UNIVERSIDADE DO ESTADO DO AMAZONAS
CENTRO DE ESTUDOS SUPERIORES DE ITACOATIARA
CURSO DE LICENCIATURA EM COMPUTAÇÃO

DANRLEY CARVALHO DOS SANTOS

MINERAÇÃO DE DADOS EDUCACIONAIS: UMA ANÁLISE SOBRE AS VARIÁVEIS QUE INFLUENCIAM NA EVASÃO ESCOLAR

ITACOATIARA/AM

2019

DANRLEY CARVALHO DOS SANTOS

MINERAÇÃO DE DADOS EDUCACIONAIS: UMA ANÁLISE SOBRE AS VARIÁVEIS QUE INFLUENCIAM NA EVASÃO ESCOLAR

Projeto de pesquisa apresentado como requisito para aprovação na disciplina de Projeto Orientado em Informática na Educação II do curso de Licenciatura em Computação do Centro de Estudos Superiores de Itacoatiara da Universidade do Estado do Amazonas. .

Orientador: Me. João da Mata Libório Filho

ITACOATIARA/AM

2019

Ficha Catalográfica
Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).
Sistema Integrado de Bibliotecas da Universidade do Estado do Amazonas.

S237m Santos, Danrley Carvalho dos
MINERAÇÃO DE DADOS EDUCACIONAIS : UMA ANÁLISE SOBRE AS VARIÁVEIS QUE INFLUENCIAM NA EVASÃO ESCOLAR / Danrley Carvalho dos Santos. Manaus : [s.n], 2019.
41 f.: color.; 30 cm.

TCC - Graduação em Licenciatura em Computação - Licenciatura - Universidade do Estado do Amazonas, Manaus, 2019.
Inclui bibliografia
Orientador: Libório Filho, João da Mata

□1. Evasão Escolar. 2. Mineração de dados educacionais. 3. Aprendizado de Máquina. I. Libório Filho, João da Mata (Orient.). II. Universidade do Estado do Amazonas. III. MINERAÇÃO DE DADOS EDUCACIONAIS

Elaborado por Jeane Macelino Galves - CRB-11/463

MINERAÇÃO DE DADOS EDUCACIONAIS: UMA ANÁLISE SOBRE AS VARIÁVEIS QUE INFLUENCIAM NA EVASÃO ESCOLAR

DANRLEY CARVALHO DOS SANTOS

Projeto de pesquisa apresentado como requisito para aprovação na disciplina de Projeto Orientado em Informática na Educação II do curso de Licenciatura em Computação do Centro de Estudos Superiores de Itacoatiara da Universidade do Estado do Amazonas, sob orientação do Prof. Msc. João da Mata Libório Filho.

João da Mata Libório Filho (Orientador)

Jhonathan Araújo Oliveira (Membro da Banca)

Kayro Figueira Pires (Membro da Banca)

Itacoatiara – Amazonas

2019

AGRADECIMENTOS

Primeiramente, agradeço a Deus pela conclusão deste trabalho. Minha vida e tudo que obtive nela devo a Ele, então a Ele toda a glória.

Em segundo lugar, agradeço ao meu alicerce que me permitiu chegar até aqui: minha família. Obrigado minha Mãe Amélia Góis e meu pai Sidney Gama, sem vocês eu não teria conseguido. Meus irmão Denis Natan e Denner Elias. Não poderia deixar de lembrar do meu primo Mauro Gama, junto a sua esposa Elizane Amazonas, por me acolherem durante três anos em sua residência. Vocês se tornaram minha segunda primeira família. Obrigado também minha filha de quatro patas Pandora, que sempre me acolheu com latidos carinhosos ao chegar em casa.

Agradeço também a algumas das pessoas mais importantes que conheci durante esta caminhada: Paula Oliveira, grande amiga de todas as horas do primeiro ao último período. Nossas histórias serão levadas em meu coração após a conclusão deste curso; Taíza Oliveira, minha grande amiga que conheci na reta final, mas que com uma sinergia surpreendente se aproximou e me ajudou a superar momentos difíceis encontrados durante a conclusão do curso: Obrigado! Você será sempre lembrada com muito carinho e tenho certeza que nossa amizade se estenderá por longos anos. Milena Vasconcelos, uma amiga que a UEA me proporcionou e que ficou sempre próxima, ajudando com conselhos sobre a pesquisa e algumas revisões imprevistas.

Estendo meu agradecimento ainda a algumas pessoas que sempre se fizeram presentes durante este processo de formação, alguns amigos que tenho como presente em minha vida: Larissa Soares, Brunna Michaella, Kelly Adriana, Augusto Lima, Árthur Souza, Maria Irlene, Alexandre Castro e Samara Santarém e valcilene Brito.

Aos meus docentes que foram quem me moldaram durante esta formação: Meu orientador João da Mata Libório Filho, melhor professor de programação. Jhonathan Araújo Oliveira, com conhecimentos de sobra em redes. Marcelo Carvalho Tavares, grande mestre de Banco de Dados. Elisangela Oliveira, a professora de humanas. Luiz Sérgio, professor famoso. Professor Kayro e demais docentes do CESIT.

E aos que por ventura esqueci, recebam meu muito obrigado!

RESUMO

A evasão escolar é um grande problema que preocupa o cenário educacional brasileiro. Neste trabalho é apresentado alguns fatores que causam essa preocupação, e uma solução para parte do problema é apresentada com a utilização de Mineração de Dados Educacionais. O objetivo consiste na identificação de variáveis relacionadas com a evasão escolar de alunos do ensino fundamental de escolas do Estado do Amazonas e para a criação de um modelo de predição da probabilidade de evasão destes alunos utilizando métodos de aprendizagem de máquina. A metodologia foi baseada no modelo CRISP-DM, bastante utilizado para mineração de dados. Foram treinados e avaliados três modelos de aprendizado para predição: Regressão Logística, Árvores de decisão e Florestas Aleatórias. Dentre as variáveis utilizadas, é notável que grande parte delas correspondem a dados de localização das escolas, equipamentos das mesmas e a faixa etária dos alunos.

Palavras-chaves: Evasão Escolar. Mineração de Dados Educacionais. Aprendizado de Máquina.

ABSTRACT

The school evasion is a big problem that concerns the Brazilian educational scenario. This paper presents some factors that cause this concern, and a solution to part of the problem is presented by using Educational Data Mining. The objective is to identify variables related to school dropout of elementary school students from Amazonas State schools and to create a model for predicting the probability of dropout using machine learning methods. The methodology was based on the CRISP-DM model, widely used for data mining. Three prediction learning models were trained and evaluated: Logistic Regression, Decision Trees and Random Forests. Among the variables used, it is noteworthy that most of them correspond to school location data, school equipment and student age.

Key-words: School evasion. Data Mining. Machine Learning.

LISTA DE ILUSTRAÇÕES

FIGURA 1 – A HIERARQUIA DO APRENDIZADO 17
FIGURA 2 – O CLASSIFICADOR À DIREITA FORNECE UMA INTERPRETAÇÃO COMPACTA DOS DADOS 18
FIGURA 3 – PROCESSO DE KDD 20
FIGURA 4 – O PROCESSO CRISP-DM 21

FIGURA 5 – METODOLOGIA 26
FIGURA 6 – DADOS BRUTOS: MATRÍCULAS DE 2014 27
FIGURA 7 – ESTADOS NO ARQUIVO DE DADOS BRUTO 28
FIGURA 8 – DADOS NULOS: MAPA DE CALOR 29
FIGURA 9 – VERIFICANDO SE O ALUNO EVADIU 29
FIGURA 10 – DIVISÃO DE DADOS DE TREINO E TESTE 30
FIGURA 11 – MATRIZ DE CONFUSÃO: FUNCIONAMENTO DA MATRIZ 32

FIGURA 12 – VARIÁVEIS SELECIONADAS 33
FIGURA 13 – SUBAMOSTRAGEM - 2014 35
FIGURA 14 – SUBAMOSTRAGEM - 2015 35
FIGURA 15 – RESULTADOS - 2014 36
FIGURA 16 – RESULTADOS - 2015 36

LISTA DE TABELAS

LISTA DE ABREVIATURAS E DE SIGLAS

CRISP-DM Cross-Industry Standard Processo of Data Mining

CSV Comma Separated Values

ECA Estatuto da Criança e do Adolescente

EJA Educação de Jovens e Adultos

EaD Ensino à distância

HTML Hypertext Markup Language

IA Inteligência Artificial

INEP O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDD Knowledge Discovery in Databases

MDE Mineração de Dados Educacionais

ML Machine Learning

NASA National Aeronautics and Space Administration

PDF Portable Document Format

Pnud Programa das Nações Unidas para o Desenvolvimento

SUMÁRIO

1 INTRODUÇÃO 9
1.1 CONTEXTUALIZAÇÃO E CARACTERIZAÇÃO DO PROBLEMA 9
1.2 JUSTIFICATIVA 10
1.3 OBJETIVO GERAL 11
1.4 OBJETIVOS ESPECÍFICOS 11
1.5 ORGANIZAÇÃO DO TRABALHO 11

2 FUNDAMENTAÇÃO TEÓRICA 13
2.1 EVASÃO ESCOLAR 13
2.1.1 FATORES QUE INFLUENCIAM À EVASÃO ESCOLAR 13
2.1.2 COMBATE À EVASÃO ESCOLAR NO BRASIL 14
2.2 APRENDIZADO DE MÁQUINA 15
2.2.1 APRENDIZADO INDUTIVO 15
2.2.2 APRENDIZADO DE MÁQUINA INDUTIVO POR EXEMPLOS 16
2.2.2.1 APRENDIZADO NÃO SUPERVISIONADO 16
2.2.2.2 APRENDIZADO SUPERVISIONADO 17
2.2.2.3 PARADIGMAS DO APRENDIZADO DE MÁQUINA 17
2.3 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS 19
2.3.1 CRISP-DM 20
2.4 MINERAÇÃO DE DADOS 22
2.4.1 TAREFAS 22
2.5 TRABALHOS RELACIONADOS 23

3 METODOLOGIA 25
3.1 REVISÃO BIBLIOGRÁFICA 25
3.2 FERRAMENTAS E BASE DE DADOS 25
3.3 ETAPAS DA PESQUISA 26

4 RESULTADOS 33

5 CONSIDERAÇÕES FINAIS 37

REFERÊNCIAS 38

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E CARACTERIZAÇÃO DO PROBLEMA

Segundo um levantamento realizado pelo Programa das Nações Unidas para o Desenvolvimento (Pnud), em 2012 o Brasil teve a terceira maior taxa de evasão escolar entre 100 países, com uma taxa de Evasão de 24,3% (UOL, 2013). Em 2017, o INEP divulgou dados que revelam que 12,9% e 12,7% dos alunos matriculados na 1ª e 2ª série do Ensino Médio, respectivamente, evadiram da escola de acordo com o Censo Escolar entre os anos de 2014 e 2015. Um dado que também é bastante preocupante dentre estes é referente ao 9º ano do ensino fundamental, que apresentou a terceira maior taxa de evasão, 7,7% (INEP, 2019).

