UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE SERRA TALHADA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

MINERAÇÃO DE DADOS APLICADA A ANÁLISE DE DESEMPENHO DE ALUNOS NO 5° ANO DO ENSINO FUNDAMENTAL

Por

Rodrigo Jacinto da Silva

Serra Talhada,
Janeiro/2019

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

RODRIGO JACINTO DA SILVA

MINERAÇÃO DE DADOS APLICADA A ANÁLISE DE DESEMPENHO DE ALUNOS NO 5° ANO DO ENSINO FUNDAMENTAL

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Serra Talhada da Universidade Federal Rural de Pernambuco como requisito parcial à obtenção do grau de Bacharel.

Orientador: Prof. Paulo Mello da Silva

Serra Talhada,
Janeiro/2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca da UAST, Serra Talhada - PE, Brasil.

S586m Silva, Rodrigo Jacinto da

Mineração de dados aplicada a análise de desempenho de alunos no 5º ano do ensino fundamental / Rodrigo Jacinto da Silva. – Serra Talhada, 2019.

50 f.: il.

Orientador: Paulo Mello da Silva

Trabalho de Conclusão de Curso (Graduação em Bacharelado em Sistemas de Informação) – Universidade Federal Rural de Pernambuco. Unidade Acadêmica de Serra Talhada, 2019.

Inclui referências e apêndices.

1. Ensino fundamental. 2. Rendimento escolar. 3. Tecnologia da informação. I. Silva, Paulo Mello da, orient. II. Título.

CDD 004

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

RODRIGO JACINTO DA SILVA

MINERAÇÃO DE DADOS APLICADA A ANÁLISE DE DESEMPENHO DE ALUNOS NO 5° ANO DO ENSINO FUNDAMENTAL

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendida e aprovada por unanimidade em 25/01/2019 pela banca examinadora.

Banca Examinadora:

Prof. Paulo Mello da Silva
Orientador
Universidade Federal Rural de Pernambuco

Prof. Sérgio de Sá Leitão Paiva Júnior
Universidade Federal Rural de Pernambuco

Prof. Arthur Diego de Godoy Barbosa
Universidade Federal Rural de Pernambuco

Este trabalho é dedicado

a toda minha família e aos mais próximos de mim.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus por me dá saúde, força e capacidade para chegar ao fim do curso que sempre sonhei em me formar, sem ele não séria possível esse feito.

Agradeço a minha família por está ao meu lado nos momentos difíceis e de superação, sempre me apoiando e ajudando que pudesse seguir em frente sem medo de falhar.

Aos meus orientadores Prof. Paulo Mello Silva e Prof. Italo Cesar de Souza Belo, que me ajudaram com seus conhecimentos e experiência para que esse trabalho pudesse ser feito. E também a todos os professores de Sistema de Informação que fizeram parte da minha graduação fazendo que me tornasse o profissional que sou hoje.

E por último e não menos importante, agradeço a todos aos meus amigos que fiz nesse tempo de faculdade, que puderam me ajudar ao longo do curso.

“Por isso não tema, pois estou com você;
não tenha medo, pois sou o seu Deus.
Eu o fortalecerei e o ajudarei;
eu o segurarei
com a minha mão direita vitoriosa.
(Bíblia Sagrada, Isaías 41,10)

RESUMO

Diferenças no desempenho educacional dos alunos podem ser visto através de provas feitas pelas escolas e avaliações dos sistemas educacionais do governo, podendo este desempenho está ligado diretamente com socioeconômico do estudante e estruturas das escolas bem como seus docentes. Uma das avaliações feitas pelo INEP (Instituto nacional de estudos de pesquisas) para avaliar a performance dos discentes do ensino fundamental, o qual este trabalho se dedica, foi criado SAEB (Sistema de Avaliação da Educação Básica). Através dos dados fornecidos pelo SAEB, é possível ser feita uma análise mais detalhada sobre o desempenho dos alunos. O objetivo desse trabalho é utilizar técnicas computacionais para analisar os dados do SAEB referente aos alunos do 5° ano do ensino fundamental das escolas públicas de estado de Pernambuco. Assim para que esse trabalho fosse realizado, foi utilizada a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) fazendo uso de todas as suas etapas para melhor entendimento e organização dos dados. O trabalho fez o uso do software Weka, que por sua vez ajudou na aplicação de técnicas e algoritmos para a análise da base de dados. A pesquisa pode trazer o desenvolvimento de um modelo de desempenho para auxiliar os gestores educacionais e professores na tomada de decisão.

Palavras-chave: análise de desempenho do ensino básico, mineração de dados educacionais, análise preditiva de alunos do ensino fundamental.

ABSTRACT

Differences in the educational performance of students can be seen through evidence made by schools and evaluations of government educational systems, and this performance is directly linked with socioeconomic of student and school structures as well as their teachers. One of the evaluations made by the INEP (National Institute of Research Studies) to evaluate the performance of elementary school students, in which this work is dedicated, was created SAEB (System of Evaluation of Basic Education). Through the data provided by SAEB, a more detailed analysis of student performance can be made. The aim of this work is to use computational techniques to analyze data from the SAEB referring to students of the 5th year of elementary school in the state public schools of pernambuco. In order to perform this work, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was used, making use of all its steps to better understand and organize the data. The work made use of the Weka software, which in turn helped in the application of techniques and algorithms for the analysis of the database. The research was able to bring the development of a performance model to assist educational managers and teachers in decision making.

Keywords: performance analysis of basic education, educational data mining, predictive analysis of primary school students.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BD	Banco de Dados
EB	Educação Básica
EF	Ensino fundamental
INEP	Instituto Nacional de Estudos e Pesquisas
MD	Mineração de Dados
SAEB	Sistema de Avaliação da Educação Básica

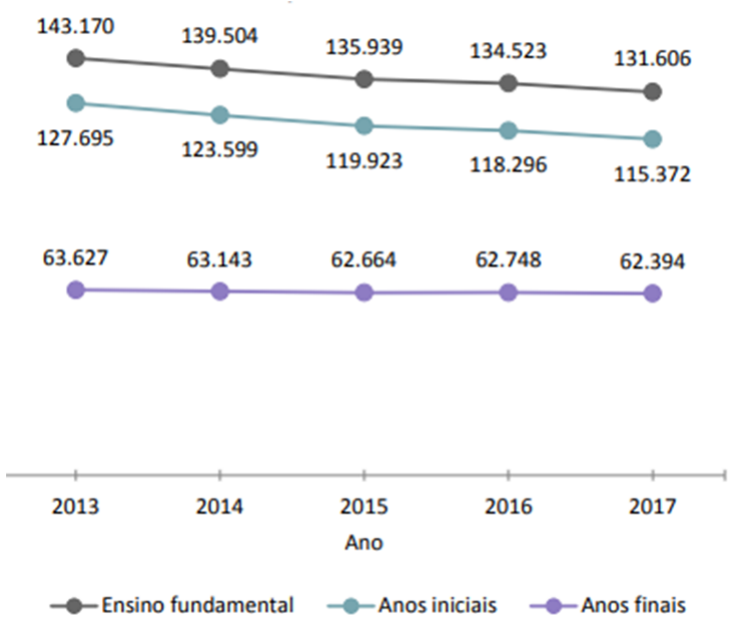
SUMÁRIO

1 Introdução

Tendo conhecimento que um dos pilares da construção de um profissional tem como inicio que o individuo tenha uma base sólida na educação básica, mostrando a grande importância que o ensino fundamental tem na vida dos estudantes. Assim como enfatiza (SILVA; MONTEIRO; RODRIGUES, 2017), a educação básica é responsável para o desenvolvimento do intelecto da criança, é onde se aprende os conceitos educacionais, os fundamentos essenciais para toda a vida estudantil, e que mais tarde se torne um profissional competente, cumprindo leis e sendo ético em suas ações.

Visto a importância das escolas de EF, podemos analisar o grande crescimento da quantidade de escolas e de qual rede fazem parte, segundo os dados levantados pelo (INEP, 2017).

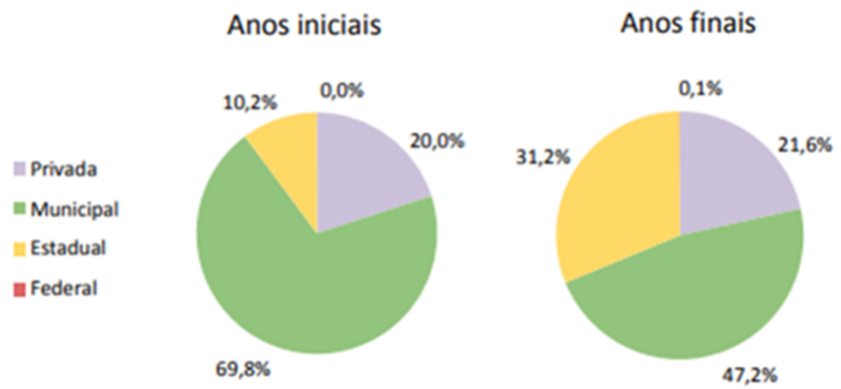
Figura 1.1 – Evolução do número de escolas do ensino fundamental (anos iniciais e anos finais) - Brasil 2013-2017



Fonte: Censo escolar INEP (2017)

Como é visto na pesquisa feita pelo censo escolar realizado pelo (INEP, 2017) as escolas que ofertam os anos inicias do ensino fundamental teve uma pequena queda 2,5 por cento comparando-se o ano 2016 com 2017. Já as escolas que oferecem os últimos anos do EF, se mantiveram estável nos últimos anos entre 2013 á 2017.

