2014 POR ESCOLA. , 2015a. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2014/nota_explicativa_enem_2014_por_escola.pdf>. Acesso em: 14/3/2017.

INEP. NOTA TÉCNICA Brasília, 05 de agosto de 2015 INDICADOR DE PERMANÊNCIA NA ESCOLA (ENSINO MÉDIO). , , n. 61, p. 2010–2015, 2015b.

INEP. Nota Técnica - Indicador de Nível Socioeconomico (Inse) das Escolas. , 2015c.

INEP, P. No Title. , 2011. Disponível em: <http://portal.inep.gov.br/>. .

J.HAN, J.PEI, M.KAMBER. **Data Mining: Concepts and Techniques**. 2012.

JOSÉ, A.; ARAÚJO, N. **Ensino Profissionalizante de Nível Médio e seus Efeitos sobre Desempenho Escolar e Inserção Produtiva: uma análise recente a partir de dados do Censo Escolar e ENEM**, 2014. Universidade Federal de Juíz de Fora - UFJF.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. An overview of free software tools for general data mining. .

K.A, R. J. C.; H.M., D.; MAHER. Product appearance inspection methods and apparatus employng low variance filter. **US Patent 523762**, 1993.

KALEGELE, K.; TAKAHASHI, H.; SVEHOLM, J.; et al. On-demand data numerosity reduction for learning artifacts. **Proceedings - International Conference on Advanced Information Networking and Applications, AINA**, , n. Ml, p. 152–159, 2012.

KAMPFF, A. J. C. **Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente .**, 2009. Universidade Federal do Rio Grande do Sul. Disponível em: <http://www.lume.ufrgs.br/handle/10183/19032>. .

KIANG, M. Y. A comparative assessment of classification methods. **Decision Support Systems**, 2003.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1–2, p. 273–324, 1997. Disponível em: <http://www.sciencedirect.com/science/article/pii/S000437029700043X>. .

KRAWCZYK, N. O Ensino Médio no Brasil. **Ação Educativa**, v. Em questão, p. 1–48, 2009. São Paulo. Disponível em: <http://www.bdae.org.br/dspace/bitstream/123456789/2342/1/emquestao6.pdf>. .

KUENZER, A. Z. O Ensino Médio agora é para a vida : Entre o pretendido , o dito e o feito. **Educação & Sociedade**, , n. 70, p. 15–39, 2000. Disponível em: <http://www.scielo.br/pdf/es/v21n70/a03v2170.pdf>. .

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. **Applied Linear Statistical Models (McGraw-Hill/Irwin Series Operations and Decision Sciences)**. McGraw-Hill/Irwin, 2004.

LIMA, M.; MENDES, I.; SILVA, D. ENSINO MÉDIO INTEGRADO NO ESPÍRITO SANTO: PERSPECTIVAS DO DEBATE ACERCA QUALIDADE A PARTIR DOS RESULTADOS DO DESEMPENHO DE ESTUDANTES NO ENEM. , 2013.

MAHAPATRA, B. Data Reduction in MANETs using Forward Feature Construction Technique. **International Conference on Man and Machine Interfacing (MAMI)**, p. 0–2, 2015.

MANHÃES, L. M. B. Predição Do Desempenho Acadêmico De Graduandos Utilizando Mineração De Dados Educacionais. , p. 140, 2015. Disponível em: <https://pdfs.semanticscholar.org/1829/75e815afa20cce5cdd9fbd164d80cd603f47.pdf>. .

MARTINS; PAULA, A. Pressupostos de Gramsci na educação profissional e tecnológica de nível médio. **#Tear: Revista de Educação, Ciência e Tecnologia**, v. 1, n. 2, 2012.

MCCUE, C. **Data mining and predictive analysis: Intelligence gathering and crime analysis**. Butterworth-Heinemann, 2014.

MEC, P. P. , 2013. Disponível em: <http://portal.mec.gov.br/component/tags/tag/32123>. .

MOHAMAD, S. K.; TASIR, Z. ScienceDirect Educational data mining: A review. **Procedia - Social and Behavioral Sciences**, v. 97, p. 320–324, 2013.

MONTGOMERY, D. C.; RUNGER, G. C.; HUBELE, N. F. **Engineering statistics**. John Wiley & Sons, 2009.

PACHECO, E. **Institutos Federais: Uma revolução na educação profissional e tecnológica**. Brasília, 2011.

PEÑA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. **Expert Systems with Applications**, v. 41, n. 4 PART 1, p. 1432–1462, 2014.

PIRES, A. Renda familiar e escolaridade dos pais: reflexões a partir dos microdados do enem 2012 do estado de são paulo. **ETD - Educação Temática Digital, Campinas,SP**, v. 17, p. 523–541, 2015. Disponível em: <http://periodicos.sbu.unicamp.br/ojs/index.php/etd/article/view/8638262>. .

PROVOST, F.; FAWCETT, T. Robust Classification for Imprecise Environments. **Machine Learning Journal**, v. 42, n. 3, p. 203–231, 2001. Disponível em: <http://people.stern.nyu.edu/fprovost/Papers/rocch-mlj.pdf>. Acesso em: 21/4/2017.

PROVOST, F. J.; FAWCETT, T.; KOHAVI, R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. Proceedings of the Fifteenth International Conference on Machine Learning. **Anais...** , ICML '98. p.445–453, 1998. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Disponível em: <http://dl.acm.org/citation.cfm?id=645527.657469>. .

QUINLAN, R. **C4. 5: programs for machine learning**. 1993.

REY, D.; NEUHÄUSER, M. Wilcoxon-Signed-Rank Test. In: M. Lovric (Ed.); **International Encyclopedia of Statistical Science**. p.1658–1659, 2011. Berlin, Heidelberg: Springer Berlin Heidelberg. Disponível em: <http://dx.doi.org/10.1007/978-3-642-04898-2_616>. .

RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S. S.; GOMES, A. S. A literatura brasileira sobre mineração de dados educacionais. **Congresso Brasileiro de Informática na Educação**, , n. 3, p. 621–630, 2014.