O abandono e evasão escolar são problemas graves para a educação, sendo considerado abandono o ato de um aluno deixar de frequentar a escola, e evasão quando um aluno abandona os estudos e não retorna no ano seguinte (BRASIL, 2015).

Os motivos de considerar as situações acima citadas como problemas graves são variados, dentre os quais pode-se citar que os mesmos causam prejuízos maiores na vida do estudante, como o atraso escolar, que ocorre quando um aluno que deveria estar no ensino médio ainda frequenta o ensino fundamental, por exemplo; dificuldade de concluir a educação básica, entre outros (CALIXTO; SEGUNDO; GUSMÃO, 2017).

As maneiras de combater a evasão no país são através de algumas políticas públicas, que ajudam com a renda de famílias carentes e exigem um certo índice de frequência escolar dos menores de idade para que o auxílio não seja bloqueado. Entretanto, mesmo com essa e outras medidas, ainda há um grupo considerável de pessoas cometendo evasão. O que se pode supor a partir disso, é que há pessoas em risco de evasão que não estão recebendo amparo social, portanto, não foram identificadas com chance de evadir no futuro. Devido a isto, faz-se necessário pesquisas para uma melhor identificação destes sujeitos, que permita reconhecer com antecedência quem pode evadir no futuro.

A Mineração de Dados Educacionais (MDE) é uma área que possui o objetivo de desenvolver métodos para exploração de dados oriundos de ambientes educacionais (BAKER; ISOTANI; CARVALHO, 2011). A MDE possibilita, através do estudo de dados, apontar caminhos objetivos para a melhoria da educação. No Brasil, a MDE está sendo explorada principalmente para estudo de dados gerados no Ensino Superior pela modalidade Ensino à distância (EaD), havendo poucos estudos sobre a educação básica (TORRES MASCHIO et al., 2018).

Uma das aplicações da MDE úteis como parte da solução de identificar sujeitos em risco de evasão, é a predição, que através de um conjunto de entrada de dados sobre alunos que já cometeram evasão, consegue apontar com certo grau de precisão se determinado aluno está com alta ou baixa probabilidade de evadir.

Diante do exposto, este trabalho utilizou Mineração de Dados Educacionais para analisar quais variáveis presentes no Censo Escolar divulgado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) tem relação com a Evasão escolar e essas variáveis serviram como base para a criação de um modelo de predição. Para isso, foram utilizados dados de alunos dos anos de 2014 e 2015 do ensino fundamental do estado do Amazonas.

Foram utilizados três algoritmos de classificação para realizar a predição: Regressão logística, Árvores de decisão e Florestas aleatórias. O desempenho de cada algoritmo pode ser encontrado na seção de resultados.

1.2 JUSTIFICATIVA

A necessidade de medidas para combater a evasão escolar no contexto educa- cional brasileiro é bastante evidente, devido a gravidade que tal problema representa para o cenário. É notável que algumas medidas já são utilizadas, porém os métodos utilizados ainda não são o ideal, pois o grupo de risco a evasão é identificado através de meios ortodoxos, que demandam muito trabalho manual por parte principalmente dos professores (FERREIRA, 2000), de modo especial no Brasil onde ainda há um índice elevado de evasão. Isso resulta em perda de desempenho nas demais atividades dos sujeitos responsáveis pela identificação.

Nesse contexto, precisão e agilidade são indispensáveis para a identificação destes alunos. Com isso, a utilização de técnicas de Machine Learning (ML) apresenta-se como uma boa solução para obtenção de respostas precisas sem custo de tempo elevado, como se pode observar no relatório do fórum econômico mundial intitulado "*This AI outperformed 20 corporate lawyers at legal work*" (Esta IA superou 20 advogados corporativos no trabalho legal, em tradução livre) (WOOD, 2018). Este relatório, além de mostrar que a IA pode ter maior eficiência para algumas tarefas, também evidencia o fato de que mesmo especialistas humanos estão sujeitos a falhas em suas áreas, e no caso da evasão, a possibilidade de identificar ou não alunos em risco, pois os docentes designados a esta tarefa precisam realizá-la em paralelo com muitas outras para as quais estão incumbidos.

Como já mencionado anteriormente, e de acordo com Calixto, Segundo e Gusmão (2017) os fatores que causam evasão são diversos, como a gravidez, baixa renda familiar (surgindo a necessidade de trabalhar durante o horário das aulas) entre

outras. Alguns desses fatores exigem medidas específicas para amenizar o problema da evasão, mas quando a causa desta não está clara, o processo de identificação se torna mais complexo, sendo necessária a análise de um conjunto de fatores distintos para identificar nos dados do aluno quais deles devem ser observados com mais cautela.

Esse processo pode ser feito através de cruzamento de informações, para verificar as variáveis que tem influência para a evasão, assim como identificar padrões em dados de um conjunto grande de alunos. A Mineração de Dados possui técnicas que possibilitam isto, e uma vez identificado um padrão, ela permite predizer (prever) a chance de determinado aluno evadir no futuro (CAMILO; SILVA, 2009).

Neste sentido, a Mineração de dados se apresentou como ferramenta com muito potencial para auxiliar no processo de identificação de sujeitos em risco de evasão, através de suas técnicas análise de dados, filtros e predição foi possível criar um modelo um tanto simples mas que já oferece um auxílio na resolução do problema apresentado. Com mais estudos a respeito do tema, não será necessário esperar que um aluno apresente uma quantidade grande de faltas na escola, ou problemas mais graves para que ele receba estímulos que o permita continuar em sua vida acadêmica.

1.3 OBJETIVO GERAL

Identificar variáveis relacionadas com a evasão escolar de alunos do ensino fundamental de escolas do Estado do Amazonas e criar um modelo de predição da probabilidade de evasão escolar.

1.4 OBJETIVOS ESPECÍFICOS

- Verificar por meio de métodos de mineração de dados quais variáveis estão relacionadas à evasão escolar de alunos do Ensino Fundamental no Estado do Amazonas;
- Criar um modelo de predição de evasão escolar utilizando métodos de aprendizagem de máquina e analisar sua acurácia.

1.5 ORGANIZAÇÃO DO TRABALHO

No capítulo introdutório, foram apresentadas as principais características deste trabalho, apresentando um contexto de aplicação, o problema, a justificativa desta pesquisa e os objetivos da mesma.

Além da introdução, este trabalho está organizado em outros três capítulos. Uma descrição geral de cada capítulo está descrita a seguir.

Capítulo 2 – Fundamentação Teórica: Apresenta alguns conceitos importantes para uma melhor compreensão desta pesquisa, aspectos teóricos nos quais ela foi embasada e ao final serão mostrados alguns trabalhos relacionados com nosso tema.

Capítulo 3 – Metodologia e Cronograma de Execução: Neste capítulo, é descrita a metodologia que será utilizada para a realização do projeto, e ao final é apresentado o cronograma para realização das atividades.

Capítulo 4 – Análise e discussão dos resultados: Neste capítulo são analisados os resultados obtidos com o desenvolvimento desta pesquisa.

Capítulo 5 – Considerações Finais: No último capítulo são efetuadas algumas considerações sobre este trabalho e seus resultados.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo estão descritos conceitos fundamentais para uma melhor compreensão deste trabalho.

2.1 EVASÃO ESCOLAR

O termo evasão escolar pode ser interpretado de maneiras diferentes (RIGO et al., 2014). Neste trabalho, adotaremos os critérios utilizados pelo Instituto Nacional de Estudos e Pesquisas (INEP), que denomina Abandono quando o aluno deixa de frequentar o ano letivo em que está matriculado, e evasão, quando o aluno além de deixar de frequentar o ano letivo, não retorna no ano seguinte (BRASIL, 2015).

Fini (1989) propõe que o fracasso escolar seja compreendido como o fracasso da escola, não do aluno. A autora aponta pesquisas que evidenciam que não há explicação cognitiva que justifique a dificuldade de aprendizagem em crianças economicamente menos favorecidas, portanto, o fracasso é da escola.

Ferreira (2000) menciona que segundo o Estatuto da Criança e do Adolescente (ECA), o problema deve ser partilhado entre família, comunidade, sociedade em geral e poder público. Um dos problemas em identificar fatores que influenciam na evasão, é que eles são dos mais variados, e um fator que era considerado importante há alguns anos atrás, pode não o ser atualmente, da mesma maneira que um problema considerado influente para a evasão na região Sul do país pode não influenciar tanto na região Norte. No entanto, existem ainda alguns consensos encontrados na bibliografia revisada. A seguir serão mostrados alguns deles.

2.1.1 Fatores que influenciam à evasão escolar

Bezerra et al. (2016) mostra que os trabalhos desenvolvidos no Brasil sobre evasão escolar costumam considerar três aspectos: 1) o indivíduo, 2) a escola, 3) o sistema de ensino. A autora enfatiza que ao se tratar do indivíduo, estamos nos referindo a sua vida no meio familiar, e neste caso o Estado só consegue agir de maneira indireta, através de políticas públicas de distribuição de renda que exigem frequência escolar de alunos menores de idade. Porém, o Estado pode agir de maneira direta nas duas outras perspectivas.

Uma classificação dos problemas gerais foi proposta por Ferreira (2000), que agrupou os problemas (ressaltando que eles não aparecem de maneira isolada, e sim

em conjunto) da seguinte maneira:

- **Escola:** Quando não é atrativa, autoritária, possui professores despreparados, etc.
- **Aluno:** Desinteressado, indisciplinado, com problema de saúde, gravidez, etc.
- **Responsáveis:** Não cumprimento do pátrio do poder, desinteresse em relação ao destino dos filhos.
- **Social:** trabalho com incompatibilidade de horário para os estudos, agressão entre alunos, violência em relação a gangues, etc.

Conforme o exposto, pode-se observar que as causas que levam à evasão são diversas, pois cada categoria mencionada possui muitos outros fatores que não foram citados, e existem até outras categorias que não foram aqui mencionadas.

Para lidar com o problema da evasão, é necessário identificar quais fatores podem influenciar nela, e uma vez que os motivos são conhecidos, é importante identificar alunos que se caracterizam nestas situações, correndo risco de cometer evasão. De modo tradicional o professor é o principal responsável por identificar e acionar a rede de combate à evasão (FERREIRA, 2000). A próxima seção abordará algumas das estratégias utilizadas para combater a evasão escolar no Brasil.

2.1.2 Combate à evasão escolar no Brasil

Como já foi mencionado ao fim da seção anterior, para que o combate à evasão seja realizado de maneira eficaz, é necessário identificar em qual ou quais das categorias está presente o problema do aluno considerado em risco de evasão. Ferreira (2000) indica algumas medidas adotadas:

- **Escola:** Quando a evasão dos alunos ocorre principalmente em razão da escola, a solução é responsabilidade principalmente da própria escola, da diretoria de ensino (Estado) e da Secretaria de Educação (no âmbito municipal), tornando, entre outras coisas, o ensino mais atraente ao aluno evadido.
- **Aluno:** Quando o problema está centrado no comportamento do próprio aluno, a intervenção direta deve ocorrer por parte da Família, Escola, Conselho tutelar, Ministério Público e Poder Judiciário, sendo mais ampla a atuação da Família e do Estado. Indiretamente, atuam o Conselho Municipal da Criança e do Adolescente, secretarias de Assistência Social e Saúde, dentro das políticas públicas que visem o regresso do aluno, incluindo programas específicos para a área (ex. reforço escolar, bolsa escola, etc.).