Figura 1.2 – Número de escolas do ensino fundamental (anos iniciais e anos finais) por dependência administrativa - Brasil 2017



Fonte: Censo escolar INEP (2017)

Analisando os gráficos de pizza no qual os dados foram coletados pelo (INEP, 2017), podemos ver que a rede que mais oferece os anos inicias do EF são as escolas da rede municipal com 69,8 por cento, e também sendo responsável pela maior quantidade de escolas que oferecem os anos finais do EF com 47,2 por cento, vindo logo atrás das escolas municipais temos as escolas estaduais que oferecem 31,2 por cento dos anos finais do ensino fundamental.

Conforme foi visto á grande quantidade de escolas para EF e seus vários tipos de redes (Privada, Municipal, Estadual e Federal), umas das formas que o governo usa para avaliar o desempenho dos alunos, é através do (SAEB, 2018) Sistema de Avaliação da Educação Básica, tem como objetivo criar diagnóstico sobre a qualidade de ensino que está sendo ofertado, tentando analisar quais fatores pode está interferindo no desempenho do aluno. Mesmo com esse sistema de avaliação, ainda existem grandes dificuldades para reconhecer quais fatores de fato, podem interferir no desempenho dos estudantes do EF.

Por esse motivo, é notado que a mineração de dados tem se tornado uma grande aliada para ajudar nessa análise de desempenho educacional. De acordo com (WEBBER et al., 2013) a existência da MD para educação, pode prover informações novas e mais precisas com objetivo de ajudar na tomada de decisão dos gestores, professores e todos envolvidos na área. “A mineração de dados educacionais é uma área recente de pesquisa que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino.” (BAKER; ISOTANI; CARVALHO, 2011)

O intuito deste trabalho é utilizar de técnicas de MD, para poder ajudar gestores na tomada de decisão, identificando variáveis fornecidas através de uma base de dados do SAEB,

quais motivos podem interferir no desempenho do aluno. Como é mostrado (MARTURANO; PIZATO, 2015) a diferença de desempenho de alunos no 5° ano do ensino fundamental podem ser afetadas através de características do próprio estudante, da família e da escola, fazendo uma análise que identifique esses fatores.

1.1 Objetivos

1.1.1 Objetivo Geral

Utilizar técnicas computacionais para analisar os dados do SAEB referente aos alunos do 5° ano do ensino fundamental das escolas públicas de estado de Pernambuco.

1.1.2 Objetivos Específicos

- Realizar revisão teórica de artigos, teses e dissertações para construção do referencial teórico.
- Extrair os dados referentes aos alunos do 5° ano do ensino fundamental das escolas públicas de Pernambuco.
- Analisar os dados de proficiência e socioeconômicos contidos na base do SAEB.
- Desenvolver um modelo de predição baseado em técnicas de mineração de dados.
- Utilizar software para realizar o processo de mineração de dados.

1.2 Motivação e Justificativa

A utilização de técnicas computacionais para melhorar área da educação vem crescendo a cada ano, mostrando a importância e benefícios que as tecnologias podem trazer para domínio educacional. Uma das áreas que fazem parte dessa crescente e que esse trabalho faz uso é da mineração de dados, que tem como um dos seus objetivos analisar, reconhecer padrões e

descobrir informações que não são possíveis em bancos convencionais (CÔRTES; PORCARO; LIFSCHITZ, 2002).

Mas ainda que exista essa crescente ainda á vários ramos na educação que não fazem grandes proveitos da computação para melhorias do ensino, Sendo esse um dos motivos que esse trabalho se dedica para prover novos recursos para educação. Observou-se que há uma grande quantidade de base de dados disponibilizadas pelo INEP (Instituto Nacional de Estudos e Pesquisas), que armazenam dados dos estudantes e não são utilizadas com novos recursos para que seja retirada informações importantes com relação ao desempenho dos alunos. A partir disso, notou-se que o SAEB (Sistema de Avaliação da Educação Básica), prover dados sobre os alunos de ensino fundamental que podem ser investigados e analisados.

A partir desse contexto, este trabalho se dedica a analisar a base de dados da 5°série do ensino fundamental de Pernambuco (SAEB, 2018). Onde foi encontrados pontos que podem ser melhorados, sendo eles:

- Desconhecimento dos fatores que influenciam o desempenho dos alunos nas avaliações de língua portuguesa do SAEB 2015.
- Pouca informação sobre o desempenho dos alunos da educação básica do estado de Pernambuco.
- Dificuldade na tomada de decisão devido aos poucos recursos de dados sobre o desempenho dos alunos.
- Necessidade do uso de técnicas para analisar grandes quantidades de dados.

Analisando esses problemas, um dos métodos que são usados na MD que podem ser aplicados para descoberta de conhecimento é a Regressão linear que por usa vez tem como objetivo descobrir quais variáveis de um determinado contexto podem influenciar em uma determinada variável escolhida (CURRAL, 1994).

Vistos todos esses fatores, esse trabalho tem como objetivo analisar a base de dados de língua portuguesa do SAEB 2015 da 5°série do EF das escolas de Pernambuco, para investigar quais atributos influenciam diretamente no desempenho do estudante, podendo servir para tomada de decisão de gestores e professores e profissionais que trabalhem nas melhorais educacionais.

1.3 Organização do Trabalho

O trabalho aqui desenvolvido está dividido em seis capítulos, detalhando suas atividades desde o levantamento bibliográfico até o seu desenvolvimento e suas conclusões finais. A seguir é mostrado um breve resumo de cada um deles.

- Capítulo 1 - Introdução: Mostra de forma resumida o trabalho aqui proposto, bem como seus objetivos e justificativas de criação deste projeto
- Capítulo 2 - Referencial Teórico: neste capítulo são realizadas discussões teóricas sobre o tema aqui abordado, alinhando o pensamento de importantes estudiosos na área ao contexto do trabalho.
- Capítulo 3 - Trabalhos Relacionados: este capítulo aborda trabalhos que possuem conteúdo e finalidade semelhantes, descrevendo-os e fazendo uma rápida comparação com o objetivo de buscar diferenças e semelhanças entre eles e o trabalho aqui proposto.
- Capítulo 4 - Materiais e Métodos: demonstra os materiais utilizados para a produção do projeto e descreve as etapas de desenvolvimento do mesmo.
- Capítulo 5 - Resultados: neste capítulo são mostrados os resultados obtidos com a elaboração da pesquisa.
- Capítulo 6 - Considerações Finais: Demonstra as metas que foram alcançadas e trabalhos futuros com podem ser feito através deste projeto.

2 Referencial Teórico

2.1 Ensino Fundamental

O ensino fundamental ou também conhecido como educação básica, é considerado entre 1° e 5° série das escolas que ensinam esse nível. Atualmente essa primeira formação escolar, é chamada de ensino fundamental 1, já que o ensino fundamental 2 é considerado ente 6° e o 9° ano (MEC, 2018).

A educação básica é considerada uma das formações mais importantes na vida estudantil do aluno, pois é a partir dela que começamos a descobrir os ensinamentos que serão levados para a vida inteira (OLIVEIRA, 2007).

É nesta fase que os conceitos psicológicos e cognitivos entram em desenvolvimento na cabeça da criança, assim mostrando que esse período é bastante importante e sendo a escola responsável por transmitir todo o conhecimento necessário para alimentar todas essas mentes desses iniciantes no mundo escolar (OLIVEIRA, 2007).

2.2 SAEB

O SAEB (Sistema de Avaliação da Educação Básica) é considerado o principal instrumento para avaliar os alunos e escolas no geral de todo o país. O sistema tem como objetivo analisar o conhecimento e habilidades dos alunos, a partir de aplicação de testes, com a finalidade também de avaliar o ensino dado em sala de aula, bem como a estrutura no geral escolar que são elas: condições de infra-estrutura das unidades escolares, perfil do diretor e recursos utilizados para gerenciamento escolar, perfil do professor e práticas pedagógicas adotadas, características dos alunos e hábitos de estudo (SAEB, 2018).

A partir desses testes aplicados pelo SAEB, é feita a análise dos resultados que permite acompanhar a evolução do desempenho dos estudantes e das escolas, e dos vários fatores que partem destes dois contextos, assim possibilitando as correções dos problemas encontrados nas escolas e da qualidade do ensino bem como o melhoramento do desempenho dos alunos

(NACIONAIS, 1998).

Essas informações são utilizadas por gestores da educação, professores, pesquisadores da área e entre outros, além de permitir que a sociedade conheça o nível das escolas públicas e privadas e do ensino aplicado (INEP, 2017).

A partir de 1995, o SAEB começou apresentar dois de resultados: a proficiência média e o aproveitamento médio. A proficiência é o resultado do conjunto de habilidades mostrada pelo aluno das disciplinas de língua portuguesa e matemática. A tabela a seguir mostra como exemplo, um resultado da proficiência média e do aproveitamento médio de língua portuguesa (NACIONAIS, 1998).

Tabela 2.1 – Proficiência média e aproveitamento médio, Brasil e regiões, leitura (Língua Portuguesa)

Região	Proficiência Média			Aproveitamento Médio		
	Séries			Séries		
	4°	8°	3°	4°	8°	3°
BR	177	252	277,0	49,4	65,9	66,1
N	154	238	262,0	44,4	61,4	61,6
NE	160	227	253,0	46,4	57,2	59,3
SE	187	262	285,0	51,5	69,3	68,6
S	181	257	283,0	50,5	68,2	67,8
CO	185	252	283,0	50,7	66,4	68,1

Fonte: NACIONAIS, 1998, p. 33

2.3 Desempenho Estudantil

O aprendizado pode ser considerado multideterminado, pois existem varias características que influenciam no aprendizado, o desempenho acadêmico do ensino fundamental podem ser atribuídas ao nível de sócio econômico a características do aluno, da família, da escola e até mesmo da vizinhança, principalmente se quando combinados com todas essas características (AIKENS; BARBARIN, 2008).