ROMERO, C.; LÓPEZ, M. I.; LUNA, J. M.; VENTURA, S. Predicting students' final performance from participation in on-line discussion forums. **Computers and Education**, 2013.

ROMERO, C.; VENTURA, S. Educational Data Mining: A Review of the State of the Art. **APPLICATIONS AND REVIEWS**, v. 40, n. 6, 2010.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 2013.

SAFAVIAN, S. R.; LANDGREBE, D. A Survey of Decision Tree Classifier Methodology. , 1991.

SHLENS, J. A tutorial on principal component analysis: derivation, discussion and singular value decomposition. **Online Note httpwww snl salk edushlenspubnotespca pdf**, v. 2, p. 1–16, 2003. Disponível em: <www.snl.salk.edu/~shlens/pca.pdf>. .

SIEGEL, S. **Estatística não-paramétrica para ciências do comportamento**. Artmed, 1975.

SIEMENS, G.; BAKER, R. S. J. D. Learning Analytics and Educational Data Mining: Torwards Communication and Collaboration. **Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12**, p. 252–254, 2012. Disponível em: <http://www.columbia.edu/~rsb2162/LAKs reformatting v2.pdf>. .

SILIPO, R.; ADAE, I.; HART, A.; BERTHOLD, M. Seven Techniques for Dimensionality Reduction. **KNIME.com**, p. 1–21, 2014.

SOUSA, S. M. Z. L. Avaliação No Currículo Escolar. **Cadernos de Pesquisa**, , n. 119, p. 175–190, 2003. Disponível em: <http://www.scielo.br/pdf/cp/n119/n119a09.pdf>. .

SPACKMAN, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. Proceedings of the sixth international workshop on Machine learning. **Anais...** . p.160–163, 1989.

SUBRATA, S. K. DAS. Feature Selection with a Linear Dependence Measure. **IEEE Transactions on Computers**, v. C-20, n. 9, p. 1106–1109, 1971.

TRAVITSKI, R. **ENEM: limites e possibilidadesTese de doutorado: Universidade de São Paulo**, 2013. Universidade de Sao Paulo.

VIANNA, H. M. Avaliações Nacionais em Larga Escala: análises e propostas. **Estudos em Avaliação Educacional**, , n. 27, p. 41–76, 2003. Disponível em: <http://www.dma.ufv.br/downloads/MAT 207/2016-I/textos/Texto complementar sobre avaliacoes sitemicas - MAT 207 - 2016-I.pdf>. .

VIGGIANO, E.; MATTOS, C. O desempenho de estudantes no Enem 2010 em diferentes regiões brasileiras. **Revista Brasileira de Estudos Pedagógicos**, v. 94, n. 237, p. 417–438, 2013. Disponível em: <http://rbep.inep.gov.br/index.php/RBEP/article/viewFile/2776/1929>. .

WANG, G.; WANG, Y. 3DM: Domain-oriented Data-driven Data Mining. **Fundam. Inform.**, v. 90, n. 4, p. 395–426, 2009. Disponível em: <http://dx.doi.org/10.3233/FI-2009-0026>. .

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)**. 2011.

XING, W.; GUO, R.; PETAKOVIC, E.; GOGGINS, S. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. **Computers in Human Behavior**, 2015.

YU, L.; LIU, H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. **International Conference on Machine Learning (ICML)**, p. 1–8, 2003. Disponível em: <http://www.aaai.org/Papers/ICML/2003/ICML03-111.pdf>. .

APÊNDICE A – DESCRIÇÃO DAS VARIÁVEIS INDEPENDENTES

Variáveis Categóricas		
ID	Variável	Descrição
1	ESCOLA_CAPITAL	Escola situada em capital
2	ESTUDA_FORA	Estuda em cidade diferente da que mora
3	id_abre_final_semana	Funciona no final de semana
4	id_aee	Atendimento educacional especializado
5	id_agua_cacimba	Abastecimento de água - cacimba
6	id_agua_filtrada	Abastecimento de água – filtrada
7	id_agua_fonte_rio	Abastecimento de água - fonte rio
8	id_agua_poco_artesiano	Abastecimento de água – poco artesiano
9	id_almoxarifado	Possui almoxarifado
10	id_alojam_aluno	Possui alojamento para o aluno
11	id_alojam_professor	Possui alojamento para o professor
12	id_area_verde	Possui área verde
13	id_auditorio	Possui auditório
14	id_banda_larga	Possui banda larga
15	id_banheiro_chuveiro	Possui banheiro com chuveiro
16	id_bercario	Possui berçário
17	id_biblioteca	Possui biblioteca
18	id_cozinha	Possui cozinha
19	id_dependencias_pne	Possui dependências para PNE
20	id_despensa	Possui despensa
21	id_energia_gerador	Possui energia – gerador
22	id_energia_outros	Possui energia - outros
23	id_escola_comp_predio	Possui prédio compartilhado com outra escola
24	id_esgoto_fossa	Possui esgoto - fossa
25	id_esgoto_inexistente	Possui esgoto - inexistente
26	id_esgoto_rede_publica	Possui esgoto - rede publica
27	id_espaco_turma_pba	Escola cede espaço para turmas do Programa Brasil Alfabetizado
28	id_internet	Possui internet
29	id_laboratorio_ciencias	Possui laboratório ciências
30	id_laboratorio_informatica	Possui laboratório informática
31	id_lavanderia	Possui lavanderia
32	id_lixo_coleta_periodica	Destinação do lixo- coleta periódica
33	id_lixo_enterra	Destinação do lixo - enterra
34	id_lixo_joga_outra_area	Destinação do lixo - joga outra área
35	id_lixo_outros	Destinação do lixo - outros
36	id_lixo_queima	Destinação do Lixo - queima
37	id_lixo_recicla	Destinação do lixo – recicla
38	id_local_func_casa_professor	Local de funcionamento da escola - casa professor
39	id_local_func_galpao	Local de funcionamento da escola - galpão
40	id_local_func_outros	Local de funcionamento da escola - – outros
41	id_local_func_predio_escolar	Local de funcionamento da escola - prédio escolar
42	id_local_func_salas_outra_esc	Local de funcionamento da escola - outra escola
43	id_local_func_templo_igreja	Local de funcionamento da escola – templo igreja
44	id_local_func_unid_prisional	Local de funcionamento da escola - unidade prisional
45	id_localizacao	Localização
46	id_localizacao_diferenciada	Localização diferenciada da escola
47	id_localizacao_esc	Localização (Escola)
48	id_material_esp_nao_utiliza	Materiais didáticos específicos para atendimento à diversidade sociocultural - Não utiliza
49	id_material_esp_quilombola	Materiais didáticos específicos para atendimento à diversidade sociocultural – Quilombolas
50	id_mod_ativ_complementar	Atividade Complementar