- **Responsáveis:** Se os responsáveis pelo aluno forem a principal razão da evasão, quem deve intervir é a Escola, junto com o conselho tutelar, Ministério público e poder judiciário. Indiretamente são a Secretaria de assistência Social e Saúde.
- **Social:** Quando se trata de questão social, como trabalho, falta de transporte, medo de violência, etc., o problema deve ser solucionado em conjunto pela família, escola, conselho tutelar, ministério público e Poder judiciário. Indiretamente, devem atuar as Secretarias de assistência social.

2.2 APRENDIZADO DE MÁQUINA

O termo Aprendizado de Máquina (ML, do inglês *Machine Learning*) pode ser definido como uma área de pesquisa da Inteligência Artificial cujo objetivo princi- pal encontra-se no desenvolvimento de técnicas sobre o aprendizado, bem como a construção de sistemas computacionais capazes de obter conhecimento (MONARD; BARANAUSKAS, 2003). A ML é muito importante para a inteligência artificial, pois a capacidade de aprender é fundamental para um comportamento inteligente (SANCHES, 2003).

Segundo Russel (2013), é possível dizer que um agente (máquina) aprende se ao executar uma tarefa, observar seu mundo e melhorar seu desempenho. Atualmente, dentre as aplicações de ML, podemos citar sistemas de reconhecimento de imagens e veículos autônomos (ALLENDE-CID, 2019). Inazawa (2019) indica que comumente a ML é utilizada para identificar padrões em dados, possibilitando automatizar tarefas complexas aos humanos ou realizar predições.

Esta área é necessária, segundo Russel, devido a três motivos principais: 1) Os projetistas não podem antecipar todas as situações possíveis em que o agente possa se encontrar, 2) os projetistas não podem antecipar todas as mudanças ao longo do tempo (existem variáveis que sofrem mudanças significativas com o passar dos anos), e 3) Por vezes, os programadores humanos não têm ideia de como programar uma solução por si só (em um programa de reconhecimento facial, por exemplo).

2.2.1 Aprendizado Indutivo

A indução é a forma de inferência lógica que permite obter conclusões gerais a partir de exemplos pontuais (SANCHES, 2003). Nela, nosso raciocínio parte do particular para o universal, sendo feita uma generalização. Sanches (2003) ainda ressalta que hipóteses geradas por uma inferência indutiva podem ser verdadeiras, e podem ser falsas também, devido a motivos como, por exemplo, as falácias (quando as premissas não sustentam a conclusão). Com isso, o que se pode dizer com maior precisão quando se há premissas verdadeiras, é que sua conclusão é provavelmente

verdadeira. No entanto, a probabilidade de uma conclusão ser verdadeira aumenta ao passo que suas premissas demonstram-se confiáveis.

Os autores Monard e Baranauskas (2003) e Sanches (2003) concordam sobre a importância da inferência para a descoberta de conhecimento novo, e ambos ressaltam que devemos utilizá-la com cautela, pois segundo eles, se o número de exemplos não for suficiente, ou se não forem bem selecionados, as hipóteses obtidas a partir dela podem ter valor pequeno ou nulo.

Um tipo especial de aprendizado indutivo que é caracterizado pela tentativa de induzir descrições gerais a partir de exemplos particulares é chamado de Aprendizado Indutivo por Exemplos (SANCHES, 2003). É importante que os exemplos representem de maneira mais fiel possível o conceito que se deseja induzir, pois caso isso não ocorra, a indução não terá grande valor. Algumas vezes acontece de dados com ruído (incorretos) serem utilizados para a indução, e isto terá pouco proveito para exemplos que não foram utilizados na indução.

2.2.2 Aprendizado de máquina indutivo por exemplos

Em ML, o aprendiz que utilizará técnicas de indução por exemplo é denotado por sistema de aprendizado ou algoritmo de aprendizado, podendo ser definido como um sistema computacional que se baseia em experiências bem sucedidas para tomar decisões futuras, isso de acordo com Sanches (2003).

O aprendizado indutivo por exemplos pode ser dividido em Supervisionado e Não-Supervisionado. No próximo tópico encontram-se características destes tipos de aprendizado, e outros tipos também serão abordados.

2.2.2.1 Aprendizado não supervisionado

Aprendizado não Supervisionado se refere ao tipo de aprendizagem em que o sistema aprende padrões na entrada, mesmo com ausência de resposta ou *feedback* explicitamente apontado (RUSSEL, 2013).

Neste tipo de aprendizado, o sistema de aprendizado recebe vetores com exemplos, sem informações de classes, e o sistema deve construir um modelo que procura regularidade nos exemplos de entrada, formando agrupamento ou *clusters* de exemplos com características similares (SANCHES, 2003), (RUSSEL, 2013). Após a determinação dos grupos, é necessário realizar uma análise para identificar o que cada grupo representa no contexto do problema em análise.

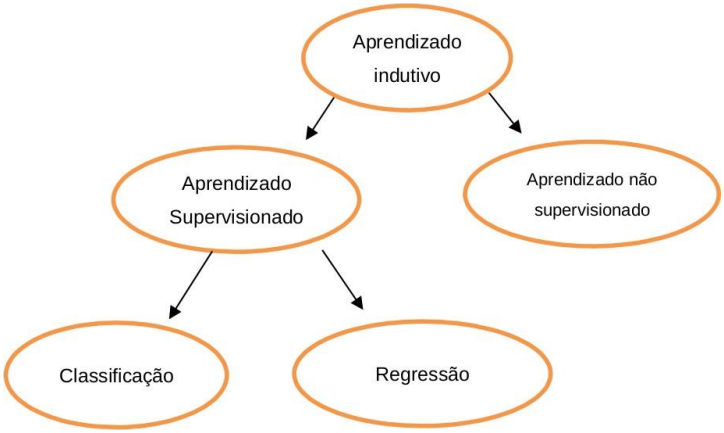
2.2.2.2 Aprendizado supervisionado

No aprendizado supervisionado, o sistema de aprendizado recebe exemplos de entrada e saída com rótulos da classe associada, e o sistema deve aprender uma função que relacione a entrada com a saída (RUSSEL, 2013), (MONARD; BARANAUSKAS, 2003)

Monard e Baranauskas (2003) define como objetivo dos algoritmos neste tipo de aprendizado construir um classificador que determine de maneira correta em qual classe exemplos não rotulados adicionados posteriormente devem ser inseridos.

Quando a classe assume valores discretos (ou qualitativos) a tarefa do aprendizado é denominada o problema é denominado como Classificação, e para valores contínuos (quantitativos, reais) ele é denominado Regressão (MONARD; BARANAUSKAS, 2003), (SANCHES, 2003).

FIGURA 1 – A HIERARQUIA DO APRENDIZADO



FONTE: Sanches (2003) e Monard e Baranauskas (2003)

2.2.2.3 Paradigmas do Aprendizado de Máquina

Existem diferentes paradigmas em ML, dentre eles, as definições de alguns deles encontram-se a seguir, de acordo com Monard e Baranauskas (2003):

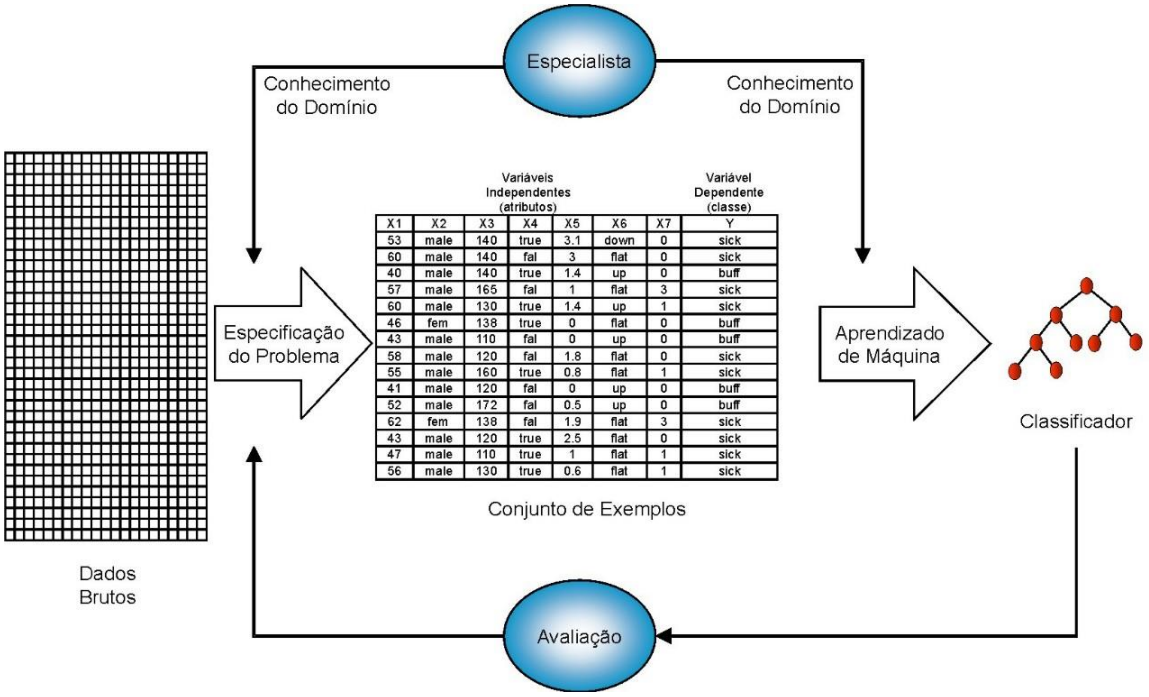
Simbólico: Os sistemas buscam aprender construindo representações simbólicas de um conceito através de análise de exemplos e contraexemplos desse conceito. As representações simbólicas estão tipicamente na forma de alguma expressão lógica, árvore de decisão, regras ou rede semântica.

Estatístico: Pesquisadores estatísticos desenvolveram diversos métodos de classificação, semelhantes aos que posteriormente foram criados pela comunidade de

Aprendizado de Máquina. Pretende-se utilizar o modelo estatístico para se aproximar do conceito induzido. Muitos modelos são paramétricos, assumindo alguma forma de modelo, e então encontrando valores apropriados para os parâmetros de modelo a partir dos dados.

Dentre os métodos estatísticos, destacam-se os de aprendizado Bayesiano, que utilizam um modelo probabilístico baseado no conhecimento prévio do problema, o qual é combinado com os exemplos de treinamento para determinar a probabilidade final de uma hipótese.