Projetos de pesquisa, que tem o objetivo de analisar características dos alunos, mostram que o desempenho escolar está relacionado diretamente com habilidades sociais e comportamentos, tanto do aluno quando da família (ACKERMAN et al., 2007), de acordo com Byrne et al. (2011) a experiência individual que a criança cria na escola também pode afetar seu desempenho, sendo assim influenciadas por propriedades do contexto escolar. É visto que a criança ver a escola por parte, como importante influenciador no nível de stress, particularmente podendo está relacionada ao desempenho (MARTURANO et al., 2009).

Como citado no (FELÍCIO; TERRA; ZOGHBI, 2012) a permanência do aluno na escola bem como sua passagem pela educação infantil, considerada a pré-escola, são fatores muito importantes para o desenvolvimento do aluno. Outro fator importante de acordo com Salles, Parente e Freitas (2010) alunos que freqüentam escolas diferentes não tem o mesmo desempenho que teria se ficasse o ensino fundamental apenas em uma, mesmo que a mudança seja para uma escola do mesmo município.

Além das habilidades sociais e comportamento que se tem entorno do aluno, a escola como qualidade de ensino a organização, o clima e segurança, sua estrutura em si, fazem parte das melhorias de desempenho do estudante, podendo dar vantagens no ciclo estudantil do alunado (HALLINGER; HECK, 2002)

2.4 Mineração de Dados e KDD

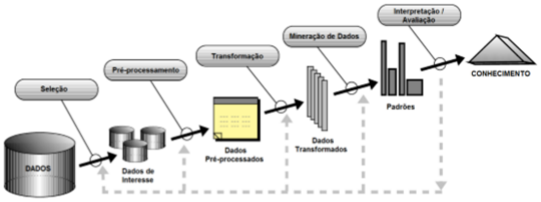
Uma das definições sobre mineração de dados, mostra que é uma forma para analisar grandes quantidades de dados, para definir grupos, identificar padrões e descobrir relações desconhecidas, que sem utilizar esse tipo de processo não é possível, e assim dar mais suporte no aperfeiçoamento da tomada de decisão (CÔRTES; PORCARO; LIFSCHITZ, 2002).

Com sua flexibilidade e facilidade de uso, a mineração de dados é utilizada em diversas áreas e profissões, como no ramo da economia que trabalha com estatísticas, dos cientistas de dados e base de dados padrões (WEIAND; PINTO,). Analisando especificamente a utilização do processo de mineração dados para os grandes banco de dados, Data Warehouses, ou outros tipos de armazenamento de dados, (MACEDO; MATOS, 2010), definiu que através do processo de MD, é capaz de extrair conhecimento sobre os dados analisados, que podem ajudar nas melhorias de apoio a decisão.

Para alguns estudiosos da área, o processo de mineração de dados é apenas uma das etapas do processo de KDD (Knowledge Discovery in Databases) que traduzido para o português significa Descoberta de Conhecimento em Bancos de Dados (FAYYAD et al., 1996).

Em virtude do que foi dito, a técnica de KDD é o processo geral de análise de dados, partindo da seleção dos dados, até o descobrimento de conhecimento, sendo MD uma parte especifica dessa cadeia de procedimentos, que tem como objetivo aplicar algoritmos para identificar padrões nos dados. Podemos analisar cada parte dos procedimentos de KDD na figura a seguir.

Figura 2.1 – Fases do KDD)



(Extraído de Fayyad et al., 1996)

2.5 Seleção

Começando pela seleção, é feita a coleta dos dados a partir de um determinado domínio especifico, que se tenha interesse de tirar conhecimento, fazendo o agrupamento dos dados de forma organizada (MACEDO; MATOS, 2010).

2.6 Pré Processamento

Partindo para segunda etapa do processo, é começado o pré-processamento, que na maioria das vezes, ocupa boa parte do tempo que se é destinado para processo de KDD, sendo recomendada ser feita, apenas por especialistas no domínio de aplicação dos dados (MANNILA, 1996). Com sua tamanha complexidade, essa etapa pode ser quebrada em partes, podendo ser definida da seguinte forma:

- **Limpeza dos Dados:** é feita a verificação dos dados, para garantir a integridade, e corrigir erros que possam comprometer a retirada de conhecimento da base.

- **Codificação dos Dados:** tem a função de organizar os dados, para que no processo de mineração de dados sejam inseridos os algoritmos de forma correta.

- **Enriquecimento dos Dados:** tem como objetivo enriquecer as informações já contidas na base, fazendo pesquisas sobre o domínio, para complementar os dados.

(Boente et al, 2008).

2.7 Transformação

A etapa de transformação tem como objetivo, preparar os dados para que sejam utilizados no processo de mineração de dados, é feita a manipulação dos dados. Segundo especialista da área, a manipulação é feita dependendo qual algoritmo será utilizado na etapa de mineração de dados, pois o manuseio dos dados é feita especificamente para que atenda os recursos do algoritmo escolhido (CASTANHEIRA, 2008).

De acordo com Silva Filho (2009), na maioria das vezes, a evolução nos dados, é feita usando alguma fórmula matemática, para que atenda de forma apropriada as posteriores modelagens, deixando quantidade necessária de informações, e diminuindo a probabilidade de erros.

2.8 Mineração de Dados

De acordo com a idéia de (CASTANHEIRA, 2008), a etapa de mineração de dados fica sendo uma das etapas mais importantes no processo de KDD. A MD é o processo de identificação de padrões encontrados na base de dados, para isso é feita em três etapas, é analisado e escolhido a técnica que melhor se encaixe com a base, podendo ser elas: Associação, Classificação, Agrupamento, Regressão, Estimativa ou Desvio. Em seguida é feito a escolha do algoritmo que será utilizado para identificar padrões na BD, por fim, é executado á mineração dos dados (FAYYAD et al., 1996).

2.9 Interpretação e Avaliação

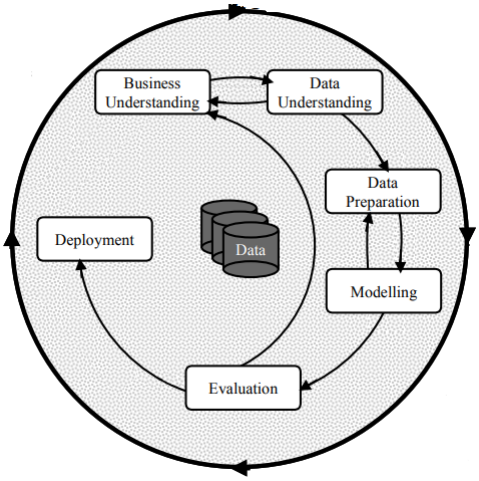
Chegando a última etapa do processo de KDD, está etapa é considerada “Pós-processamento”, onde pode ser analisado quais foram os resultados obtidos, se eles foram satisfatórios e de serventia para o domínio (MACEDO; MATOS, 2010), Caso a mineração dos dados não tenha retornado um resultado satisfatório, Silva Filho (2009) afirma que pode ser refeito algumas partes ou todas as etapas de KDD, para tentar corrigir em qual parte houve alguma inconsistência no procedimento.

2.10 Modelo de Processo Crisp-DM

O modelo Crisp-DM utilizado frequentemente em empresas de tecnologia, que significa Cross Industry Standard Process for Data Mining, que trazendo para o português é entendido “processo padrão da indústria cruzada para mineração de dados”.

Assim como citado no (SANTOS; RAMOS, 2006) Foi baseado no modelo padrão de KDD, mas tem como objetivo ser mais produtivo e flexível para aplicação de projetos de mineração de dados. O modelo possui 6 fases bem definidas, que norteia o processo de criação de um projeto de mineração de dados bem sucedido (CHAPMAN et al., 1999).

Figura 2.2 – Fasess do modelo crisp-dm.



Fonte: CHAPMAN et al., 1999, p. 7

2.11 Compreensão do Negócio

A primeira fase tem como objetivo saber qual o problema que deverá ser resolvido, tomar conhecimento sobre o domínio no qual vai ser desenvolvido o trabalho, é nessa fase que se define qual o objetivo da criação do projeto (MORO; LAUREANO; CORTEZ, 2011)

2.12 Compreensão dos Dados

De acordo com (MORO; LAUREANO; CORTEZ, 2011) a fase de compreensão, fica sendo a busca por dados disponíveis para que sejam analisados e documentados. O cientista

de dados dever validar os dados para saber se as informações atendem aos objetivos definidos, deve ser feita a organização dos dados de forma que fiquem claras de ser entendidas (SANTOS; RAMOS, 2006).

2.13 Preparação dos dados

Segundo (IBM, 2016) está é uma das fases mais importantes do modelo, pois é nesta fase que é feita o pré-processamento dos dados, é verificado (ruídos, outliers, campos ausentes), e até mesmo que seja criada uma nova base dados caso o banco disponível não atenda aos objetivos escolhidos. É nesta fase que devem ser criado o dicionário de dados, onde explica como estão organizados os dados, o que seus valores representam (IBM, 2016).

2.14 Modelagem

A fase de modelagem aplica as técnicas de mineração na base de dados, é nesta fase que deve ser escolhido um software no qual aplica os métodos de MD. (IBM, 2016). De acordo com (MORO; LAUREANO; CORTEZ, 2011) os dados minerados podem ser usados para alimentar algoritmos que tentam prever conhecimento sobre o domínio no qual o projeto foi feito.