51	id_mod_eja	Modalidade EJA
52	id_patio_coberto	Patio coberto
53	id_patio_descoberto	Patio descoberto
54	id_proposta_pedag_alternancia	Escola com proposta pedagógica de formação por Alternância
55	id_quadra_esportes_coberta	Dependências existentes na escola - Quadra esportes coberta
56	id_quadra_esportes_descoberta	Dependências existentes na escola - Quadra esportes descoberta
57	id_refeitorio	Dependências existentes na escola - Refeitório
58	id_reg_medio_integrado	Ensino Regular - Ensino Médio – Integrado
59	id_reg_medio_medio	Ensino Regular - Ensino Médio - Médio
60	id_reg_medio_prof	Ensino Regular - Ensino Médio- Ensino Profissional
61	id_sala_atendimento_especial	Dependências existentes na escola - - Sala atendimento especial
62	id_sala_diretoria	Dependências existentes na escola - Sala diretoria
63	id_sala_leitura	Dependências existentes na escola - Sala leitura
64	id_sala_professor	Dependências existentes na escola - Sala professor
65	id_sanitario_dentro_predio	Dependências existentes na escola - Sanitário dentro prédio
66	id_sanitario_ei	Dependências existentes na escola - Banheiro adequado à educação infantil
67	id_sanitario_fora_predio	Dependências existentes na escola – Sanitário fora prédio
68	id_sanitario_pne	Dependências existentes na escola - Sanitário PNE
69	id_secretaria	Dependências existentes na escola - Sala de secretaria
70	in_acesso	Indicador de solicitação de sala de fácil acesso – ENEM
71	in_ampliada_18	Indicador de solicitação de prova super ampliada com fonte tamanho 18 – ENEM
72	in_ampliada_24	Indicador de solicitação de prova super ampliada com fonte tamanho 24 – ENEM
73	in_autismo	Indicador de autismo
74	in_baixa_visao	Indicador de baixa visão
75	in_braille	Indicador de solicitação de prova impressa em braille
76	in_certificado	Indicador de necessidade de certificado
77	in_deficiencia_auditiva	Indicador de deficiência auditiva
78	in_deficiencia_fisica	Indicador de deficiência física
79	in_deficiencia_mental	Indicador de deficiência mental
80	in_deficit_atencao	Indicador de deficit atenção
81	in_dislexia	Indicador de dislexia
82	in_gestante	Indicador de gestante
83	in_idoso	Indicador de Possui oso
84	in_lactante	Indicador de lactante
85	in_ledor	Indicador de ledor
86	in_leitura_labial	Indicador de leitura labial
87	in_libras	Indicador de libras
88	in_mesa_cadeira_rodas	Indicador de mesa cadeira rodas
89	in_mesa_cadeira_separada	Indicador de mesa cadeira separada
90	in_sabatista	Indicador de sabatista
91	in_status_redacao	Status da redação
92	in_surdez	Indicador de surdez
93	in_transcricao	Indicador de solicitação de transcrição
94	inse	Indicador Sócio Econômico
95	ipe	Indicador Permanência na Escola
96	nacionalidade	Nacionalidade
97	q001	Até quando seu pai estudou?
98	q002	Até quando sua mãe estudou?
99	q003	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares)
100	q005	A resPossui ência de sua família é

101	q006	A resPossui ência de sua família está localizada em
102	q007	Você tem TV em cores em sua casa?
103	q008	Você tem Possui videocassete e/ou DVDVocê tem em sua casa?
104	q009	Você tem Rádio em sua casa?
105	q010	Microcomputador Você tem em sua casa?
106	q011	Automóvel Você tem em sua casa?
107	q012	Máquina de lavar roupa Você tem em sua casa?
108	q013	Geladeira Você tem em sua casa?
109	q014	Freezer (aparelho independente ou parte da geladeira duplex)Você tem em sua casa?
110	q015	Telefone fixo Você tem em sua casa?
111	q016	Telefone celular Você tem em sua casa?
112	q017	Acesso à Internet Você tem em sua casa?
113	q018	TV por assinatura Você tem em sua casa?
114	q019	Aspirador de pó Você tem em sua casa?
115	q020	Empregada mensalista Você tem em sua casa?
116	q021	Banheiro Você tem em sua casa?
117	q022	Você exerce ou já exerceu atividade remunerada?
118	q030	Quantos anos você levou para concluir o Ensino Fundamental?
119	q031	Você deixou de estudar durante o Ensino Fundamental?
120	q032	Em que tipo de escola você cursou o Ensino Fundamental?
121	q033	Quantos anos você levou para concluir o Ensino Médio?
122	q034	Você deixou de estudar durante o Ensino Médio?
123	q035	Em que tipo de escola você cursou o Ensino Médio?
124	q036	Caso você ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades Pró-Uni (Programa UniversPossui ade para Todos)
125	q037	Caso você ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades? Bolsa de estudos da própria Instituição de Ensino Superior
126	q038	Caso você ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades? Bolsa de estudos da empresa onde trabalho
127	q039	Caso você ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades? Auxílio do Programa de Financiamento Estudantil – FIES
128	q041	Quantas horas semanais você trabalha ou trabalhou aproximadamente?
129	q047	Indique se você frequenta ou frequentou: Curso profissionalizante
130	q048	Indique se você frequenta ou frequentou: Curso preparatório para vestibular
131	q049	Indique os cursos que você frequenta ou frequentou: Curso superior
132	q050	Indique os cursos que você frequenta ou frequentou: Curso de língua estrangeira
133	q051	Indique os cursos que você frequenta ou frequentou: Curso de computação ou informática
134	q052	Indique os cursos que você frequenta ou frequentou: Curso preparatório para concursos públicos
135	q053	Indique os cursos que você frequenta ou frequentou: :Outro curso
136	q054	Você cursa ou já cursou a Educação de Jovens e Adultos – EJA?
137	q055	Como é ou era o principal curso de EJA que você frequenta ou frequentou?
138	q056	Sobre EJA
139	q057	Sobre EJA
140	q058	Sobre EJA
141	q059	Sobre EJA
142	q060	Sobre EJA
143	q061	Sobre EJA
144	q062	Sobre EJA
145	q063	Sobre EJA
146	q064	Você já frequentou o ensino regular?