FIGURA 2 – O CLASSIFICADOR À DIREITA FORNECE UMA INTERPRETAÇÃO COMPACTA DOS DADOS



X1	X2	X3	X4	X5	X6	X7	Y
53	male	140	true	3.1	down	0	sick
60	male	140	fal	3	flat	0	sick
40	male	140	true	1.4	up	0	buff
57	male	165	fal	1	flat	3	sick
60	male	130	true	1.4	up	1	sick
46	fem	138	true	0	flat	0	buff
43	male	110	fal	0	up	0	buff
58	male	120	fal	1.8	flat	0	sick
55	male	160	true	0.8	flat	1	sick
41	male	120	fal	0	up	0	buff
52	male	172	fal	0.5	up	0	buff
62	fem	138	fal	1.9	flat	3	sick
43	male	120	true	2.5	flat	0	sick
47	male	110	true	1	flat	1	sick
56	male	130	true	0.6	flat	1	sick

FONTE: Monard e Baranauskas (2003)

A FIGURA 2 apresenta um exemplo de como funciona o aprendizado estatístico. De modo geral, o que existe no início do processo são os dados brutos. O primeiro passo deve ser o conhecimento dos dados e do problema a ser solucionado. Em uma próxima etapa, o algoritmo de aprendizado é chamado e este recebe os dados (contendo as variáveis independentes que irão dar suporte para a criação do modelo, e a variável dependente, que é a saída desejada que o modelo identifique) como parâmetro. Após isso é gerada a classificação, e deve ser feita uma avaliação para estabelecer se o modelo obtém um bom desempenho.

Baseado em exemplos: A classificação de um exemplo é realizada através da comparação com outro similar que já possui a classe conhecida, assumindo que o novo exemplo terá a mesma classe. Esse tipo de sistema é denominado *lazy* (preguiçoso). Sistemas *lazy* necessitam manter os exemplos na memória para classificar novo exemplos, em oposição aos sistemas *eager* (gulosos), que utilizam os exemplos para induzir o modelo, descartando-os em seguida. Saber quais exemplos (casos) devem ser memorizados em sistemas *lazy* é muito importante.

Conexionista: Redes Neurais são construções matemáticas simplificadas inspiradas no modelo biológico do sistema nervoso. A representação de uma Rede Neural envolve unidades altamente interconectadas e, por esse motivo, o nome conexionismo e utilizado para descrever a área de estudo.

Genético: Um classificador genético consiste de uma população de elementos de classificação que competem para fazer a predição. Elementos que possuem uma performance fraca são descartados, enquanto os elementos mais fortes proliferam, produzindo variações de si mesmos.

2.3 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS

Alguns autores notaram que o armazenamento de dados é uma das preocupações existentes desde o princípio da utilização de sistemas computacionais Rabelo et al. (2017) e Camilo e Silva (2009). É observado o fato de a diminuição no custo de aquisição de hardware ter influenciado no aumento da capacidade de dados armazenados.

Em Camilo e Silva (2009), alguns exemplos de estruturas de armazenamento novas e mais complexas são citadas: banco de dados, Data *Warehouses*, Bibliotecas Virtuais, *Web* e outras. Camilo e Silva (2009) exemplifica o enorme volume de dados gerados ainda naquela década: Satélites da NASA: cerca de um terabyte de dados por dia; Projeto Genoma: milhares de bytes para cada uma das bilhões de bases genéticas, etc.

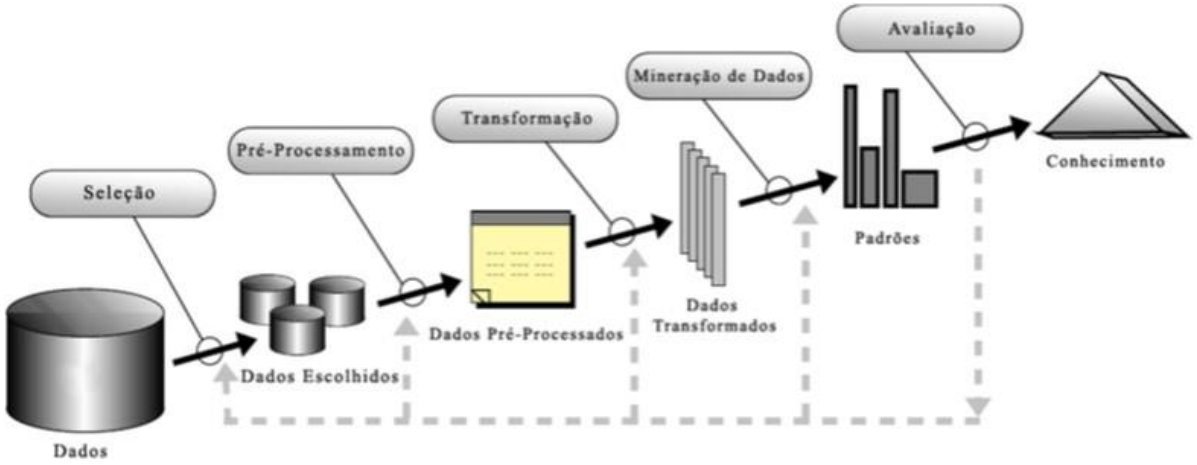
Ao citar Baranauskas (2001), Sanches (2003) discorre sobre a maneira clássica de analisar os dados: analistas são especialistas humanos que conhecem os dados atuando com um sistema computacional. Entretanto, ainda segundo Sanches (2003), na medida que a quantidade de dados aumenta, torna-se inviável efetuar uma análise manual destes dados, sendo a tarefa de induzir conhecimentos novos praticamente improvável.

Para que esse grande volume de dados fossem utilizados na obtenção de conhecimento, surge uma área denominada Descoberta de Conhecimento em Bancos de Dados (KDD, do inglês *Knowledge Discovery in Databases*), que tenta resolver um

problema da chamada Era da Informação: A sobrecarga de dados.

Há diferentes definições para KDD presentes na bibliografia. Em algumas, ela é sinônimo de Mineração de Dados, enquanto em outras, Mineração de dados é uma etapa de KDD (CAMILO; SILVA, 2009). No entanto, ainda de acordo com Camilo e Silva (2009), é consenso que o processo de mineração é interativo, iterativo e dividido em fases, como se pode observar na figura 3.

FIGURA 3 – PROCESSO DE KDD



FONTE: Camilo e Silva (2009)

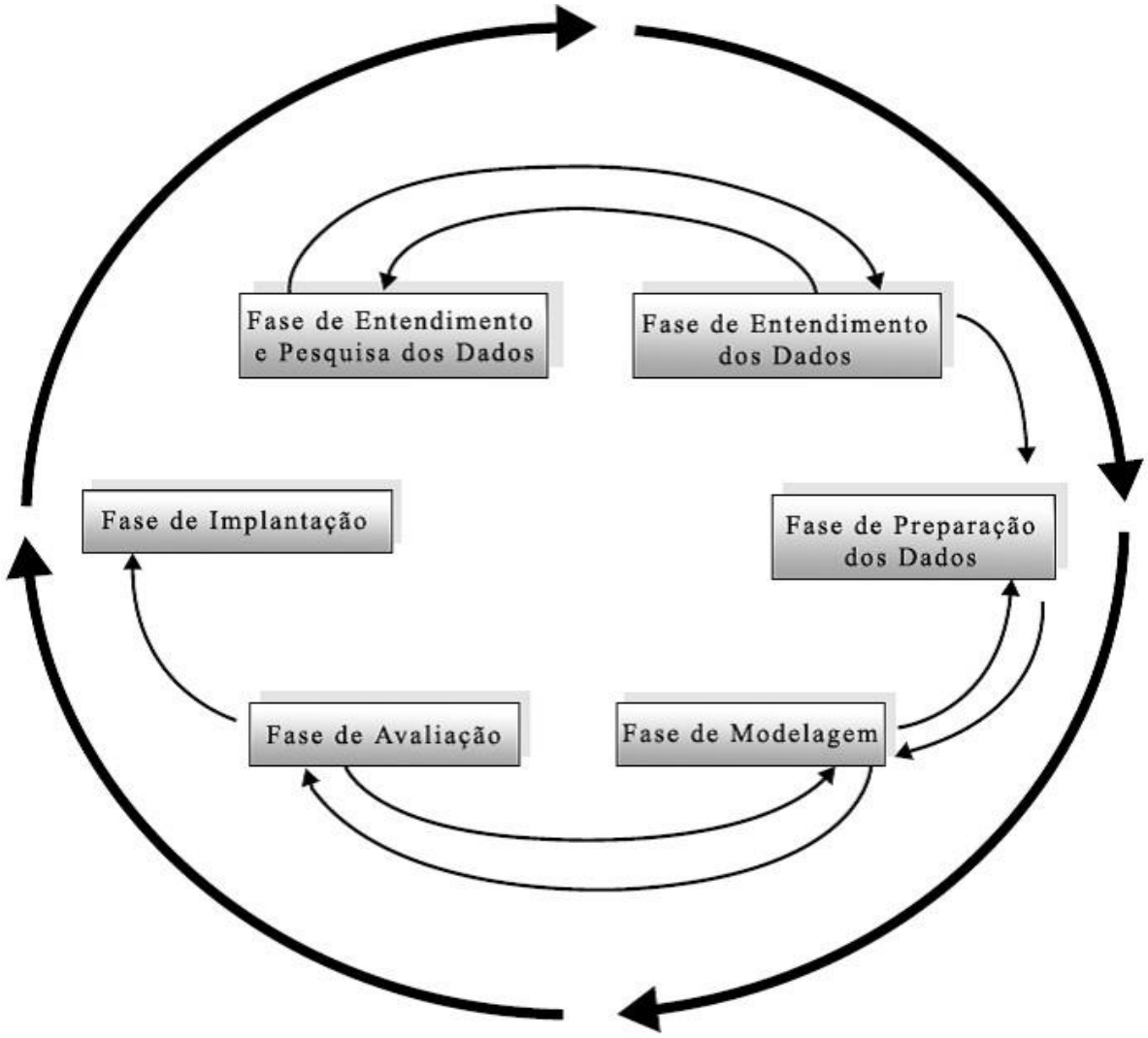
Existem processos para definir e padronizar cada fase e atividade da Mineração de Dados. Utilizaremos o *Cross-Industry Standard Processo of Data Mining* (CRISP-DM), pois este é bastante utilizado na bibliografia, e é considerado um padrão com boa aceitação (CAMILO; SILVA, 2009).

2.3.1 CRISP-DM

CRISP-DM é um modelo que apresenta boas práticas para se extrair resulta- dos melhores no processo de mineração, este modelo é documentado e disponível livremente (CALIXTO; SEGUNDO; GUSMÃO, 2017).

O processo CRISP-DM é constituído de seis fases organizados de maneira cíclica, podendo ir e voltar durante as fases (CAMILO; SILVA, 2009). A Figura 4 mostra as fases do processo.

FIGURA 4 – O PROCESSO CRISP-DM



FONTE: Camilo e Silva (2009)

De acordo com Camilo e Silva (2009) as fases do processo CRISP-DM são:

1) Entendimento dos negócios: Deve-se entender qual objetivo se deseja atingir com a mineração de dados. Esta fase irá ajudar nas seguintes.