2.15 Avaliação

A avaliação dos resultados, devem ser apresentados para os envolvidos no domínio através de reuniões ou qualquer outra forma de apresentação.(SANTOS; RAMOS, 2006) . Os resultados devem analisados de acordo com o que foi definido na fase de compreensão do domínio, para saber se alcançaram os objetivos propostos.(IBM, 2016)

2.16 Implementação dos Modelos

A última fase do modelo de processo crisp-dm, fica por conta de aplicar os modelos criados na empresa. (IBM, 2016) Assim como é citado por (SANTOS; RAMOS, 2006), é através dos resultados obtidos que a empresa pode mudar toda sua forma de pensar e agir no ramo de negócios, a partir das informações contidas no modelo que os processos da organização pode ser modificados.

2.17 Atividades e Algoritmos de Mineração de Dados

A mineração de dados possui vários tipos de métodos onde são classificadas suas atividades bem como seus algoritmos, eles são divididos em grupos como é mostrado (QUILICI-GONZALEZ; ZAMPIROLLI, 2015).

Figura 2.3 – Classificação da natureza das tarefas de mineração de dados.

Fonte: QUILICI-GONZALEZ; ZAMPIROLLI, 2015, p. 27

Os grupos são divididos pelas atividades descritivas e preditivas como mostra a figura 2.3, onde a atividade descritiva possui algoritmos que tentam reconhecer similaridades nos dados, assim criando grupos a partir de sua semelhança. Já as atividades preditivas nas qual esse trabalho fez uso, utilizam algoritmos que reconhecem padrões nos dados e predizem valores de atributos futuros, também podendo mostrar quais atributos podem influenciar diretamente em um atributo específico. (QUILICI-GONZALEZ; ZAMPIROLLI, 2015); (GALVÃO; MARIN, 2009).

Assim como foi dito por Galvão e Marin (2008), atividades com seus algoritmos existem independente de utilizar-se a mineração de dados, mas fazendo uso do KDD, e aplicando as atividades já mencionadas, produzem bom resultados.

2.18 Correlação

Na estatística a correlação é definida como a relação entra duas variáveis, onde podem existir alguns tipos de correlação, a mais conhecida é o coeficiente de correlação de Pearson, produto-momento ou (r) de Pearson, ele é responsável por medir a relação entre duas variáveis lineares quantitativas (FERREIRA, 2008).

Este coeficiente é na maioria das vezes representado pela letra "r" e assume apenas valores entre -1 e 1, quando valor de r = 1, isso significa uma relação perfeita positiva entre as variáveis em questão, em que no caso quando uma aumenta a outra também é aumentada, já se o coeficiente for r = -1, demonstra uma relação negativa perfeita entre as variáveis, Isto é, se uma aumenta, a outra sempre diminui. Em casos do valor r = 0 significa que as variáveis não dependem linearmente uma da outra, e no caso podem ser investigados por outros meios (FERREIRA, 2008).

2.19 Regressão

Este trabalho faz uso do algoritmo regressão, onde pode ser definido como um dos algoritmos mais simples e fácil utilização, mas dependendo da quantidade de atributos pode se tornar complexo (IBM, 2010). Como foi definido (IBM, 2010), Regressão tem como objetivo principal selecionar quais atributos (variáveis independentes) podem mais influenciar em uma variável escolhida (variável dependente), assim criando um modelo no qual pode predizer valores futuros para a variável dependente, podendo mostrar bom resultados, sobre como certas variáveis podem influenciar diretamente no valor de uma variável escolhida.

Existem dois tipos de regressão, simples e múltipla, onde a simples tenta predizer o valor de uma única variável, já a regressão múltipla deixa o algoritmo um pouco mais complexo podendo fazer a previsão de várias variáveis dependentes (GALVÃO; MARIN, 2009). Uma característica importante do algoritmo de regressão, que só podem ser usados base de dados com valores quantitativos, onde BD que possuem variáveis não numéricas devem ser transformadas em valores numéricos.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_k X_k + \xi$$

3 Trabalhos Relacionados

3.1 Preditores de Desempenho Escolar no 5°Ano do Ensino Fundamental (Trabalho 1)

O trabalho elaborado por (MARTURANO; PIZATO, 2015), analisa diferenças no desempenho dos alunos no 5° ano do ensino fundamental, onde essas diferenças podem está relacionado com as características do aluno, da família e da escola. Assim foi definido, "O objetivo deste estudo prospectivo foi testar um modelo de predição de desempenho no 5° ano do ensino fundamental – EF, tendo como preditores habilidades acadêmicas e sociais, problemas de comportamento e percepção de estressores escolares no 3° ano, bem como o tempo de exposição à educação infantil – EI e a condição socioeconômica do alunado da escola de EF" (MARTURANO; PIZATO, 2015, pp. 1).

O projeto coletou dados de 248 alunos de 4 escolas públicas de uma cidade do estado de São Paulo (A qual não foi informada), É relatado que os pais e professores foram informados sobre as metas do trabalho, os procedimentos e a livre participação, mães e professores participaram com informações sobre os alunos, onde foram aplicado o modelo de predição de desempenho acadêmico para investigar quais fatores influenciam.

Os critérios para avaliar os estudantes do 3° ano do EF, foi o nível de socioeconômico, as características da escola de ensino fundamental baseado em três critérios: mobilidades dos alunos (Mudança de cidade) e porcentagem dos alunos sem acesso prévio á ensino infantil e habilidades acadêmicas avaliadas com o teste de desempenho escolar. Já dos alunos do 5° ano foram avaliados por desempenho acadêmico através de dois instrumentos, avaliação coletiva de português e matemática, e avaliação realizada pela escola na 4° série (Atual 5° ano). Outros fatores foram a Escala de Competência Acadêmica do Sistema de Avaliação de Habilidades Sociais – SSRS-BR e as habilidades sociais, os comportamentos externalizantes e os comportamentos internalizantes também foram avaliados pelo professor com o SSRS-BR.

Os resultados da pesquisa como é mostrado na Figura 3.1, detectou diferenças de gênero no 3° ano, as meninas possuem mais habilidades sociais e os meninos com mais comportamentos

externalizantes. Já no 5° ano as meninas tiveram melhor desempenho e melhor competência acadêmica.

Tabela 3.1 – Médias e desvios-padrão das variáveis do estudo, por gênero, anos na EI e escola de EF

3*Variáveis	Gênero		Anos na Educação Infantil		Escola de Ensino Fundamental			
	Meninos	Meninas	1	2	Alfa	Beta	Gama	Delta
	(N: 125)	(N: 123)	(N: 74)	(N: 174)	(N: 81)	(N: 56)	(N: 50)	(N: 61)
Desempenho TDE 3º	80,8 (28,9)	86,7 (27,5)	78,7 (28,3)	85,8 (28,1)	72,1 (34,4)	76,4 (27,4)	95,4 (16,8)	96,2 (17,2)
Habilidades sociais 3º	46,5 (14,0)	51,1 (12,0)	48,1 (10,9)	49,0 (14,1)	46,6 (11,4)	46,3 (13,0)	46,7 (12,2)	55,5 (14,4)
Externalização 3º	9,4 (7,0)	5,4 (6,2)	7,5 (6,5)	7,4 (7,1)	10,8 (6,9)	7,8 (6,2)	2,7 (5,3)	6,5 (6,2)
Internalização 3º	4,2 (3,3)	3,3 (2,8)	3,7 (3,2)	3,5 (3,1)	5,1 (2,6)	2,4 (2,4)	1,7 (2,2)	4,2 (3,7)
Stress escolar 3º	10,6 (7,7)	9,0 (7,1)	9,1 (7,1)	10,1 (7,6)	11,5 (7,3)	11,4 (8,5)	7,2 (7,2)	6,0 (0,8)
Desempenho avaliação coletiva 5º	9,4 (4,1)	10,7 (3,9)	9,2 (3,5)	10,4 (4,2)	8,6 (3,9)	8,3 (3,3)	11,4 (3,5)	12,5 (3,7)
Competência acadêmica 5º	33,7 (9,4)	36,9 (8,6)	35,5 (7,1)	35,1 (9,9)	35,3 (9,1)	34,8 (7,6)	36,5 (9,1)	34,5 (10,5)

Nota: diferenças estatisticamente significativas entre grupos, p<0,05: 1 para gênero; 2 para anos na educação infantil; 3 para escola. Fonte: MARTURANO; PIZATO, 2015, p. 19

A Tabela 3.2, mostra a correlação entre os resultados das variáveis dos alunos no 3° ano e 5° ano do EF, as medidas de competência dos estudantes correlacionam positivamente umas com as outras e também as medidas adaptativas, mostrando que comportamentos no 3° ano influenciam diretamente nos resultados quando o estudante estiver cursando o 5 ano.

Tabela 3.2 – Correlações entre as variáveis avaliadas no 3° e no 5° ano do ensino fundamental

Variáveis	1	2	3	4	5	6	7
1. Desempenho TDE 3º ano	-						
2. Habilidades sociais 3º ano	0,35**	-					
3. Externalização 3º ano	-0,35**	-0,49**	-				
4. Internalização 3º ano	-0,30**	-0,43**	0,52**	-			
5. Stress escolar 3º ano	-0,40**	-0,28**	0,26**	0,16*	-		
6. Desemp. av. coletiva 5º ano	0,73**	0,48**	-0,42**	-0,35**	-0,47**	-	
7. Competência acad. 5º ano	0,55**	0,45**	-0,32**	-0,42**	-0,38**	0,63**	-
8. Nível socioeconômico	0,26**	0,20**	-0,23**	-0,17**	-0,22*	0,33**	0,12

Fonte: MARTURANO; PIZATO, 2015, p. 20

3.2 Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados (Trabalho 2)

O trabalho (GOTTARDO; KAESTNER; NORONHA, 2012) através do Ambiente Virtual de Aprendizado (AVA), e com a grande quantidade de dados que esses ambientes armazenam com relação as atividades dos alunos. O objetivo deste trabalho é aplicar técnicas de mineração de dados para retirar informações que auxiliem os professores no gerenciamento

de processo de ensino, além da definição de atributos para criação de um modelo que consiga analisar o desempenho do aluno.