147	q065	Indique o que levou você a deixar de cursar o ensino regular: Falta de vaga em escola pública
148	q066	Indique o que levou você a deixar de cursar o ensino regular: Ausência de escola perto de casa
149	q067	Indique o que levou você a deixar de cursar o ensino regular: Dificuldades após reprovação
150	q068	Indique o que levou você a deixar de cursar o ensino regular: Falta de interesse em estudar
151	q069	Indique o que levou você a deixar de cursar o ensino regular: Falta de condições adequadas na escola
152	q070	Indique o que levou você a deixar de cursar o ensino regular: Trabalho, falta de tempo para estudar
153	q071	Indique o que levou você a deixar de cursar o ensino regular: Motivos pessoais, casamento/filhos etc.
154	q072	Indique o que levou você a deixar de cursar o ensino regular:Falta de apoio familiar.
155	q073	Indique o que levou você a deixar de cursar o ensino regular: Problemas de saúde ou possui ente comigo ou familiares
156	q074	Indique o que levou você a deixar de cursar o ensino regular:Discriminação/Preconceitos (sexo, raça, Possui ade, classe etc.)
157	q075	Indique o que levou você a deixar de cursar o ensino regular: Medo de sofrer violência
158	q076	Quantos anos você tinha quando deixou de frequentar o ensino regular?
159	REGIAO_ESCOLA	Região geográfica da Escola
160	tp_cor_raca	Cor/Raça
161	tp_estado_civil	Em que tipo de escola você cursou o Ensino Fundamental?
162	tp_lingua	Tipo de Língua Estrangeira
163	tp_sexo	Sexo
164	uf_entidade_certificacao	Sigla da Unidade da Federação da Entidade Certificadora
165	uf_residencia	Sigla da Unidade da Federação de residência
166	id_agua_rede_publica	Abastecimento de água – rede publica
167	in_cegueira	Indicador de cegueira
	Variáveis Numéricas	
	Variável	**Descrição**
168	idade	idade
169	IFD	Índice de Formação Docente
170	ITD	Índice de Titulação Docente
171	num_comp_administrativos	Número de comp administrativos
172	num_comp_alunos	Número de comp alunos
173	num_computadores	Número de computadores
174	num_equip_copiadora	Número de equip copiadora
175	num_equip_dvd	Número de equip dvd
176	num_equip_fax	Número de equip fax
177	num_equip_foto	Número de equip foto
178	num_equip_impressora	Número de equip impressora
179	num_equip_multimidia	Número de equip multimidia
180	num_equip_parabolica	Número de equip parabolica
181	num_equip_retro	Número de equip retro
182	num_equip_som	Número de equip som
183	num_equip_tv	Número de equip tv
184	num_equip_videocassete	Número de equip videocassete
185	num_funcionarios	Número de funcionarios
186	num_salas_existentes	Número de salas existentes
187	num_salas_utilizadas	Número de salas utilizadas
188	q004	Quantas pessoas moram em sua casa (incluindo você)?
189	q023	Indique os motivos que levaram você a participar do ENEM: Testar meus conhecimentos
190	q024	Indique os motivos que levaram você a participar do ENEM: Aumentar a possibilidade de conseguir um emprego

191	q025	Indique os motivos que levaram você a participar do ENEM: Progredir no meu emprego atual
192	q026	Indique os motivos que levaram você a participar do ENEM: Ingressar na Educação Superior Pública
193	q027	Indique os motivos que levaram você a participar do ENEM: Ingressar na Educação Superior Privada
194	q028	Indique os motivos que levaram você a participar do ENEM: Conseguir uma bolsa de estudos (ProUni, outras)
195	q029	Indique os motivos que levaram você a participar do ENEM: Participar do Programa de Financiamento Estudantil – FIES
196	q040	Com que idade você começou a exercer uma atividade remunerada?
197	q042	Indique a importância de cada um dos motivos abaixo na sua decisão de trabalhar: Ajudar meus pais nas despesas com a residência
198	q043	Indique a importância de cada um dos motivos abaixo na sua decisão de trabalhar: Sustentar minha família (esposo/a, filhos/as etc.)
199	q044	Indique a importância de cada um dos motivos abaixo na sua decisão de trabalhar: Ser independente/ganhar meu próprio dinheiro
200	q045	Indique a importância de cada um dos motivos abaixo na sua decisão de trabalhar: Adquirir experiência
201	q046	Indique a importância de cada um dos motivos abaixo na sua decisão de trabalhar: Custear/pagar meus estudos