2) Entendimento dos dados: As fontes fornecedoras dos dados podem vir de diversos locais e possuírem formatos diferentes. É necessário conhecer os dados visando:

- Descrever de forma mais clara o problema;
- Identificar os dados relevantes para o problema em questão;
- Certificar-se de que as variáveis relevantes para o projeto não são independentes.

3) Preparação dos dados: Dependendo da qualidade dos dados, algumas ações podem ser necessárias antes de aplicar o processo de mineração. Geralmente este processo envolve filtrar, combinar e preencher valores vazios.

4) Modelagem: É nesta fase que se aplicam as técnicas de mineração. Para escolher as técnicas, deve-se considerar os objetivos desejados.

5) Avaliação: Nesta etapa são utilizadas ferramentas gráficas para visualização e análise dos resultados, além da execução de testes e validações com intuito de obter a confiabilidade nos modelos, e a obtenção de indicadores para auxiliar na análise dos resultados.

6) Distribuição: Mostrar os resultados aos envolvidos.

2.4 MINERAÇÃO DE DADOS

A MD, trata-se da disciplina que tem como principal objetivo descobrir novas informações presentes em grandes quantidades de dados (BAKER; ISOTANI; CARVALHO, 2011), extrair novos conhecimentos implícitos em bases de dados, para prever ocorrências de padrões úteis para determinadas atividades Rigo et al. (2014).

No contexto de MD, “novas informações” ou “conhecimentos implícitos” tendem a se referir a relações existentes entre variáveis do conjunto de dados (BAKER; ISOTANI; CARVALHO, 2011).

Existem muitas ferramentas que facilitam a execução dos algoritmos de Mineração, porém, a análise humana ainda é necessária, mas ainda assim, muito trabalho manual é evitado, pois os especialistas só precisam dedicar seu tempo com partes mais significativas dos dados (CAMILO; SILVA, 2009).

2.4.1 Tarefas

Costuma-se dividir mineração de dados em algumas tarefas, dentre as quais as mais comuns encontram-se a seguir, de acordo com (CAMILO; SILVA, 2009):

- **Descrição:** Tarefa utilizada para descrever os padrões e tendências revelados pelos dados. Geralmente a descrição oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.
- **Classificação:** Esta tarefa visa identificar a qual classe um determinado registro pertence. Nela, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo indicação para a classe à qual pertence, com intuito de

‘aprender’ como classificar um novo registro, utilizando valores categóricos. Utiliza-se aprendizado supervisionado.

- **Estimação (ou Regressão):** A tarefa da classificação é similar ao que é realizado pela classificação, divergindo no tipo de dados utilizado por cada uma. A classificação é utilizada quando o registro possui valores numéricos, não categóricos.
- **Predição:** A predição é parecida com a classificação e estimação, mas ela visa descobrir um valor futuro de determinado atributo. Podem ser utilizados alguns métodos de classificação e regressão para esta tarefa, com as devidas considerações.
- **Agrupamento:** O agrupamento visa identificar e aproximar registros similares. Um agrupamento é uma coleção de registros similares entre si, e diferentes de outros registros nos demais agrupamentos. Utiliza aprendizado não supervisionado, portanto, não necessita que os registros sejam identificados previamente. Ela não pretende classificar, estimar ou predizer algum valor, mas somente identificar os grupos similares.
- **Associação:** Consiste em identificar quais atributos estão relacionados. Apresentam a forma: Se atributo X ENTÃO atributo Y.

2.5 TRABALHOS RELACIONADOS

Nesta seção serão apresentados trabalhos científicos envolvendo mineração de dados educativos, nos quais seus autores compartilhavam de objetivos próximos aos deste projeto, servindo como suporte no desenvolvimento do mesmo.

Rabelo et al. (2017) desenvolveram um trabalho que objetivou descrever a aplicação de técnicas de Mineração de Dados Educacionais por meio de árvores de decisão, através de um ambiente virtual de aprendizagem educacional, a plataforma Moodle, durante a realização de cursos de graduação à distância pela UFRN. Os autores também pretendiam predizer se o aluno teria sucesso ou insucesso acadêmico no decorrer do curso, avaliando principalmente seu nível de participação, interação ou desempenho na plataforma, sendo a hipótese da pesquisa a seguinte: o desempenho do aluno no ambiente virtual, influencia em sua aprovação na disciplina. A mineração de dados foi realizada com o auxílio da Ferramenta Weka, empregando a técnica de classificação de árvores de decisão.

O resultado foi de uma precisão de acertos de 93,97% com a utilização do classificador ID3, e de 96% utilizando o J48, com o mesmo modelo preditivo e com os mesmos indicadores. A hipótese foi validada analisando uma turma que possuía 268 alunos, obtendo um resultado do teste de Qui-Quadrado que permitiu afirmar que houve evidências de que o desempenho ou participação efetiva do aluno no ambiente virtual influência no resultado final ou aprovação do aluno.

Bezerra et al. (2016) realizaram uma pesquisa com alunos propensos a evasão no 9º (nono) ano do ensino fundamental, com o objetivo de trazer informações que contribuíssem para o desenvolvimento de políticas públicas visando reduzir a evasão escolar nesse momento de transição da formação escolar. A base de dados utilizada foi a do Censo escolar de 2012, divulgada pelo INEP. A preparação de dados foi efetuada utilizando linguagem SQL, no SQLServer. Neste trabalho foram utilizadas as técnica "árvore de decisão", devido a sua alta legibilidade; indução de regras, por ser uma das mais importantes em mineração de dados; e Regressão Logística.

Os autores concluíram com a análise do resultado obtido pela indução de regras e árvore de decisão que os atributos mais determinantes para a evasão foram o turno das aulas e a idade dos alunos. Observaram também utilizando regressão logística que os fatores mais fortes com relação aos docentes são a média de escolas em que eles trabalham e o numero de docentes pós-graduados.

Calixto, Segundo e Gusmão (2017) tiveram como objetivo a construção de um estudo analítico a fim de identificar as variáveis concernentes à evasão escolar e comparar a relação entre os estados analisados entre os anos de 2014, 2015 e 2016 dos estados de Ceará e Sergipe. Os dados utilizados foram do censo escolar distribuído pelo INEP, assim como no trabalho acima mencionado, mas estes utilizaram dos anos de 2014, 2015 e 2016, escolhendo os estados do Sergipe e Ceará por serem os estados com maior e menor taxa no ranking do índice do Desenvolvimento da Educação Básica para os estados da região Nordeste.

As técnicas utilizadas foram a indução de regra, regressão logística, obtendo nesta última acurácia de classificação de 87.4 no Ceará e 86.8 em Sergipe. Em comparação entre estados, as etapas do ensino médio influenciam de maneira positiva na evasão, isso quer dizer que elas são variáveis que há maior probabilidade de evasão nelas. Outra variável que influenciou positivamente nos dois estados foi a idade, isso significa que a cada ano a probabilidade de evasão aumenta para cada aluno. No estado do Ceará, a raça negra foi a única que influenciou positivamente em todos os anos, enquanto no Sergipe isso não ocorreu. No estado do Ceará, no par 2014-2015, a escola possuir laboratório de informática influenciou de maneira positiva, mas no ano seguinte o valor apresentou-se invertido. No Sergipe, esta variável apresentou-se como influência negativa em ambas as comparações.

3 METODOLOGIA

Esta pesquisa tem uma abordagem quantitativa, pois seu desenvolvimento e resultado são completamente baseados em dados dados mensuráveis estatisticamente. A seguir estão descritas as ferramentas, base de dados utilizada e as etapas da pesquisa.

3.1 REVISÃO BIBLIOGRÁFICA

Nesta etapa foi realizada uma revisão bibliográfica, com o objetivo de obter um embasamento teórico para a fundamentação e elaboração deste trabalho. Foram realizadas pesquisas em sites, artigos científicos, monografias, teses, livros entre outros modelos de pesquisas que abordam tema como: Evasão escolar, Mineração de dados educacionais e Aprendizado de máquina.

3.2 FERRAMENTAS E BASE DE DADOS

Nesta seção são descritas as ferramentas que serão utilizadas nesta pesquisa.

- **Base do Censo Escolar**[1]**:** O Censo Escolar é um levantamento de dados estatístico-educacionais de âmbito nacional realizado todos os anos e coordenado pelo Inep. Trata-se do principal instrumento de coleta de informações da educação básica, ensino regular (educação Infantil e ensinos fundamental e médio), educação especial e Educação de Jovens e Adultos (EJA).
- **Jupyter Notebook:** Ambiente Computacional web que permite entender e visualizar dados e resultados de análises. Facilita a expereimentação, colaboração e publicação online. Os documentos Jupyter Notebooks, que por padrão tem formato ".ipynb"podem ser convertidos em outros formatos como HTML, slides, Latex, PDF, Python, etc.
- **Pandas:** Pandas é uma biblioteca Python, que fornece ferramentas de análise e visualização de dados e estruturas de dados de alta performance, com uma fácil utilização.
- **Numpy:** Biblioteca Python para realização de cálculos em arrays multidimensionais com alta eficiência.
- **Matplotlib:** Biblioteca Python que permite criação de gráficos em 2D para vizualização de dados. Dentre os gráficos encontram-se o de barras, linha, pizza, histogramas e muitos outros. Esta biblioteca permite personalizar os gráficos de maneira simples.

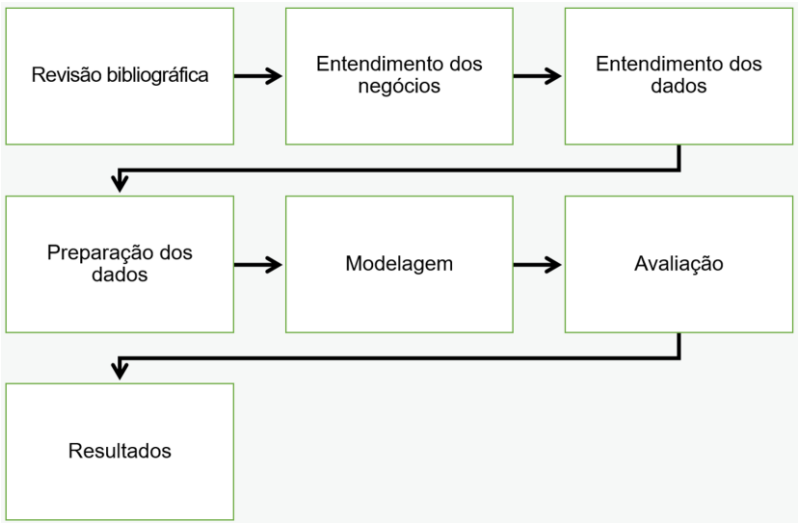
[1] http://dados.gov.br/dataset/microdados-do-censo-escolar

- **Seaborn:** O Seaborn é outra biblioteca de visualização de dados Python, sendo esta baseada no matplotlib. Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos, os quais precisariam ser criados de maneira um pouco complexa utilizando apenas o Matplotlib.
- **Scikit-Learning:** Ferramenta de machine learning para Python. Ela inclui vários algoritmos de classificação, regressão e agrupamento incluindo máquinas de vetores de suporte, florestas aleatórias, entre outros.