(GOTTARDO; KAESTNER; NORONHA, 2012) explica que foram criadas 3 tipos de dimensões possíveis de serem extraídos atributos para que fosse criado um modelo de previsão de desempenho, sendo elas descritos na tabela a seguir:

Tabela 3.3 – Dimensões para extração de Atributos

1 Dimensão: perfil geral de uso do AVA:	nesta dimensão o objetivo é identificar dados que representem aspectos de planejamento, organização e gestão do tempo do estudante para a realização do curso. Para isso definiu-se indicadores gerais de quantidade e tempo médio de acessos aos recursos do AVA. Foram incluídos também atributos que representem atividades rotineiras e regulares dos acessos dos aprendizes
2 Dimensão: interação Estudante-Estudante:	com esta dimensão pretende-se verificar se os estudantes interagem entre si usando as ferramentas disponíveis, como fóruns, chats, envio ou recebimento de mensagens. Espera-se, com estes atributos, identificar a existência de colaboração e cooperação entre estudantes. Fato esse denominado por Schrire (2006) como "aprendendo com os outros" (em inglês learning with others).
3 Dimensão: interação Estudante-Professor:	nesta dimensão o objetivo é averiguar como professores ou tutores interagem com estudantes no contexto do AVA. Este tipo de interação tem sua importância destacada por Holliman e Scanlon (2006). Esses autores ressaltam que professores ou tutores têm um papel fundamental no sentido de facilitar e incentivar a colaboração entre estudantes.

Fonte: GOTTARDO; KAESTNER; NORONHA, 2012, p. 5

Com objetivo de analisar e testar o conjunto de atributos extraídos, foi realizado experimentos com uma base de dados do moodle, com informações de alunos que já haviam feitos cursos a distância ou que estavam cursando. A partir da base de dados selecionou uma disciplina, considerando com a maior quantidade de estudantes que já tinha concluído a disciplina. A partir desses critérios foi selecionada uma disciplina com 155 alunos concluintes em quatro turmas

diferentes.

Através da extração de atributos e formação da base de dados, foi aplicada a técnica conhecida como discretização, está tarefa foi executada através de discretização não supervisionada disponível na ferramenta Weka. Através da aplicação e uso da ferramenta foram criados três classes, apresentando o rótulo e a descrição de cada classe, mostrando o número de estudantes bom como suas notas para cada uma das classes

Tabela 3.4 – Distribuição das classes obtidas pelo processo de discretização

Título da Classe	Descrição	Número de Estudantes	Intervalo de Notas
A	Estudantes com desempenho superior	22	87-97
B	Estudantes com desempenho intermediário	109	77-87
C	Estudantes com desempenho inferior	24	67-77

Fonte: GOTTARDO; KAESTNER; NORONHA, 2012, p. 7

Os resultados obtidos mostram que é preciso realizar certas correções na forma de ensino, para melhorar o desempenho do aluno. A pesquisa pode ser útil para professores, acompanhar o aluno de maneira individual que utiliza o ambiente virtual AVA e para definir estratégias de ensino que busquem diminuir a quantidade reprovações.

Tabela 3.5 – Comparativo entre os trabalhos relacionados

FATORES	Este Trabalho	Trabalho 1	Trabalho 2
Próprio para tomada de decisão de Professores e Gestores educacionais	SIM	SIM	SIM
Utilização de informações socioeconômicas do estudante	SIM	SIM	NÃO
Divisão da Base dados, para gerar resultados mais específicos	SIM	NÃO	NÃO
Utilização de técnicas computacionais para prover novos recursos na educação	SIM	SIM	SIM
Criação de modelo genérico para analisar desempenho dos alunos no EF	SIM	NÃO	NÃO
Utilização do Modelo de Processo CRISP-DM	SIM	NÃO	NÃO

Fonte: elaborado pelo autor

4 Metodologia

Tendo em vista as metas a serem compridas por esse trabalho, foram escolhidas processos e métodos para que a pesquisa e os resultados fossem alcançados, a partir disso podemos destacar a sequência dos fatos que foram realizados para chegar a determinado resultado.

Abordagem e metodologia

Neste trabalho será usado o método indutivo de abordagem, sendo referenciado pela metodologia de pesquisa quantitativa. Esse tipo de pesquisa quantitativa considera que tudo pode ser traduzido para a forma de valores numéricos, ou seja, significa transformar em números opiniões, informações ou qualquer outro tipo de dados, para que sejam classificados e analisados. Métodos e técnicas estatísticas devem ser utilizadas (MINAYO; SANCHES, 1993).

Tendo conhecimento do conceito anterior, este trabalho tem o objetivo dentro de uma abordagem quantitativa, fazendo uso dos modelos de processo, métodos e técnicas, propor o desenvolvimento de um modelo para determinar quais fatores relacionados aos estudantes do ensino fundamental, podem influenciar em seu desempenho.

4.1 Aplicando Modelo CRISP-DM

4.2 Compreensão do Negócio

A primeira fase do CRISP-DM consiste em saber qual domínio será trabalhado, e qual o objetivo de usar a mineração de dados nesse contexto, qual problema será resolvido. No caso deste trabalho foi escolhido a área da educação, e tem como objetivo investigar quais fatores influenciam no desempenho do aluno, vindo a servir na tomada de decisão de gestores e professores.

4.3 Compreensão dos Dados

A pós escolher a área na qual vai ser desenvolvido o projeto, e seus objetivos, a segunda fase é feita a pesquisa para encontrar os dados e qual serão as ferramentas utilizadas, além da análise e documentação dos dados. A partir dos objetivos mencionados foi adquirido através do site do governo (http://inep.gov.br/dados) a base de dados do SAEB 2015, prova de língua portuguesa dos alunos do 5° ano do ensino fundamental das escolas de Pernambuco, pois fazendo a análise do banco disponível, existem informações a onde podem ser feitas análise de desempenho dos estudantes, como por exemplo, a base disponibiliza as notas dos alunos bem como o formulário de socioeconômico.

Foram utilizados algumas ferramentas de tecnologia da informação para elaboração deste trabalho.

- Weka Ferramenta utilizada para o processamento dos dados, Waikato Environment for Knowledge Analysis, mais conhecida como Weka, é um software de código livre, onde podem ser usada gratuitamente. Tem como objetivo aplicar técnicas de mineração de dados em base de dados, e aceita vários tipos de formatos (CSV, ARFF, JSON), tem como função a utilização e criação de aprendizagem de máquina onde podem ser aplicada nas bases de dados utilizadas (WITTEN et al., 2016). O weka tem como pré-processamento de dados os métodos mais utilizados na mineração de dados como, regressão, classificação, agrupamento entre outras.

- SPSS Software pertencente a IBM, tem como objetivo á análise estatísticas dos dados, possui métodos e técnicas para aplicação de mineração de dados, possui funções para criar, editar e manusear todos os tipos de bases de dados válidas de vários formatos (CSV, JSON, XLSX), possui grande poder para criar relatórios sobre os tipos de BD escolhidos (PESTANA; GAGEIRO, 2003).

- Overleaf Ferramenta de pré-processamento de texto, utilizada para a criação de artigos, trabalhos de conclusão de curso, no qual este trabalho fez uso.

4.4 Preparação dos Dados

Nesta fase será feita o pré-processamento dos dados, na qual pode ser considerada uma das fases mais importantes do processo, aqui é feito todas as correções na base de dados (ruídos, outliers e linhas em branco) e até desenvolvida uma nova base de dados a partir da base já existente caso seja necessário.

Tendo em vista que será aplicado o método de regressão na próxima fase do CRISP-DM, onde no método utilizado só são aceitos dados numéricos, foi observado que algumas colunas da base, era do tipo não numéricos e que algumas linhas possuíam valores em branco.

A partir desses erros o BD foi redefinido utilizando uma lógica a onde valores não numéricos pudessem ser representados por números em forma de sequência. A tabela a seguir mostra uma parte do dicionário dos dados e de como foi feita essa redefinição.

Tabela 4.1 – Dicionário de Dados redefinida

Variável	Tipo	Descrição	Código de Preenchimento
ID_TURNO	Numérico	Turno em que a turma estuda	Matutino - 1 Vespertino - 2 Noturno - 3 Intermediário - 4
TELEVISÃO	Numérico	Na sua casa tem televisão?	Não Possui - 0 Sim, Uma - 1 Sim, Duas - 2 Sim, Três - 3 Sim, Quatro ou mais - 4
MORA_MAE	Numérico	Você mora com sua mãe?	Não - 0 Sim - 1 Outro Responsável - 2

Fonte: elaborado pelo autor

Como é visto na tabela 4.1, as colunas que possuem valores não numéricos foram redefinidas, como no caso da coluna ID TURNO o aluno preencheu o turno em qual ele estudava, já para base de dados foi redefinida como uma escala de 1 à 4 dependo de qual período ele escolheu.

Outro erro detectado é que alguns alunos não preencheram algumas partes do questionário a onde ficaram linhas em branco, para resolver esse problema foi utilizada a ferramenta SPSS para corrigir esses erros.

O software SPSS possui uma função a onde podem ser definidos, valores para as linhas em branco, acessando o programa ele possui a função "substituir valores ausentes", podendo

escolher a coluna na qual possui valores em branco, o SPSS faz uma busca atrás de campos vazios, caso encontre você pode escolher qual opção será usada para preencher o campo, no caso deste trabalho foi utilizado "média entre dois pontos mais próximos".

4.5 Modelagem

Nesta próxima etapa, é feita aplicação das técnicas de mineração de dados, para que sejam mostrados os resultados obtidos. O software utilizado neste trabalho foi o Weka, no qual possui todas as técnicas necessárias para aplicação do método de regressão. Mas antes de fazer uso do programa, a base de dados foi dividida seguindo a tabela a seguir.