APÊNDICE B – ESTATÍSTICA DESCRITIVA DAS VARIÁVEIS NUMÉRICAS

ID	Variáveis numéricas	Preenchim.	G. Inf.	Min	Max	Média	Desvio Padrão	Variância	Coef. Varia.
1	idade	99,99	0,01	14	65	17,98	1,45	2,10	0,08
2	IFD	96,35	0,13	6,9	97,3	72,45	14,97	224,02	0,21
3	ITD	100		1,189	3,19	2,224	0,327	0,11	
4	num_comp_admin	99,75	0,08	1	2000	147,04	217,54	47325,68	1,48
5	num_comp_alunos	96,04	0,10	3	2779	209,91	281,37	79171,43	1,34
6	num_computadores	100,00	0,08	1	4779	351,40	466,29	217429,50	1,33
7	num_equip_copiadora	100,00	0,01	0	107	5,69	12,35	152,43	2,17
8	num_equip_dvd	100,00	0,03	0	138	7,82	14,91	222,17	1,91
9	num_equip_fax	100,00	0,01	0	25	2,64	3,90	15,22	1,48
10	num_equip_foto	100,00	0,04	0	167	7,84	15,58	242,61	1,99
11	num_equip_impressora	100,00	0,06	0	290	31,96	43,77	1915,38	1,37
12	num_equip_multimidia	100,00	0,06	0	270	32,62	37,24	1387,05	1,14
13	num_equip_parabolica	100,00	0,02	0	13	1,29	1,79	3,21	1,39
14	num_equip_retro	100,00	0,02	0	73	6,39	12,99	168,75	2,03
15	num_equip_som	100,00	0,03	0	138	6,51	15,73	247,55	2,42
16	num_equip_tv	100,00	0,06	0	128	15,99	20,88	436,12	1,31
17	num_equip_videocas.	100,00	0,02	0	63	3,55	9,96	99,15	2,81
18	num_funcionarios	100,00	0,12	32	1368	245,58	257,03	66065,46	1,05
19	num_salas_existe.	100,00	0,07	2	2810	41,63	149,68	22402,98	3,60
20	num_salas_utili.	100,00	0,05	2	265	31,67	30,49	929,91	0,96
21	q004	100,00	0,00	1	20	4,09	1,24	1,54	0,30
22	q023	100,00	0,00	0	5	4,04	1,38	1,92	0,34
23	q024	100,00	0,01	0	5	3,89	1,53	2,35	0,39
24	q025	0,93	0,00	0	5	2,37	1,87	3,48	0,79
25	q026	100,00	0,00	0	5	4,90	0,52	0,27	0,11
26	q027	100,00	0,01	0	5	3,02	1,79	3,21	0,59
27	q028	100,00	0,04	0	5	4,03	1,49	2,21	0,37
28	q029	100,00	0,04	0	5	3,17	1,83	3,35	0,58
29	q040	20,69	0,00	13	25	16,14	1,53	2,35	0,09
30	q042	20,69	0,00	0	5	3,24	1,80	3,25	0,56
31	q043	20,69	0,00	0	5	0,84	1,58	2,49	1,89
32	q044	20,69	0,00	0	5	4,31	1,20	1,44	0,28
33	q045	20,69	0,00	0	5	4,57	0,91	0,83	0,20
34	q046	20,69	0,00	0	5	2,96	2,00	4,01	0,68

ID	1° decil	9° decil	1°Quartil	3°Quartil	Dist. S.interqua.	Coef.Var. Separ.	Ausentes	Mediana
1	17	19	17	18	0,5	0,03	2	18
2	58,10	87,90	67,20	81,90	7,35	0,10	809	73,9
3							0	
4	23	296	50	175	62,5	0,69	56	91
5	47	440	78	215	68,5	0,49	879	139
6	80	748	140	369	114,5	0,48	0	241
7	0	11	1	5	2	1,00	0	2
8	0	17	1	9	4	1,33	0	3
9	0	6	0	3	1,5	1,50	0	1
10	0	21	2	8	3	1,00	0	3
11	1	68	7	33	13	0,65	0	20
12	4	72	12	39	13,5	0,68	0	20
13	0	3	0	2	1	1,00	0	1
14	0	17	0	7	3,5	3,50	0	1
15	0	12	1	5	2	1,00	0	2
16	1	43	3	20	8,5	1,21	0	7
17	0	8	0	2	1	0,00	0	0
18	80	504	106	252	73	0,45	0	162
19	11	67	16	41	12,5	0,50	0	25
20	10	67	15	36	10,5	0,46	0	23
21	3	5	3	5	1	0,25	0	4
22	2	5	3	5	1	0,20	0	5
23	1	5	3	5	1	0,20	0	5
24	0	5	0	4	2	1,00	21981	2
25	5	5	5	5	0	0,00	0	5
26	0	5	2	5	1,5	0,50	0	3
27	2	5	3	5	1	0,20	0	5
28	0	5	2	5	1,5	0,38	0	4
29	14	18	15	17	1	0,06	17596	16
30	0	5	2	5	1,5	0,38	17596	4
31	0	4	0	1	0,5	0,00	17596	0
32	3	5	4	5	0,5	0,10	17596	5
33	3	5	5	5	0	0,00	17596	5
34	0	5	1	5	2	0,67	17596	3