3.3 ETAPAS DA PESQUISA

De acordo com o que abordado na seção 2.3.1, a metodologia desta pesquisa será baseada no processo CRISP-DM. Pode-se observar as etapas na a figura 5:

FIGURA 5 – METODOLOGIA



FONTE: Autor

- **Entendimento dos negócios:** Análise bibliográfica sobre Evasão em trabalhos específicos e em trabalhos que utilizam MDE para predizer risco de evasão escolar. Devido ao acesso livre e a grande quantidade de dados disponíveis e a grande uso desta base encontrados na bibliografia, escolheu-se utilizar dados oriundos do Censo Escolar divulgados pelo INEP em seu portal. Essa divulgação é realizada anualmente, desde 1995, contendo dados do ensino infantil ao superior de todos os estados do país. Nestes dados existem uma quantidade grande de informações acerca de alunos, docentes, matrículas e turmas de todos os estados brasileiros.

- **Entendimento dos dados:** Nesta etapa foi realizada uma análise no dicionário de dados do Censo divulgados pelo próprio INEP. Observou-se a organização dos dados, os possíveis valores de cada campo, as mudanças que ocorrem em nomes de variáveis em diferentes anos, entre outras características. Para uma melhor compreensão foi realizada também uma análise exploratória. Para isto, utilizou-se o ambiente Jupyter Notebook. Os dados brutos continham um grande volume de informações que não pertenciam ao objeto de interesse da pesquisa, como escolas, docentes e alunos de outros estados e de fora do ensino fundamental, escolas paralisadas. A princípio desejava-se utilizar os dados mais recentes fornecidos pelo INEP, mas isto não foi possível devido a uma alteração no código de identificação única dos alunos a partir de 2018, o que impossibilitou a verificação mencionada anteriormente. Com isto, foram selecionados os anos de 2014 e 2015.

  Os dados do Censo Escolar são disponibilizados em quatro Arquivos formato Arquivos Separados por Vírgulas (CSV, do inglês *Comma Separated Values*), a saber: Matrículas, Docentes, Escolas e Turmas. O acesso a dados de desempenho de alunos somente é autorizado para as instituições em que os mesmos estudam, o que inviabilizou a utilização de tais informações nesta pesquisa. Junto aos arquivos é fornecido um dicionário de dados com detalhes sobre cada variável presente. Neste dicionário é apresentada uma breve descrição de cada variável, além disso é informado quando uma delas passa por alteração na nomenclatura, ou quando é removida do levantamento.

FIGURA 6 – DADOS BRUTOS: MATRÍCULAS DE 2014

```
matriculas14.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6000886 entries, 0 to 6000885
Data columns (total 85 columns):
```

FONTE: Autor

A FIGURA 6 mostra informações de dados brutos de de um arquivo de 2014. Existem apenas neste arquivo de Matrículas 85 colunas com informações a respeito dos alunos, somando um total de 6.000.886 linhas com dados sobre alunos da região norte.

- **Preparação dos dados:** Nesta seção serão descritas as etapas realizadas na fase de preparação dos dados. Esta fase é muito importante, pois é ela quem

permite a criação de um bom modelo. Nela, os ruídos presentes nos dados são tratados, as variáveis de interesse são selecionadas e ao final dela os dados estão prontos para a utilização no modelo. A fase de preparação foi dividida em: Filtro de dados de interesse – Dados relacionados a alunos do ensino fundamental do estado do amazonas, tratamento de valores nulos ou irregulares. Essas etapas foram realizadas com auxílio das bibliotecas Pandas e Seaborn, executadas no ambiente Jupyter Notebook.

- **Filtro de dados de interesse:** Os arquivos em estado bruto foram importados para o ambiente utilizado. Métodos de Python, observamos que o arquivo da região norte selecionado estava preenchido com dados de 7 estados (cada um identificado por um código único do tipo inteiro: 11, 12, 13 (amazonas), 14, 15, 16 e 17). Com isso, foi aplicado o primeiro filtro para selecionar os dados do amazonas em cada um dos arquivos utilizados.

FIGURA 7 – ESTADOS NO ARQUIVO DE DADOS BRUTO

FONTE: Autor

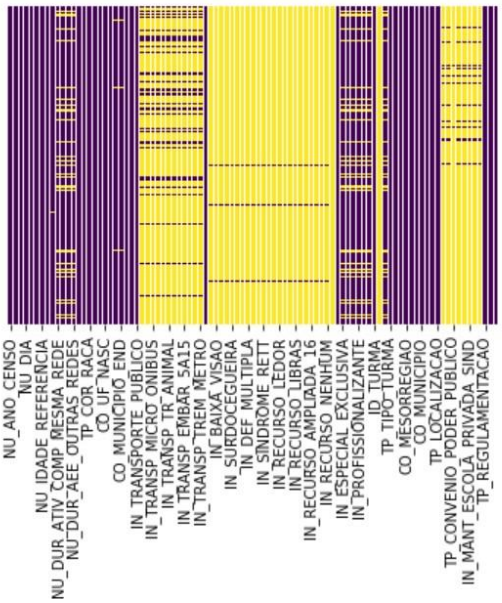
A etapa seguinte do tratamento foi a verificação de alunos do ensino fundamental. Os dados brutos possuíam informações sobre todas as turmas do ensino infantil ao ensino médio. Selecionamos apenas as turmas de ensino fundamental, excluindo as linhas correspondentes a turmas de ensino fundamental ou infantil.

- **Tratamento de valores nulos:** Após os filtros citados, começamos a analisar os dados levando em consideração a quantidade de valores nulos que possuíam. Isto foi feito com auxílio da biblioteca Seaborn, que retorna um mapa de calor. Na imagem pode-se observar a utilização desta função.

A FIGURA 8 mostra um conjunto de colunas dos dados, cada coluna presente corresponde a um campo do arquivo. A cor amarela representa os valores nulos. Pode-se observar que em algumas colunas a quantidade de valores nulos é disparadamente superior aos valores não nulos. Para estas colunas não é recomendável realizar nenhum tipo de tratamento para recuperação, pois isso poderia alterar o resultado do modelo criado. Com isso, optou-se por remover

FIGURA 8 – DADOS NULOS: MAPA DE CALOR

```
sns.heatmap(matricula15.isnull(),yticklabels=False,cbar=False,cmap='viridis')
<matplotlib.axes._subplots.AxesSubplot at 0xb214cc0>
```



FONTE: Autor

estas colunas do arquivo de dados. Também existiam colunas em que a quantidade de linhas com dados ausentes era relativamente baixa, e com algumas delas calculamos a média e substituímos os valores ausente pela média, para outras utilizamos a moda, ou simplesmente excluímos as linhas com valores nulos (quando a coluna possuía uma quantidade demasiadamente baixa de valores nulos).

- **Criação da coluna EVASÃO:** Por se tratar de predição, que utiliza de aprendizado supervisionado, a criação do modelo necessita que exista uma variável com as classes de saída, e neste caso, essa classe é a responsável por informar se o aluno teve ou não evasão. Para a criação desta variável, será realizada análise par a par entre os anos, e será considerado evasor, o aluno matriculado em alguma escola do estado do Amazonas no ano x, e não possui matrícula no ano x+1. Este processo foi realizado com o trecho de código da imagem abaixo.

FIGURA 9 – VERIFICANDO SE O ALUNO EVADIU

```
evasao = []
for i in juncao16['CO_PESSOA_FISICA'].values:
    evasao.append(1)
    if i in matricula17['CO_PESSOA_FISICA'].values:
        evasao[len(evasao)-1] = 0
        continue
```

FONTE: Autor

O código cria um vetor vazio entra em um laço de repetição que percorre cada elemento de matricula de um arquivo com dados que já foram selecionados alunos do ensino fundamental do estado do amazonas. A cada elemento que é percorrido, é adicionado um valor novo no vetor. Como padrão é adicionado o valor 0, assumindo que o aluno não está matriculado no ano seguinte. Após essa atribuição, é realizada uma busca pelo código de identificação única do aluno no arquivo de dados puro (sem filtros, ou seja, a busca foi realizada verificando se no ano seguinte o aluno estava matriculado nos 7 estados da região norte presentes no arquivo). Após este processo, o vetor foi integrado ao DataFrame como uma nova coluna denominada EVASAO.

- **Divisão de arquivos em treino e teste:** A última etapa da preparação dos dados foi a divisão dos arquivos em treino e teste. O arquivo foi dividido da seguinte maneira: 70% para treino e 30%de para teste, acordo com o recomendado na bibliografia. Os dados de treino são os que fornecem a fonte para que o modelo seja gerado. É com base nestes dados que ele criará a função que relaciona os dados de entrada com a saída. A imagem abaixo mostra como esta etapa foi implementada.

FIGURA 10 – DIVISÃO DE DADOS DE TREINO E TESTE

```
from sklearn.model_selection import train_test_split

X = juncao14.drop('EVASAO',axis=1)
y = juncao14['EVASAO']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30)
```

FONTE: Autor

Na primeira linha da imagem estamos importando o método de divisão da biblioteca ScikitLearning. Nas linhas do meio, definimos como valores de X todas as colunas do DataFrame que exceto a coluna EVASAO, e definimos como y apenas a coluna EVASAO Por fim, na última linha criamos 4 variáveis: 2 que receberão os valores de treino e teste de X e duas para os valores de y. Após esta divisão, nosso modelo já poderia ser treinado.

- **Modelagem:** A modelagem é a fase em que se aplicam os métodos de aprendizado de máquina para selecionar o melhor modelo de predição, com intuito de identificar se, com base nas circunstancias do aluno, ele vai evadir ou não. Foi utilizada a biblioteca ScikitLearning da linguagem Python. Os métodos utilizados para a criação do modelo foram os de regressão logística, árvore de decisão e

florestas aleatórias. Uma breve descrição do que é realizado por cada algoritmo será apresentada a seguir.

Regressão Logística

Algoritmo de classificação utilizado para estimar valores discretos (categóricos) baseado em um grupo de variáveis independentes. Isto é, o algoritmo prevê a probabilidade da ocorrência de um evento, ajustando os dados a uma função logística. Como ele prevê a probabilidade, seus valores de saída são algo esperado entre 0 e 1. Igualmente como na regressão linear é necessário aplicar pesos onde ajustam-se aos dados de treinamento do algoritmo, porém a regressão logística não procura a melhor reta que se ajuste aos dados, mas sim a melhor curva. A regressão logística calcula uma razão de probabilidade da variável alvo, que posteriormente é convertida em uma variável de base logarítmica, permitindo assim a classificação com base na aproximação de um dos valores (WITTEN; FRANK, 2005)

Árvores de Decisão

Árvores de decisão é um tipo de algoritmo de aprendizagem supervisionada, bastante utilizado em problemas de classificação. Seu funcionamento abrange entradas contínuas e categóricas, de entrada e de saída. Estas últimas são as classes. De acordo com (Gama, 2004) nestes modelos, um problema complexo é decomposto em problemas menores e recursivamente esta técnica é reaplicada a cada sub problema. Este algoritmo está entre os mais populares de inferência. A capacidade de discriminação de uma árvore vem da divisão do espaço definido pelos atributos em sub espaços e a cada sub espaço é associada uma classe.