Tabela 4.2 – Divisão da base de dados com relação a Proficiência

Grupo	Proficiência_LP
Base 1	Maior igual á 0 e menor igual á 199
Base 2	Maior igual á 200 e menor igual á 300
Base 3	Maior igual á 300

Fonte: elaborado pelo autor

A tabela a seguir mostra a quantidade de registros (Informações dos alunos) da base dados do SAEB 2015 completa, e depois do tratamento com a divisão em grupos.

Tabela 4.3 – Quantidade de Registros antes e depois do Tratamento

Base de Dados	(N°) Quantidade de Registros
Base Completa	85.410
Base 1	53.291
Base 2	31.602
Base 3	517

Fonte: elaborado pelo autor

Como mostrado na tabela 4.2, a base de dados foi dividida com relação ao desempenho dos alunos, assim dando mais eficácia nos resultados, para fique nítido quais variáveis mais influenciam no desempenho (Proficiência) dos estudantes.

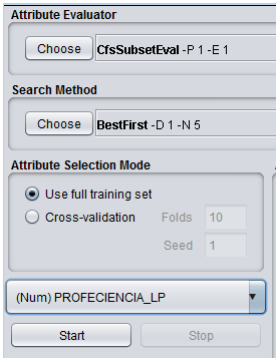
As bases de dados possuem 49 variáveis cada, foi utilizada a opção do Weka Select Atributte, para que pudesse ser feita seleção das variáveis mais influentes com relação a proficiência, o programa nos dá duas opções para que seja executado a função, sendo elas:

CfsSubsetEval - “Avalia o valor de um subconjunto de atributos, considerando a capacidade preditiva individual de cada recurso, juntamente com o grau de redundância entre eles”

(Weka) .

BestFirst – Ranqueia as variáveis com maior influencia com relação a variável escolhida, que neste trabalho é a Proficiência.

Figura 4.1 – Função para seleção de Variáveis



Fonte: Sistema Weka

4.6 Variáveis selecionadas pelo Weka

Tabela 4.4 – Variáveis Selecionadas Base de dados 1

(N) Variável	Nome da Variável	Descrição
1°	ID_DEPENDENCIA_ADM	Dependência Administrativa (Escola)
2°	ID_LOCALIZACAO	Turno da Turma
3°	FREEZER	Na sua casa tem freezer?
4°	QUARTOS	Na sua casa tem quartos para dormir?
5°	DOMESTICA	Na sua casa trabalha alguma empregada doméstica?
6°	MORA_MAE	Você mora com sua mãe?
7°	MAE_ALFABETIZADA	Sua mãe ou a mulher responsável por você sabe ler e escrever?
8°	PAI_ALFABETIZADO	Seu pai ou homem responsável por você sabe ler e escrever?
9°	INCENTIVO_ESTUDAR	Seus pais ou responsáveis incentivam você a estudar?
10°	INCENTIVO_DEVER	Seus pais ou responsáveis incentivam você a fazer o dever de casa e os trabalhos da escola?
11°	INCENTIVO_LER	Seus pais ou responsáveis incentivam você a ler?
12°	TRABALHA_FORA	Você trabalha fora de casa?
13°	TIPO_ESCOLA	Desde a primeira série em que tipo de escola você estudou?
14°	REPROVACAO	Você já foi reprovado?
15°	ABANDONO	Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?
16°	DEVER_LP	Você faz o dever de casa de língua portuguesa?
17°	CORRIGE_LP	O professor corrige o dever de casa de língua portuguesa?

Fonte: elaborado pelo autor

Tabela 4.5 – Variáveis Selecionadas Base de dados 2

(N) Variável	Nome da Variável	Descrição
1°	IDADE	Idade do Estudante
2°	TELEVISAO	Na sua casa tem televisão?
3°	FREEZER	Na sua casa tem freezer?
4°	MORA_MAE	Você mora com sua mãe?
5°	MAE_ALFABETIZADA	Sua mãe ou a mulher responsável por você sabe ler e escrever?
6°	INCENTIVO_ESTUDAR	Seus pais ou responsáveis incentivam você a estudar?
7°	INCENTIVO_DEVER	Seus pais ou responsáveis incentivam você a fazer o dever de casa e os trabalhos da escola?
8°	INCENTIVO_LER	Seus pais ou responsáveis incentivam você a ler?
9°	INCENTVO_ESCOLA	Seus pais ou responsáveis incentivam você a ir a escola e não faltar às aulas?
10°	LER_JORNAIS	Você lê: Jornais (inclusive os de distribuição gratuita).
11°	IR_CINEMA	Você Costuma: Ir ao cinema.
12°	IR_ESPETACULOS	Você Costuma: Ver apresentações musicais ou de dança.
13°	TRABALHA_FORA	Você trabalha fora de casa?
14°	REPROVACAO	Você já foi reprovado?
15°	ABANDONO	Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?
16°	DEVER_LP	Você faz o dever de casa de língua portuguesa?
17°	USA_BIBLIOTECA	Você utiliza a biblioteca ou sala de leitura da sua escola

Fonte: elaborado pelo autor

Tabela 4.6 – Variáveis Selecionadas Base de dados 3

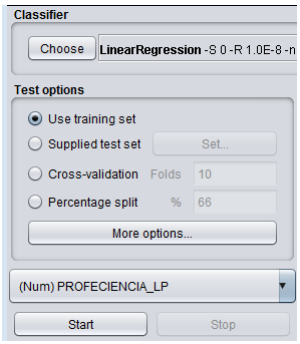
(N) Variável	Nome da Variável	Descrição
1°	ID_TURNO	Turno da Turma
2°	IDADE	Idade do Estudante
3°	TELEVISAO	Na sua casa tem televisão?
4°	COMPUTADOR	Na sua casa tem computador?
5°	MORA_MAE	Você mora com sua mãe?
6°	MAE_ALFABETIZADA	Sua mãe ou a mulher responsável por você sabe ler e escrever?
7°	PAI_ESTUDOU	Até que série seu pai ou homem responsável por você estudou?
8°	INCENTVO_ESCOLA	Seus pais ou responsáveis incentivam você a ir a escola e não faltar às aulas?
9°	IR_BIBLIOTECA	Você Costuma: Frequentar bibliotecas.
10°	TRABALHA_FORA	Você trabalha fora de casa?
11°	CORRIGE_LP	O professor corrige o dever de casa de língua portuguesa?
12°	CORRIGE_MT	O professor corrige o dever de casa de matemática?

Fonte: elaborado pelo autor

Depois de utilizar weka para selecionar atributos com relação a Proficiência dos alunos, chega a vez de aplicarmos o método de regressão para seja investigado os resultados obtidos, e consiga ser feito uma interpretação com valores retornados pela aplicação do método.

Como é mostrado na Figura 4.2, vai ser utilizado o método de regressão linear, e foi selecionado a opção “Use Training set” tendo como função utilizar apenas a base dados usada.

Figura 4.2 – Uso do Método de Regressão Linear



Fonte: Sistema Weka

5 Resultados e Avaliação

Os resultados da criação dos modelos utilizando o método de regressão linear estão nas tabelas a seguir. O método descarta variáveis que não formam um bom modelo.

5.1 Base de dados 1 (Proficiência_LP: $\geq 0 \leq 199$)

Tabela 5.1 – Resultado da Regressão Linear (Base 1)

Variável	***Descrição***	***Resultado***
INCENTIVO_DEVER	Seus pais ou responsáveis incentivam você a fazer o dever de casa e os trabalhos da escola?	3.0566
INCENTIVO_ESTUDAR	Seus pais ou responsáveis incentivam você a estudar?	2.6353
MAE_ALFABETIZADA	Sua mãe ou a mulher responsável por você sabe ler e escrever?	2.6926
DEVER_LP	Você faz o dever de casa de língua portuguesa?	2.1273
MORA_MAE	Você mora com sua mãe?	1.4101
ABANDONO	Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?	- 4.2658
TRABALHA_FORA	Você trabalha fora de casa?	- 3.5909
REPROVACAO	Você já foi reprovado?	- 3.3185

Fonte: elaborado pelo autor.

Antes de fazer a interpretação dos fatores selecionados, é importante entender o significado dos números que foram retornados pela aplicação da técnica de regressão linear, assim sendo mais especifico com relação aos resultados. Podemos utilizar como exemplo, o fator INCENTIVO_ESTUDAR, que além de selecionado obteve o valor positivo 2.6353, onde mostra que o fator INCENTIVO_ESTUDAR aumenta 2.6353 na proficiência em língua portuguesa do aluno, outro fator que pode ser utilizado como exemplo, é o atributo REPROVACAO que por sua vez obteve o valor negativo -3.3185, que significa a diminuição de -3.3185 na proficiência do estudante. A exemplificação anterior pode ser utilizada para explicar todos os valores obtidos nas bases dados 1, 2 e 3.

Analisando os resultados da primeira base de dados, foram selecionadas as variáveis que mais influenciam no desempenho do aluno com relação a prova do SAEB de português,

fazendo a interpretação sobre as informações contidas, fica evidente que pais que incentivam os filhos a estudar, resolver as atividades e trabalhos escolares, e ainda sendo mais especifico, que incentivam a fazer o dever de case da disciplina de língua portuguesa, tem grande participação nos resultados do desempenho do estudante de ensino fundamental.

Observando os outros fatores contidos na tabela 5.1, estudantes do EF que possuem mães alfabetizadas que no caso saibam ler e escrever, podem influenciar diretamente no desempenho do aluno e também outro fator que pode ser importante é se o estudante morar com ela, interpretando assim, mães com nível de escolaridade, incentivam seus filhos a estudar, a resolverem deveres e trabalhos escolares, refletindo diretamente no que foi dito no parágrafo anterior. Portanto a mãe é uma variável muito importante na vida estudantil do aluno.