APÊNDICE C – ESTATÍSTICA DESCRITIVA DAS VARIÁVEIS CATEGÓRICAS

ID	Variável	Preenchimento	G. de Inf.	Ausentes	Moda	F. Moda %	Nº categ.
1	ESCOLA_CAPITAL	100,00%	0,036	0	0	70,95	2
2	ESTUDA_FORA	100,00%	0,001	0	1	67,78	2
3	id_abre_final_semana	100,00%	0,002	0	0	72,73	2
4	id_aee	100,00%	0,005	0	0	92,66	2
5	id_agua_cacimba	100,00%	0,001	0	0	96,34	2
6	id_agua_filtrada	100,00%	0,002	0	1	95,52	2
7	id_agua_fonte_rio	100,00%	0,011	0	0	92,13	2
8	id_agua_poco_artesiano	100,00%	0,015	0	0	61,36	2
9	id_almoxarifado	100,00%	0,008	0	1	87,38	2
10	id_alojam_aluno	100,00%	0,037	0	0	73,64	2
11	id_alojam_professor	100,00%	0,001	0	0	90,95	2
12	id_area_verde	100,00%	0,002	0	1	74,29	2
13	id_auditorio	100,00%	0,005	0	1	81,95	2
14	id_banda_larga	99,62%	0,007	84	1	96,90	3
15	id_banheiro_chuveiro	100,00%	0,006	0	1	76,20	2
16	id_bercario	100,00%	0,000	0	0	99,74	2
17	id_biblioteca	100,00%	0,013	0	1	94,24	2
18	id_cozinha	100,00%	0,001	0	1	80,37	2
19	id_dependencias_pne	100,00%	0,000	0	1	83,44	2
20	id_despensa	100,00%	0,002	0	0	63,83	2
21	id_energia_gerador	100,00%	0,000	0	0	89,30	2
22	id_energia_outros	100,00%	0,001	0	0	99,57	2
23	id_escola_comp_predio	99,41%	0,001	131	0	98,63	3
24	id_esgoto_fossa	100,00%	0,038	0	0	50,85	2
25	id_esgoto_inexistente	100,00%	0,001	0	0	99,06	2
26	id_esgoto_rede_publica	100,00%	0,049	0	1	60,98	2
27	id_espaco_turma_pba	100,00%	0,001	0	0	98,72	2
28	id_internet	100,00%	0,001	0	1	99,62	2
29	id_laboratorio_ciencias	100,00%	0,003	0	1	90,46	2
30	id_laboratorio_informatica	100,00%	0,012	0	1	95,65	2
31	id_lavanderia	100,00%	0,016	0	0	81,03	2
32	id_lixo_coleta_periodica	100,00%	0,001	0	1	98,30	2
33	id_lixo_enterra	100,00%	0,003	0	0	97,51	2
34	id_lixo_joga_outra_area	100,00%	0,002	0	0	98,09	2
35	id_lixo_outros	100,00%	0,000	0	0	97,34	2
36	id_lixo_queima	100,00%	0,004	0	0	97,01	2
37	id_lixo_recicla	100,00%	0,000	0	0	76,86	2
38	id_local_func_casa_professor	100,00%	0,001	0	0	99,68	2
39	id_local_func_galpao	100,00%	0,000	0	0	98,86	2
40	id_local_func_outros	100,00%	0,003	0	0	96,16	2
41	id_local_func_predio_escolar	100,00%	0,007	0	1	94,01	4
42	id_local_func_salas_outra_esc	100,00%	0,004	0	0	96,57	2
43	id_local_func_templo_igreja	100,00%	0,001	0	0	99,70	2
44	id_local_func_unid_prisional	100,00%	0,000	0	0	99,73	2
45	id_localizacao	100,00%	0,026	0	1	82,51	2

46	id_localizacao_diferenciada	100,00%	0,001	0	0	99,35	2
47	id_localizacao_esc	100,00%	0,026	0	1	82,51	2
48	id_material_esp_nao_utiliza	100,00%	0,000	0	1	99,90	2
49	id_material_esp_quilombola	100,00%	0,000	0	0	99,90	2
50	id_mod_ativ_complementar	100,00%	0,009	0	0	95,26	2
51	id_mod_eja	100,00%	0,002	0	1	58,42	2
52	id_patio_coberto	100,00%	0,001	0	1	57,35	2
53	id_patio_descoberto	100,00%	0,000	0	1	63,15	2
54	id_proposta_pedag_alternancia	100,00%	0,004	0	0	96,66	2
55	id_quadra_esportes_coberta	100,00%	0,001	0	1	71,60	2
56	id_quadra_esportes_descoberta	100,00%	0,005	0	0	58,67	2
57	id_refeitorio	100,00%	0,000	0	1	67,21	2
58	id_reg_medio_integrado	100,00%	0,011	0	1	97,07	2
59	id_reg_medio_medio	100,00%	0,019	0	0	90,42	2
60	id_reg_medio_prof	100,00%	0,006	0	1	86,21	2
61	id_sala_atendimento_especial	100,00%	0,010	0	0	72,95	2
62	id_sala_diretoria	100,00%	0,017	0	1	95,45	2
63	id_sala_leitura	100,00%	0,000	0	1	55,45	2
64	id_sala_professor	100,00%	0,014	0	1	94,38	2
65	id_sanitario_dentro_predio	100,00%	0,000	0	1	99,71	2
66	id_sanitario_ei	100,00%	0,000	0	0	99,38	2
67	id_sanitario_fora_predio	100,00%	0,000	0	0	55,45	2
68	id_sanitario_pne	100,00%	0,001	0	1	92,80	2
69	id_secretaria	100,00%	0,002	0	1	87,42	2
70	in_acesso	100,00%	0,000	0	0	99,91	2
71	in_ampliada_18	100,00%	0,000	0	0	99,98	2
72	in_ampliada_24	100,00%	0,000	0	0	99,98	2
73	in_autismo	100,00%	0,000	0	0	100,00	2
74	in_baixa_visao	100,00%	0,000	0	0	99,90	2
75	in_braille	100,00%	0,000	0	0	99,98	2
76	in_certificado	57,94%	0,000	9332		42,06	3
77	in_deficiencia_auditiva	100,00%	0,000	0	0	99,93	2
78	in_deficiencia_fisica	100,00%	0,000	0	0	99,90	2
79	in_deficiencia_mental	100,00%	0,000	0	0	99,99	2
80	in_deficit_atencao	100,00%	0,000	0	0	99,95	2
81	in_dislexia	100,00%	0,000	0	0	99,98	2
82	in_gestante	100,00%	0,000	0	0	99,97	2
83	in_idoso	100,00%	0,000	0	0	100,00	2
84	in_lactante	100,00%	0,000	0	0	99,97	2
85	in_ledor	100,00%	0,000	0	0	99,93	2
86	in_leitura_labial	100,00%	0,000	0	0	99,99	2
87	in_libras	100,00%	0,000	0	0	99,97	2
88	in_mesa_cadeira_rodas	100,00%	0,000	0	0	99,99	2
89	in_mesa_cadeira_separada	100,00%	0,000	0	0	99,96	2
90	in_sabatista	100,00%	0,000	0	0	99,05	2
91	in_status_redacao	100,00%	0,003	0	7	99,31	8