A construção de uma árvore de decisão pode ser descrita da seguinte maneira: deve-se selecionar em primeiro lugar o atributo que será o nó-raiz e fazer um ramo para cada possível valor de entrada. Isto divide o problema em subconjuntos, um para cada valor do atributo. A partir disto, o processo pode ser repetido recursivamente para cada ramo. Se ocorrer de os exemplos em um nó possuírem a mesma classificação, o desenvolvimento da árvore é interrompido neste trecho. A escolha do atributo é baseada em qual irá gerar uma árvore menor e que tenha mais chances de classificar melhor.

Florestas aleatórias

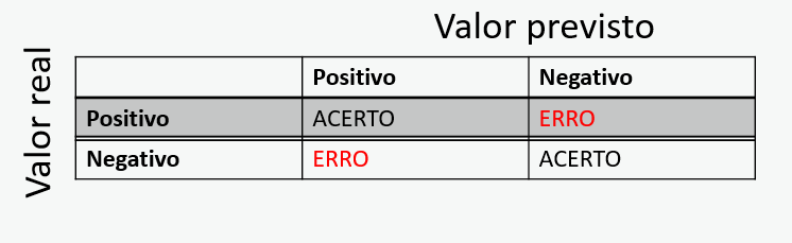
Florestas aleatórias é um algoritmo de aprendizagem de máquina flexível e de fácil utilização que produz um resultado significativamente bom, mesmo sem ajuste de hiperparâmetros. É amplamente utilizado por vários motivos, dentre os quais citamos sua simplicidade a capacidade de utilizá-lo em tarefas de classificação e também de regressão. Ele também aplica métodos de redução dimensional, trata valores faltantes, valores anómalos (‘outliers’) e outras etapas essenciais da

exploração de dados. No geral, ele aoresenta um desempenho muito bom. É um tipo de método de aprendizado de ‘ensemble’, onde um grupo de modelos fracos são combinados para formar um modelo mais forte.

Na floresta aleatória, crescemos múltiplas árvores ao invés de uma única árvore no modelo do CART. Para classificar um novo objeto baseado em atributos, cada árvore dá uma classificação, que é como se a árvore desse “votos” para essa classe. A floresta escolhe a classificação que tiver mais votos (de todas as árvores da floresta) e, em caso de regressão, considera a média das saídas por árvores diferentes.

- **Avaliação:** Foi avaliado qual modelo obteve melhor acurácia. Utilizou-se para auxiliar neste processo a matiz de confusão, que informa os tipos de erros e acertos cometidos pelo modelo. Na FIGURA 11 é possível verificar quais os tipos de acertos e erros a matriz classifica.

FIGURA 11 – MATRIZ DE CONFUSÃO: FUNCIONAMENTO DA MATRIZ



FONTE: Autor

A matriz é classificada em 4 partes: Positivo verdadeiro (TP) , Falso positivo (FP), Falso negativo (FN) e Negativo verdadeiro (VN).

Positivo verdadeiro: ocorre quando o modelo informa que o aluno cometerá evasão e isso acontece.

Falso positivo: Quando o modelo diz que o aluno vai evadir e isso não acontece.

Falso negativo: Quando o modelo aponta que o aluno não tem risco de evasão e a evasão acontece.

Negativo verdadeiro: Quando o modelo informa que o aluno não tem risco de evasão e o aluno realmente não encontra-se matriculado no ano seguinte.

4 RESULTADOS

Os resultados obtidos são oriundos da aplicação de algoritmos pertencentes à biblioteca Scikit-Learn. Foi possível realizar diversos testes com a ferramenta, explorando e validando os dados, nesta essência a biblioteca foi utilizada como uma ferramenta exploratória, na qual os atributos selecionados foram utilizados.

VARIÁVEIS RELACIONADAS COM A EVASÃO

Com a criação da coluna Evasão e sua inclusão no arquivo de dados, foi possível utilizar métodos de correlação da linguagem Python (classificam de 1 a -1, com influência positiva ou negativa sobre a EVASÂO, sendo os valores mais próximos de 0 os que não possuem correlação alguma), para verificar quais dentre todas as colunas possuíam um valor significativo para continua sendo utilizados. Foram mantidas apenas variáveis com o valor de correlação superior a 0.04 em módulo (sem considerar o sinal). Essa seleção resultou em 26 variáveis no ano de 2014 e 20 variáveis no ano de 2015, as quais podem ser observadas na FIGURA 12.

FIGURA 12 – VARIÁVEIS SELECIONADAS

Ano de 2014

```
Data columns (total 26 columns)
NUM_IDADE
ID_ZONA_RESIDENCIAL
ID_N_T_E_P
FK_COD_ETAPA_ENSINO_x
ID_ETAPA_AGREGADA_MAT
ID_LOCALIZACAO_ESC
EVASAO
ID_LOCALIZACAO_x
ID_AGUA_FONTE_RIO
ID_ENERGIA_REDE_PUBLICA
ID_LIXO_COLETA_PERIODICA
ID_LIXO_QUEIMA
ID_SALA_PROFESSOR
ID_QUADRA_ESPORTES_DESCOBERTA
ID_DEPENDENCIAS_PNE
ID_PATIO_COBERTO
NUM_EQUIP_MULTIMIDIA
NUM_EQUIP_FOTO
NUM_COMPUTADORES
NUM_FUNCIONARIOS
ID_ALIMENTACAO
ID_REG_MEDIO_MEDIO
NUM_MATRICULAS
ID_LINGUA_LITERAT_INGLES
ID_ENSINO_RELIGIOSO
ID_MANT_ESCOLA_PRIVADA_S_FINS
```

Ano de 2015

```
Data columns (total 20 columns):
NU_IDADE
TP_ZONA_RESIDENCIAL
IN_TRANSPORTE_PUBLICO
TP_LOCALIZACAO_x
TP_LOCALIZACAO_DIFERENCIADA_x
IN_EDUCACAO_INDIGENA_x
EVASAO
IN_AGUA_FONTE_RIO
IN_ENERGIA_REDE_PUBLICA
IN_LIXO_COLETA_PERIODICA
IN_LIXO_QUEIMA
IN_EQUIP_MULTIMIDIA
IN_COMPUTADOR
IN_INTERNET
TP_ATIVIDADE_COMPLEMENTAR
NU_MATRICULAS
IN_DISC_EDUCACAO_FISICA
TP_LOCALIZACAO
TP_LOCALIZACAO_DIFERENCIADA
```

FONTE: Autor

Ressaltamos que correlação não implica causalidade. Isto ocorre devido a fatores interligados as variáveis presentes no censo. Por exemplo, no ano de 2014 o campo “Quadra de esportes coberta” e “Quadra de esportes descoberta” aparecem como correlação com a evasão. Mas isso não quer dizer com uma certa confiabilidade que a ausência ou presença de uma quadra coberta influência na evasão. Pode ser que a quadra esteja relacionada com outros fatores, como o local em que a escola se encontra, os investimentos da administração responsável neste lugar entre muitos outros possíveis fatores. O fato de outras variáveis envolvendo o tipo de localização na lista sustenta parte do argumento acima.

CRIAÇÃO E AVALIAÇÃO DOS MODELOS

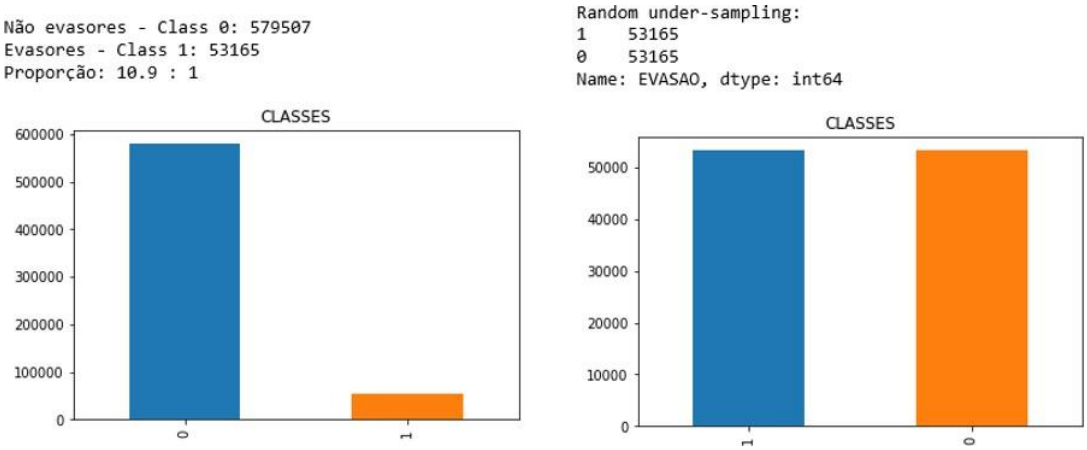
O conjunto de dados de entrada foi dividido em conjunto de treinamento e conjunto de testes, esse processo foi realizado utilizando a função traintestsplit da biblioteca ScikitLearn. Assim os algoritmos foram submetidos ao conjunto de treinamento, gerando um modelo preditivo.

Na primeira fase de modelagem, observou-se que a base de dados possuía a característica de desbalanceamento, o que gerou um vício do modelo de cada algoritmo em classificar apenas a classe majoritária (com mais elementos). Este problema, denominado classes desbalanceadas, surge principalmente porque os algoritmos tradicionais assumem diferentes erros como igualmente importantes, supondo que as distribuições são relativamente equilibradas (Vapnik 1995; Bishop 2006; He and Garcia 2009;Monard e Batista 2002). Essa premissa de assumir custos iguais é fiel ao modelo probabilístico adotado, mas ocasiona problemas em um cenário que apresenta desbalanceamento, como o favorecimento da classe majoritária em suas regras de decisão.

Para solucionar o problema, foi adotado um método de amostragem: suba-mostragem (undersapling), na qual se removem amostras da classe majoritária e as classes se tornam parecidas (Elrahman Abraham 2013). Este método foi selecionado devido a uma exigência menor de processamento para a criação dos modelos após sua implementação. As FIGURAS mostram a quantidade de dados em cada classe antes e depois do tratamento, onde pode-se observar que a quantidade de elementos foi reduzida de maneira significativa, mas como resultado foi adquirido uma semelhança na quantidade de atributos em cada classe, solucionando o problema mencionado.