Outros atributos que foram selecionados como influenciadores, que podem justificar o desempenho baixo no SAEB de língua portuguesa desse primeiro grupo de estudantes, é que variáveis como abandono, que representam o abandono do aluno de ir a escola, a reprovação do estudante no ensino fundamental estão relacionados diretamente com o desempenho do aluno. A variável trabalha fora que também foi selecionada indica que os estudantes desse primeiro grupo, tem indícios que já executam alguma função, podendo influenciar em seus estudos, tendo menos tempo para se dedicar a escola, afetando seu desempenho, e podendo está relacionado diretamente com os motivos do abandono do aluno na escola.

5.2 Base de dados 2 (Proficiência_LP: $\geq 200 \leq 300$)

Tabela 5.2 – Resultado da Regressão Linear (Base 2)

Variável	***Descrição***	***Resultado***
INCENTIVO_ESTUDAR	Seus pais ou responsáveis incentivam você a estudar?	3.7933
DEVER_LP	Você faz o dever de casa de língua portuguesa?	3.4845
INCENTIVO_DEVER	Seus pais ou responsáveis incentivam você a fazer o dever de casa e os trabalhos da escola?	1.9093
MAE_ALFABETIZDA	Sua mãe ou a mulher responsável por você sabe ler e escrever?	1.6428
MORA_MAE	Você mora com sua mãe?	1.4101
IDADE	Idade do Estudante	0.7557
LER_JORNAIS	Você lê: Jornais (inclusive os de distribuição gratuita).	- 1.1092
USA_BIBLIOTECA	Você utiliza a biblioteca ou sala de leitura da sua escola?	- 1.6843
TRABALHA_FORA	Você trabalha fora de casa?	- 2.346
REPROVACAO	Você já foi reprovado?	- 3.2998

Fonte: elaborado pelo autor.

Semelhante ao primeiro grupo, a segunda base de dados também selecionou como importantes influenciadores no desempenho, os atributos em que os estudantes são incentivados pelos pais a estudar, resolverem atividades e trabalhos escolares, principalmente sendo de língua portuguesa, e que morar com a mãe e que ela seja alfabetizada fazem toda a diferença no desempenho escolar do aluno.

No grupo 2 foi selecionado o atributo idade, que por sua vez pode demonstrar que a idade do aluno podem influenciar diretamente no desempenho, assim alunos que estão cursando o ensino fundamental em seu tempo certo, podem ter performance melhor do que estudantes que estão cursando o ensino fundamental com idade avançada, demonstrando reprovação nas séries do EF. Como prova do que foi dito, mostrando confiança nos resultados obtidos, este grupo também selecionou como fator importante para o desempenho, a variável reprovação e trabalha fora que podem ser fatores resultantes do grupo ter nota razoável no SAEB.

Também foram selecionados os atributos ler jornais e ir a biblioteca, que demonstram estudantes que fazem o exercício de ler jornais ou livros da biblioteca, podem influenciar diretamente no seu desempenho de português, que no caso do grupo 2 esses atributos podem estar demonstrando a falta de leitura de jornais e livros, resultando em maioria dos estudantes ter

obtido nota razoável no SAEB.

5.3 Base de dados 3 (Proficiência_LP: > 300)

Tabela 5.3 – Resultado da Regressão Linear (Base 3)

Variável	***Descrição***	***Resultado***
MAE_ALFABETIZADA	Sua mãe ou a mulher responsável por você sabe ler e escrever?	2.4156
MORA_MAE	Você mora com sua mãe?	2.3284
IR_BIBLIOTECA	Você Costuma: Frequentar bibliotecas?	1.6321
CORRIGE_LP	O professor corrige o dever de casa de língua portuguesa?	1.4671
IDADE	Idade do Estudante	- 0.8428
PAI_ESTUDOU	Até que série seu pai ou homem responsável por você estudou?	- 0.8983

Fonte: elaborado pelo autor.

Analisando o último grupo, a onde estão os alunos que obtiveram maiores notas no SAEB, assim como nos outros grupos aqui também a mãe é um dos fatores que influenciam no desempenho do aluno, mostrando que a mãe alfabetizada que sabe ler e escrever e o fato do aluno morar com a mãe, incentiva desde dos primeiros anos escolares da vida do estudante.

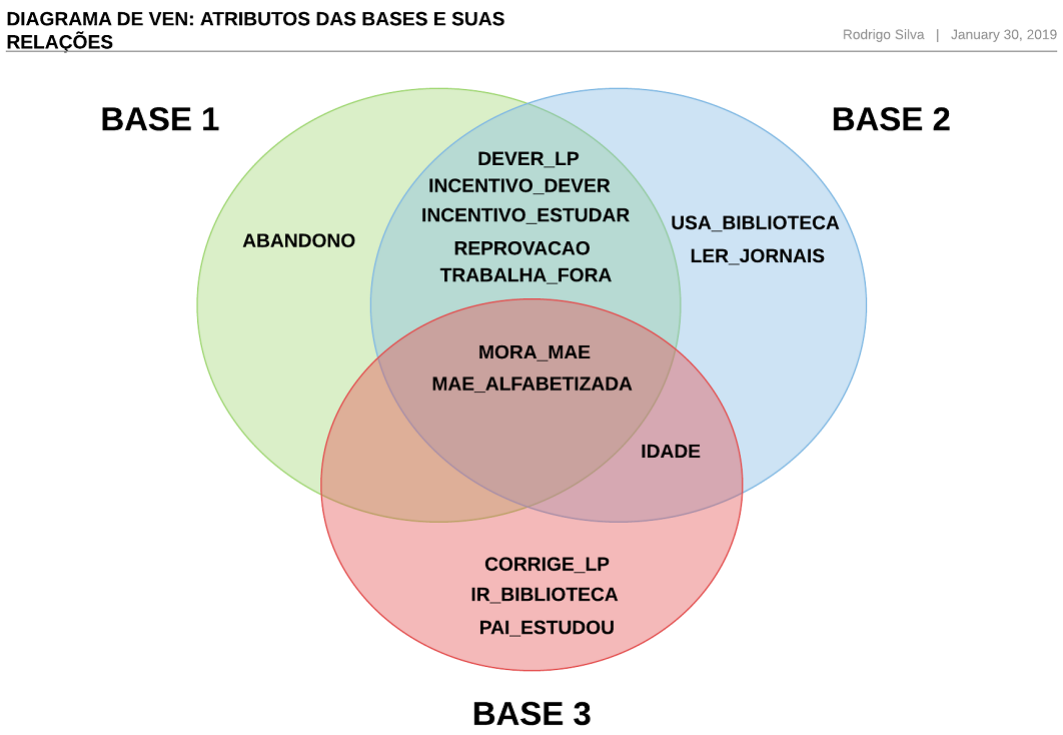
O grupo 3 possui a variável ir a biblioteca, que pode ser visto como os alunos contidos nesse grupo, fazem uso frequente da biblioteca para ler livros, mostrando grandes influencias no seu desempenho. Outra variável aqui selecionada é se o professor corrige o dever de língua portuguesa, podendo ser analisado que alunos que fazem as atividades de português, e são corrigidas pelo professor, também faz grande importância para a evolução do aluno. Estas características podem ser um dos fatores que fazem este grupo ter bom desempenho no SAEB.

As variáveis idade e nível de escolaridade do pai, também fazem parte do modelo do grupo 3. O atributo idade demonstra a importância do aluno está na série certa na escola referente a sua idade, já a questão do nível de escolaridade do pai, aqui nesse caso pode ser interpretado como não fazer diferença no desempenho do aluno, mas este fato não pode ser analisado como certeza absoluta, já que foram utilizadas técnicas estatísticas que tentam dar um embasamento sobre os dados.

5.4 Modelo de Relação entre os Grupos

Através do diagrama de Venn é possível analisar quais fatores cada modelo das bases de dados possui, bem como suas relações. Assim como é definido, o Diagrama de Venn é um método de conjuntos, que por sua vez também podem ser representado por figuras curvas fechadas (Elipse) que representam as relações que existem entre os grupos especificados. (MEDEIROS, 2016)

Figura 5.1 – Diagrama de Ven



Fonte: Elaborado Pelo Autor

Como é mostrado na figura 5.1 podemos ver os fatores de cada base de dados obteve através das técnicas de mineração de dados, criando assim um modelo. É possível verificar que as bases 1 e 2, possuem fatores únicos em seus modelos, mas existe grandes quantidades de fatores que se igualam nas duas bases referidas, mostrando a grande influência que esses atributos possuem para esses alunos, que fazem parte desses grupos. É mostrado também, os fatores únicos que influenciam a base 3, na qual está os alunos que obtiveram melhor desempenho. A onde pode ser observado, que as três bases de dados possuem os fatores (MAE_ALFABETIZADA e MORAR_MAE) presentes em seus modelos, mostrando que além de seus fatores únicos, todos os alunos independente em qual grupo está encaixado, possuem características iguais que

influenciem em seu desempenho escolar.

De acordo com a análise feita anteriormente, é notado que além das peculiaridades que cada grupo possui, existem semelhanças entres esses conjuntos, em que a partir dessas igualdades de atributos podem ser criado um modelo único que mostra independente dos grupos formados, quais variáveis podem diagnosticar o desempenho escolar do estudante no ensino fundamental.

Tabela 5.4 – Modelo Unificado da Relação entre os Grupos

Variável	Descrição
INCENTIVO_DEVER	Seus pais ou responsáveis incentivam você a fazer o dever de casa e os trabalhos da escola?
INCENTIVO_ESTUDAR	Seus pais ou responsáveis incentivam você a estudar?
MAE_ALFABETIZADA	Sua mãe ou a mulher responsável por você sabe ler e escrever?
MORA_MAE	Você mora com sua mãe?
TRABALHA_FORA	Você trabalha fora de casa?
REPROVACAO	Você já foi reprovado?