92	in_surdez	100,00%	0,000	0	0	99,97	2
93	in_transcricao	100,00%	0,000	0	0	99,90	2
94	inse	97,65%	0,099	522	Médio Alto	30,86	7
95	ipe	100,00%	0,004	0	80% ou mais	87,43	5
96	nacionalidade	100,00%	0,001	0	1	99,59	4
97	q001	100,00%	0,041	0	E	30,27	9
98	q002	100,00%	0,025	0	E	33,10	9
99	q003	100,00%	0,074	0	B	18,98	17
100	q005	100,00%	0,001	0	A	67,54	5
101	q006	100,00%	0,010	0	B	89,73	4
102	q007	100,00%	0,030	0	A	52,81	4
103	q008	100,00%	0,012	0	A	65,60	4
104	q009	100,00%	0,010	0	A	55,07	4
105	q010	100,00%	0,045	0	A	58,22	4
106	q011	100,00%	0,015	0	A	47,90	4
107	q012	100,00%	0,015	0	A	73,98	4
108	q013	100,00%	0,004	0	A	94,42	4
109	q014	100,00%	0,016	0	D	62,82	4
110	q015	100,00%	0,042	0	D	49,98	4
111	q016	100,00%	0,016	0	C	56,61	4
112	q017	100,00%	0,033	0	A	73,52	4
113	q018	100,00%	0,035	0	D	69,55	4
114	q019	100,00%	0,023	0	D	77,03	4
115	q020	100,00%	0,005	0	D	93,42	4
116	q021	100,00%	0,024	0	A	58,83	4
117	q022	100,00%	0,001	0	C	79,31	3
118	q030	100,00%	0,008	0	B	70,81	8
119	q031	41,61%	0,001	12955		58,39	6
120	q032	41,61%	0,027	12955		58,39	7
121	q033	100,00%	0,002	0	F	49,15	7
122	q034	41,64%	0,001	12949		58,36	6
123	q035	41,64%	0,000	12949		58,36	6
124	q036	100,00%	0,013	0	A	92,05	2
125	q037	100,00%	0,003	0	A	90,93	2
126	q038	100,00%	0,006	0	B	56,73	2
127	q039	100,00%	0,014	0	A	77,00	2
128	q041	20,69%	0,003	17596		79,31	6
129	q047	20,69%	0,001	17596		79,31	3
130	q048	20,69%	0,007	17596		79,31	3
131	q049	20,69%	0,001	17596		79,31	3
132	q050	20,69%	0,008	17596		79,31	3
133	q051	20,69%	0,003	17596		79,31	3
134	q052	20,69%	0,001	17596		79,31	3
135	q053	20,69%	0,001	17596		79,31	3
136	q054	29,30%	0,000	15686		70,70	3
137	q055	0,44%	0,001	22089		99,56	5

138	q056	0,44%	0,000	22089		99,56	3
139	q057	0,44%	0,000	22089		99,56	3
140	q058	0,44%	0,000	22089		99,56	3
141	q059	0,44%	0,001	22089		99,56	3
142	q060	0,45%	0,001	22088		99,55	3
143	q061	0,44%	0,000	22089		99,56	3
144	q062	0,44%	0,000	22089		99,56	3
145	q063	0,45%	0,000	22088		99,55	3
146	q064	29,13%	0,001	15723		70,87	3
147	q065	25,09%	0,000	16620		74,91	3
148	q066	25,09%	0,000	16620		74,91	3
149	q067	25,09%	0,000	16620		74,91	3
150	q068	25,09%	0,000	16620		74,91	3
151	q069	25,09%	0,000	16620		74,91	3
152	q070	25,09%	0,000	16620		74,91	3
153	q071	25,09%	0,000	16620		74,91	3
154	q072	25,09%	0,000	16620		74,91	3
155	q073	25,09%	0,000	16620		74,91	3
156	q074	25,09%	0,000	16620		74,91	3
157	q075	25,09%	0,000	16620		74,91	3
158	q076	25,06%	0,002	16628		74,94	7
159	REGIAO_ESCOLA	100,00%	0,053	0	nordeste	33,30	5
160	tp_cor_raca	100,00%	0,010	0	3	42,60	6
161	tp_estado_civil	100,00%	0,001	0	0	99,48	3
162	tp_lingua	100,00%	0,059	0	0	53,37	2
163	tp_sexo	100,00%	0,004	0	M	51,13	2
164	uf_entidade_certificacao	29,19%	0,016	15711		70,81	28
165	uf_residencia	100,00%	0,065	0	MG	14,37	27
166	id_agua_rede_publica	100,00%	0,038	0	1	75,1	2
167	in_cegueira	100,00%	0	0	0	99,97	2

APÊNDICE D – VARIÁVEIS RETIRADAS PELO PROCESSO DE REDUÇÃO SUPERVISIONADA

Valores Ausentes	Baixa Variância	Alta Correlação
q025	num_salas_existentes	id_sala_diretoria
q040	num_salas_utilizadas	id_laboratorio_informatica
q041	num_equip_tv	id_biblioteca
q042	num_equip_videocassete	num_equip_dvd
q043	num_equip_dvd	num_computadores
q044	num_equip_parabolica	num_comp_alunos
q045	num_equip_copiadora	uf_esc
q046	num_equip_retro	id_localizacao_esc
q047	num_equip_impressora	
q048	num_equip_som	
q049	num_equip_multimidia	
q050	num_equip_fax	
q051	num_equip_foto	
q052	num_computadores	
q053	num_comp_administrativos	
q065	num_comp_alunos	
q066	num_funcionarios	
q067	idade	
q068	q004	
q069	q025	
q070	q026	
q071	q040	
q072	q043	
q073		
q074		
q075		
q076		