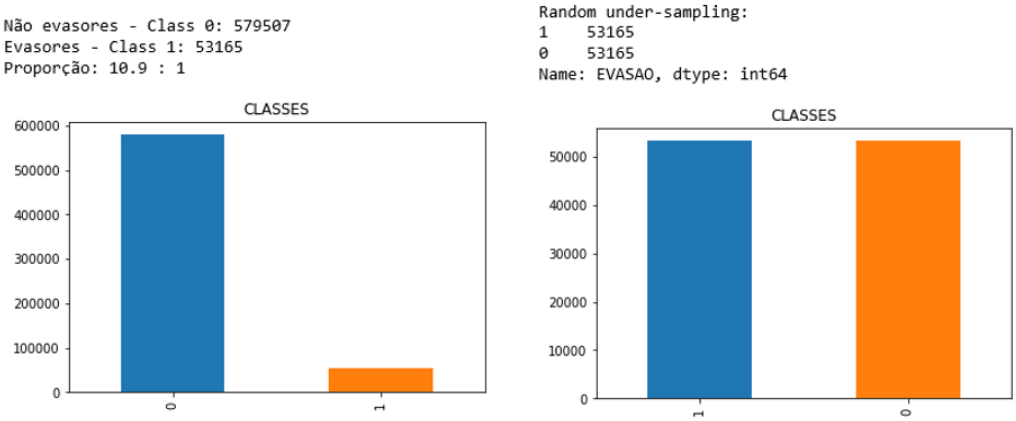
Para a avaliação de desempenho dos algoritmos foram geradas as matrizes de confusão e calculadas a acurácia de cada modelo (FIGURAS 13 E 14). Comparando

FIGURA 13 – SUBAMOSTRAGEM - 2014



FONTE: Autor

FIGURA 14 – SUBAMOSTRAGEM - 2015



FONTE: Autor

os diferentes modelos nestas métricas para concluir, aquele de melhor desempenho para a proposta do trabalho.

Nos modelos gerados com dados de 2014, o que obteve melhor acurácia foi criado com o algoritmo Florestas aleatórias, seguido pelo de Árvores de decisão e por último o de Regressão Logística. O desempenho do algoritmo de florestas

FIGURA 15 – RESULTADOS - 2014

REGRESSÃO LOGÍSTICA

	CLASSE PREDITA	
CLASSE ATUAL	1	0
1	2862	1470
0	1415	2768

CLASSE	PREDIÇÃO
0	58%
1	64%
ACURÁCIA:	61%

ÁRVORES DE DECISÃO

	CLASSE PREDITA	
CLASSE ATUAL	1	0
1	2869	1463
0	1679	2504

CLASSE	PREDIÇÃO
0	63%
1	63%
ACURÁCIA:	63%

FLORESTAS ALEATÓRIAS

	CLASSE PREDITA	
CLASSE ATUAL	1	0
1	98	42
0	70	74

CLASSE	PREDIÇÃO
0	63%
1	65%
ACURÁCIA:	66%

FONTE: Autor

FIGURA 16 – RESULTADOS - 2015

REGRESSÃO LOGÍSTICA

	CLASSE PREDITA	
CLASSE ATUAL	1	0
1	344	174
0	277	269

CLASSE	PREDIÇÃO
0	55%
1	61%
ACURÁCIA:	58%

ÁRVORES DE DECISÃO

	CLASSE PREDITA	
CLASSE ATUAL	1	0
1	11585	4436
0	8929	6949

CLASSE	PREDIÇÃO
0	56%
1	61%
ACURÁCIA:	59%

FLORESTAS ALEATÓRIAS

	CLASSE PREDITA	
CLASSE ATUAL	1	0
1	385	133
0	288	258

CLASSE	PREDIÇÃO
0	57%
1	66%
ACURÁCIA:	62%

FONTE: Autor

aleatórias se enquadra no que era esperado após a revisão bibliográfica, pois o mesmo utiliza combinações de algoritmos mais simples para uma melhor solução. Observa-se por meio das matrizes de confusão que a maioria de classificações apontadas pelos modelos, nos 3 algoritmos utilizados, são do tipo TP, seguidos pelo TN (ambas classificações corretas).

Igual ao que ocorreu nos com dados de 2014, os modelos construídos utilizando dados do ano de 2015 o que obteve melhor acurácia foi o que foi criado com o algoritmo de Florestas Aleatórias, seguido por Árvores de Decisão e Regressão Logística. A mesma tendência de tipos de classificação que mais ocorreu é parcialmente observada: TP foi o que mais foi apontado pelos 3 modelos criados, já o segundo tipo de classificação apontado, se repetiu TN apenas no algoritmo Árvore de decisão, e foi alterado para FN nos dois outros algoritmos.

A acurácia dos modelos criados, girando em torno de 63% no ano de 2014 e de 60% no ano de 2015 apresenta-se como positiva, pois apresenta uma taxa de acerto acima de 50%, mas ainda pode ser mais bastante melhoradas com estudos que explorem outras características dos dados, como o desbalanceamento, por exemplo.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa trouxe como resultado 3 modelos de predição de evasão para alunos do ensino fundamental do estado do Amazonas. Os dados utilizados para a construção dos modelos foram dos anos de 2014 e 2015, oriundos do Censo Escolar fornecido pelo INEP. Após diversos métodos de filtros e tratamentos utilizados nos dados, foi realizada uma análise sobre correlação com dados referentes a evasão de alunos. Observamos que a maioria das variáveis influentes estão relacionadas a Faixa etária, recursos das escolas e localização. Com todas as análises, restaram 26 variáveis nos dados de 2014 e 20 nos dados de 2015 para criação dos modelos. Foi adotado o método de subamostragem para solucionar o problema de desbalanceamento de classes ocasionado pela desproporção de quantidade de alunos que evadiram (classe minoritária) em relação aos que não evadiram (classe majoritária).

Dos modelos utilizados, o que apresentou melhor acurácia foi o que apresentou melhor acurácia foi o que se construiu utilizando Florestas Aleatórias, que utiliza uma combinação de um número alto de árvores de decisão na sua composição seguido pelo algoritmo Árvores de decisão e Regressão logística, respectivamente. A acurácia dos modelos foi positiva, mas ainda pode ser bastante melhoroada com mais estudos considerando fatores mais específicos do conjunto de dados.

Como trabalhos futuros, pretende-se analisar outras formas de tratamento para a base de dados desbalanceadas, pois o método utilizado causa uma diminuição da amostra, o que pode ocasionar em perda de informação e, consequentemente, em uma acurácia pior. Também pretende-se realizar testes com mais algoritmos ensemble.

Esperamos que os resultados e discussões desta pesquisa possam forne- cer uma pequena contribuição no processo de solução de um problema tão sério e complexo que é a evasão escolar.

REFERÊNCIAS

ALMEIDA, Eliana Silva de (Ed.). MACHINE LEARNING: CATALISADOR DA CIÊNCIA. [S.l.]: SBC, jan. 2019. Disponível em: <http://sbc.org.br/images/flippingbook/computacaobrasil/computa_39/pdf/CompBrasil_39_180.pdf>. Citado 1 vez na página 15.

BAKER, Ryan Shaun Joazeiro de; ISOTANI, Seiji; CARVALHO, Adriana Maria Joazeiro Baker de. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, 2011. Citado 3 vezes nas páginas 9, 22.

BEZERRA, Camila et al. Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes. **Anais do XXVII Simpósio Brasileiro de Informática na Educação**, 2016. Citado 2 vezes nas páginas 13, 24.

Bishop, C.M. (2006). **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Springer.

BRASIL. **Índice de abandono escolar é três vezes maior no 6º ano do ensino fundamental**. [S.l.: s.n.], 2015. Disponível em: <http://www.brasil.gov.br/noticias/educacao-e-ciencia/2012/05/indice-de-abandono-escolar-e-tres-vezes-maior-no-6o-ano-do-ensino fundamental>. Acesso em: 3 maio 2019. Citado 2 vezes nas páginas 9, 13.

CALIXTO, Kennet E. A; SEGUNDO, Caetano V. N.; GUSMÃO, Renê P. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. **Anais do XXVIII Simpósio Brasileiro de Informática na Educação**, 2017. Citado 4 vezes nas páginas 9, 10, 20, 24.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. [S.l.], 2009. Citado 11 vezes nas páginas 11, 19–22.

Elrahman, S. M. A. and Abraham, A. (2013). A review of class imbalance problem. Journal of Network and Innovative Computing, 1(2013):332–340.

FERREIRA, Luiz Antonio Miguel. Evasão escolar. **Encontros Pela Educação**, 2000. Citado 5 vezes nas páginas 10, 13, 14.

FINI, Lucila Diehl Tolaine. **Psicologia: Ciência e Profissão**. [S.l.]: Conselho Federal de Psicologia, 1989. Citado 1 vez na página 13.

GAMA, J. Árvores de Decisão, 2000.

He, H. & Garcia, E.A. (2009). Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, 21, 1263–1284.

INEP. **Inep divulga dados inéditos sobre fluxo escolar na educação básica**. [S.l.: s.n.]. Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/b4aqv9zfy7bv/content/inep-divulga-dados-ineditos-sobre-fluxo-escolar-na-educacao-basica/21206>. Acesso em: 3 maio 2019. Citado 1 vez na página 9.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de M´aquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, 2003. Citado 6 vezes nas páginas 15–18.

RABELO, Humberto et al. Utilização de técnicas de Mineração de Dados Educacionais para a predição de desempenho de alunos de EaD em Ambientes Virtuais de Aprendizagem. **Anais do XXVIII Simpósio Brasileiro de Informática na Educação**, 2017. Citado 2 vezes nas páginas 19, 23.

RIGO, Sandro J. et al. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, v. 22, n. 1, 2014. Citado 2 vezes nas páginas 13, 22.

RUSSEL, Stuart J. **Inteligência artificial**. 3. ed. [S.l.]: Elsevier, 2013. Citado 4 vezes nas páginas 15–17.

SANCHES, André Rodrigo. **Uma Visão Geral Sobre Mineração de Dados**. [S.l.], 2003. Citado 10 vezes nas páginas 15–17, 19.

TORRES MASCHIO, Pedro de et al. Um Panorama Acerca da Mineração de Dados Educacionais no Brasil. **Anais do XXIX Simpósio Brasileiro de Informática na Educação**, 2018. Citado 1 vez na página 9.

UOL. **Brasil tem 3ª maior taxa de evasão escolar entre 100 países, diz Pnud**. [S.l.: s.n.], mar. 2013. UOL. Disponível em: <https://educacao.uol.com.br/noticias/2013/03/14/brasil- tem- 3- maior- taxa- de- evasao- escolar- entre- 100- paises- diz-pnud.htm>. Acesso em: 10 maio 2019. Citado 1 vez na página 9.

Vapnik, V.N. (1995). **The nature of statistical learning theory**. Springer-Verlag New York, Inc.

WOOD, Johnny. **This AI outperformed 20 corporate lawyers at legal work.** [S.l.: s.n.], nov. 2018. THE WORLD ECONOMIC FORUM. Disponível em: <www.weforum.org/agenda/2018/11/this-ai-outperformed-20-corporate-lawyers-at-legal-work/>. Acesso em: 6 maio 2019. Citado 1 vez na página 10.