Fonte: elaborado pelo autor.

Assim com a criação do modelo unificado entre os grupos, fica evidente que o incentivo dos pais para estudar e a resolução de atividades escolares do estudante são um fator imprescindível para a evolução estudantil do aluno. Outro fator que se mostrou importante em todos os grupos é o fato de morar com a mãe e dela ser alfabetizada, mostrando influência direta para os alunos. Os atributos trabalha fora e reprovação se mostraram relevantes em questão ao desempenho do aluno, que foi compreendido que alunos que executam alguma ação com relação a trabalho, podem ter menos tempo para se dedicar ao estudos. Já no caso da reprovação podem mostrar desinteresse do aluno com relação escola, influenciando diretamente o desempenho.

6 Considerações Finais

6.1 Dificuldades Encontradas

Na elaboração desta pesquisa, teve como dificuldades analisar boa parte das bases de dados disponíveis pelo INEP e escolher através dos atributos e valores contidos nesses bancos, qual base de dados apresentava melhor qualidade para que fosse feito o processo de mineração, a onde foram feitos vários testes o que levou bastante tempo e trabalho árduo para que fosse escolhida a base certa.

Outro fator que dificultou o elaboração do projeto, foi com as grandes quantidades de dados, o computador utilizado sofreu com alguns travamentos, onde tiveram que ser feito reajustes nas configurações do sistema operacional para que utilizasse menos recursos e pudesse ser executado as bases dados juntos com os programas escolhidos (Weka, SPSS).

6.2 Conclusão

Este trabalho apresentou a criação de um modelo computacional para identificar quais fatos (Variáveis) influenciam no desempenho do aluno, a partir da análise de dados dos alunos no 5 ° ano do ensino fundamental (SAEB 2015), e técnicas de mineração de dados.

O modelo criado pode ser considerado um modelo genérico, pois os atributos que foram selecionados com a utilização das técnicas de mineração de dados, dizem respeito a fatos que influenciam o estudante do EF no geral. Assim este trabalho não contribui apenas como uma proposta sobre o desempenho do aluno em português, e sim sobre o desempenho do aluno em seja qual for á disciplina do ensino fundamental, vindo a servir para professores de todas as disciplinas do EF, administradores e gestores de ensino, bem como pesquisadores e profissionais que se interessam pela área da educação e desempenho educacional. Os achados deste trabalho contribuem de forma convincente, tanto do ponto de vista da modelagem, quanto da criação de um modelo de desempenho educacional.

6.3 Contribuições deste trabalho

O Trabalho aqui proposto, teve o objetivo de contribuir para da apoio a decisão e estudos de gestores, professores e todos os profissionais que se dedicam a estudar o desempenho dos alunos. sendo os esses os pontos alcançados:

- Desenvolvimento de um modelo de desempenho para auxiliar os gestores educacionais e professores na tomada de decisão.
- Referência para pesquisadores na área de educação.
- Publicação de artigos científicos na área de mineração de dados educacionais.

6.4 Proposta para trabalhos futuros

- implementar o modelo em forma de aplicação para tomada de decisão.
- Verificar o desempenho dos alunos do ensino fundamental de pernambuco fazendo a divisão de gênero.
- Fazer Análise de desempenho dos alunos, adicionando fatores estruturais da escola e qualidade ensino dos seus docentes.

REFERÊNCIAS BIBLIOGRÁFICAS

ACKERMAN, B. P. et al. Relation between reading problems and internalizing behavior in school for preadolescent children from economically disadvantaged families. *Child Development*, Wiley Online Library, v. 78, n. 2, p. 581–596, 2007.

AIKENS, N. L.; BARBARIN, O. Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology*, American Psychological Association, v. 100, n. 2, p. 235, 2008.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, v. 19, n. 02, p. 03, 2011.

BYRNE, D. G. et al. Stressor experience in primary school-aged children: Development of a scale to assess profiles of exposure and effects on psychological well-being. *International Journal of Stress Management*, Educational Publishing Foundation, v. 18, n. 1, p. 88, 2011.

CASTANHEIRA, L. G. Aplicação de técnicas de mineração de dados em problemas de classificação de padrões. *Belo Horizonte: UFMG*, 2008.

CHAPMAN, P. et al. The crisp-dm user guide. In: *4th CRISP-DM SIG Workshop in Brussels in March*. [S.l.: s.n.], 1999. v. 1999.

CÔRTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. *Mineração de dados-funcionalidades, técnicas e abordagens*. [S.l.]: PUC, 2002.

CURRAL, J. Statistics packages: A general overview. *Maths&Stats Newsletter*, 1994.

FAYYAD, U. M. et al. Knowledge discovery and data mining: Towards a unifying framework. In: *KDD*. [S.l.: s.n.], 1996. v. 96, p. 82–88.

FELÍCIO, F. d.; TERRA, R.; ZOGHBI, A. C. The effects of early childhood education on literacy scores using data from a new brazilian assessment tool. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 42, n. 1, p. 97–128, 2012.

FERREIRA, D. F. *Estatística multivariada*. [S.l.]: Editora Ufla Lavras, 2008.

GALVÃO, N. D.; MARIN, H. d. F. Técnica de mineração de dados: uma revisão da literatura. *Acta Paulista de Enfermagem*, Escola Paulista de Enfermagem, Universidade Federal de São Paulo (UNIFESP), 2009.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R. V. Avaliação de desempenho de estudantes em cursos de educação a distância utilizando mineração de dados. In: *Anais do Workshop de Desafios da Computação Aplicada à Educação*. [S.l.: s.n.], 2012. p. 30–39.

HALLINGER, P.; HECK, R. H. What do you call people with visions? the role of vision, mission and goals in school leadership and improvement. In: *Second international handbook of educational leadership and administration*. [S.l.]: Springer, 2002. p. 9–40.

IBM. Notas estatísticas: Censo escolar 2017. 2016. Disponível em: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf>.

INEP. Censo escolar 2017: Notas estatísticas. 2017. Disponível em: <http://download.inep.gov.br/educacao_basica/censo_escolar/notas_estatisticas/2018/notas_estatisticas_Censo_Escolar_2017.pdf>.

MACEDO, D. C.; MATOS, S. N. Extração de conhecimento através da mineração de dados. *Revista de Engenharia e Tecnologia*, v. 2, n. 2, p. Páginas–22, 2010.

MANNILA, H. Data mining: machine learning, statistics, and databases. In: IEEE. *Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on*. [S.l.], 1996. p. 2–9.

MARTURANO, E. M.; PIZATO, E. C. G. Preditores de desempenho escolar no 5º ano do ensino fundamental. *Psico*, Pontificia Universidade Catolica do Rio Grande, v. 46, n. 1, p. 16–24, 2015.

MARTURANO, E. M. et al. Estresse cotidiano na transição da 1ª série: percepção dos alunos e associação com desempenho e ajustamento. *Psicologia: Reflexão e Crítica*, Curso de Pós-Graduação em Psicologia da Universidade Federal do Rio Grande ..., v. 22, n. 1, p. 93–101, 2009.

MEC, M. d. E. Ministério da educação. 2018. Disponível em: <https://www.mec.gov.br>.

MEDEIROS, R. J. J. Matemática ii. 2016.

MINAYO, M. C. d. S.; SANCHES, O. Quantitativo-qualitativo: oposição ou complementaridade? *Cadernos de saúde pública*, SciELO Public Health, v. 9, p. 237–248, 1993.

MORO, S.; LAUREANO, R.; CORTEZ, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: EUROSIS-ETI. *Proceedings of European Simulation and Modelling Conference-ESM'2011*. [S.l.], 2011. p. 117–121.

NACIONAIS, I. A. P. C. terceiro e quarto ciclos do ensino fundamental. *Brasília: MEC-Secretaria de Educação Fundamental*, 1998.

OLIVEIRA, R. Portela de. Da universalização do ensino fundamental ao desafio da qualidade: uma análise histórica. *Educação & Sociedade*, SciELO Brasil, v. 28, n. 100, 2007.

PESTANA, M. H.; GAGEIRO, J. N. Análise de dados para ciências sociais: a complementaridade do spss. Sílabo Lisboa, 2003.

QUILICI-GONZALEZ, J. A.; ZAMPIROLLI, F. de A. *Sistemas inteligentes e mineração de dados*. [S.l.: s.n.], 2015.

SAEB, E. B. Sistema de avaliação da educação básica. 2018. Disponível em: <http://portal.inep.gov.br/educacao-basica/saeb>.

SALLES, J. F. d.; PARENTE, M. A. d. M. P.; FREITAS, L. B. d. L. Leitura/escrita de crianças: comparações entre grupos de diferentes escolas públicas. *Paidéia: cadernos de educação da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto. Ribeirão Preto, SP. Vol. 20, n. 47 (2010), p. 335-344*, 2010.

SANTOS, M. Y.; RAMOS, I. *Business Intelligence: Tecnologias da informação na gestão de conhecimento*. [S.l.]: FCA-Editora de Informática, Lda, 2006.

SILVA, S. da; MONTEIRO, S. S.; RODRIGUES, M. F. A importância da educação infantil para o pleno desenvolvimento da criança. *Revista Mosaico*, v. 8, n. 2, p. 30–38, 2017.

WEBBER, C. G. et al. Utilização de algoritmos de agrupamento na mineração de dados educacionais. *RENOTE*, v. 11, n. 1, 2013.

WEIAND, A.; PINTO, A. dos S. O aluno de ead em fóruns do ava moodle: um estudo sobre suas competências. *CEP*, v. 95520, p. 000.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.