APÊNDICE E – VARIÁVEIS SIGNIFICATIVAS (P-VALOR < 0,05) PARA O MODELO DE REGRESSÃO LOGÍSTICA

Variável	Estimate	Std. Error	z value	Pr(>|z|)
inseBaixo	-2,60	0,82	-3,19	0,001
q026	1,56	0,35	4,49	0,000
tp_cor_raca5	-1,54	0,75	-2,04	0,041
q004	-1,42	0,53	-2,70	0,007
id_localizacao_diferenciada4	1,27	0,52	2,45	0,014
inseMedio Baixo	-1,25	0,21	-6,01	0,000
q030F	-1,25	0,55	-2,27	0,023
REGIAO_ESCOLAsudeste	1,14	0,14	7,99	0,000
id_proposta_pedag_alternancia1	-1,09	0,25	-4,45	0,000
ipeDe 20% a 40%	1,01	0,30	3,42	0,001
id_esgoto_inexistente1	0,96	0,42	2,31	0,021
inseMedio	-0,91	0,13	-7,01	0,000
q030D	-0,91	0,30	-3,06	0,002
num_funcionarios	0,84	0,20	4,21	0,000
tp_cor_raca4	-0,83	0,32	-2,58	0,010
num_equip_multimidia	0,83	0,25	3,34	0,001
itd	-0,82	0,25	-3,28	0,001
REGIAO_ESCOLAnordeste	0,76	0,16	4,78	0,000
tp_cor_raca2	-0,75	0,25	-2,94	0,003
num_equip_fax	-0,75	0,20	-3,67	0,000
q032C	0,74	0,11	6,72	0,000
q028	-0,69	0,12	-5,78	0,000
id_lixo_coleta_periodica1	-0,68	0,27	-2,54	0,011
id_local_func_salas_outra_esc1	-0,68	0,22	-3,05	0,002
id_escola_comp_predio1	0,66	0,33	1,99	0,047
tp_lingua1	-0,62	0,07	-9,56	0,000
id_agua_fonte_rio1	-0,60	0,18	-3,29	0,001
num_equip_videocassete	-0,60	0,26	-2,30	0,022
idade(18,65]	-0,59	0,10	-5,90	0,000
num_equip_parabolica	-0,54	0,22	-2,42	0,015
q032D	0,53	0,16	3,34	0,001
tp_cor_raca3	-0,51	0,24	-2,12	0,034
tp_cor_raca1	-0,51	0,24	-2,11	0,035
id_agua_filtrada1	0,50	0,21	2,36	0,018
inseMedio Alto	-0,49	0,09	-5,23	0,000
q064B	-0,49	0,16	-3,05	0,002
ifd	0,48	0,22	2,15	0,032
q011C	-0,46	0,22	-2,06	0,039
q017B	-0,41	0,17	-2,47	0,014
id_energia_gerador1	0,39	0,13	2,93	0,003
q010D	-0,39	0,10	-3,92	0,000

id_banheiro_chuveiro1	-0,39	0,09	-4,10	0,000
id_laboratorio_ciencias1	0,38	0,14	2,82	0,005
id_alojam_aluno1	-0,37	0,12	-3,05	0,002
id_sanitario_pne1	-0,36	0,16	-2,19	0,029
q029	-0,36	0,10	-3,52	0,000
ESCOLA_CAPITAL1	0,35	0,10	3,38	0,001
q010C	0,33	0,12	2,82	0,005
q016B	0,33	0,10	3,28	0,001
q016C	0,33	0,09	3,55	0,000
q030B	0,32	0,11	3,09	0,002
q010B	0,32	0,08	3,99	0,000
id_abre_final_semana1	0,30	0,09	3,52	0,000
tp_sexoM	0,30	0,06	5,01	0,000
q037B	-0,29	0,10	-2,81	0,005
q007C	-0,27	0,10	-2,76	0,006
id_despensa1	-0,26	0,08	-3,12	0,002
q038B	0,25	0,06	3,96	0,000
idade(17,18]	-0,21	0,08	-2,76	0,006
id_sala_leitura1	0,19	0,08	2,45	0,014
q011D	0,19	0,07	2,52	0,012
id_sala_atendimento_especial1	0,17	0,08	2,10	0,036
id_patio_descoberto1	0,15	0,08	2,06	0,040
inseBaixo	-2,60	0,82	-3,19	0,001
q026	1,56	0,35	4,49	0,000
tp_cor_raca5	-1,54	0,75	-2,04	0,041
q004	-1,42	0,53	-2,70	0,007
id_localizacao_diferenciada4	1,27	0,52	2,45	0,014
inseMedio Baixo	-1,25	0,21	-6,01	0,000
q030F	-1,25	0,55	-2,27	0,023
REGIAO_ESCOLAsudeste	1,14	0,14	7,99	0,000
id_proposta_pedag_alternancia1	-1,09	0,25	-4,45	0,000
ipeDe 20% a 40%	1,01	0,30	3,42	0,001
id_esgoto_inexistente1	0,96	0,42	2,31	0,021
inseMedio	-0,91	0,13	-7,01	0,000
q030D	-0,91	0,30	-3,06	0,002
num_funcionarios	0,84	0,20	4,21	0,000
tp_cor_raca4	-0,83	0,32	-2,58	0,010
num_equip_multimidia	0,83	0,25	3,34	0,001
itd	-0,82	0,25	-3,28	0,001
REGIAO_ESCOLAnordeste	0,76	0,16	4,78	0,000
tp_cor_raca2	-0,75	0,25	-2,94	0,003
num_equip_fax	-0,75	0,20	-3,67	0,000

q032C	0,74	0,11	6,72	0,000
q028	-0,69	0,12	-5,78	0,000
id_lixo_coleta_periodica1	-0,68	0,27	-2,54	0,011
id_local_func_salas_outra_esc1	-0,68	0,22	-3,05	0,002
id_escola_comp_predio1	0,66	0,33	1,99	0,047
tp_lingua1	-0,62	0,07	-9,56	0,000
id_agua_fonte_rio1	-0,60	0,18	-3,29	0,001
num_equip_videocassete	-0,60	0,26	-2,30	0,022
idade(18,65]	-0,59	0,10	-5,90	0,000
num_equip_parabolica	-0,54	0,22	-2,42	0,015
q032D	0,53	0,16	3,34	0,001
tp_cor_raca3	-0,51	0,24	-2,12	0,034
tp_cor_raca1	-0,51	0,24	-2,11	0,035
id_agua_filtrada1	0,50	0,21	2,36	0,018
inseMedio Alto	-0,49	0,09	-5,23	0,000
q064B	-0,49	0,16	-3,05	0,002
ifd	0,48	0,22	2,15	0,032
q011C	-0,46	0,22	-2,06	0,039
q017B	-0,41	0,17	-2,47	0,014
id_energia_gerador1	0,39	0,13	2,93	0,003
q010D	-0,39	0,10	-3,92	0,000
id_banheiro_chuveiro1	-0,39	0,09	-4,10	0,000
id_laboratorio_ciencias1	0,38	0,14	2,82	0,005