UNIVERSIDADE FEDERAL DO CEARÁ

***CAMPUS* DE SOBRAL**

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E COMPUTAÇÃO

MESTRADO ACADÊMICO EM ENGENHARIA ELÉTRICA E COMPUTAÇÃO

IGOR ANTÔNIO GOMES TELES

APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DO ABANDONO E EVASÃO DOS ESTUDANTES DO ESTADO DO CEARÁ

SOBRAL

2023

IGOR ANTÔNIO GOMES TELES

APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DO ABANDONO E EVASÃO DOS ESTUDANTES DO ESTADO DO CEARÁ

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia Elétrica e Computação do Programa de Pós-Graduação em Engenharia Elétrica e Computação do *Campus* de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em . Área de Concentração: .

Orientador: Prof. Dr. Carlos Alexandre Rolim Fernandes.
Coorientadora: Profª. Drª. Alesandra de Araújo Benevides.

SOBRAL

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

T272a Teles, Igor Antônio Gomes.
APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DO ABANDONO E EVASÃO DOS ESTUDANTES DO ESTADO DO CEARÁ / Igor Antônio Gomes Teles. – 2023.
115 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Campus de Sobral, Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Sobral, 2023.
Orientação: Prof. Dr. Carlos Alexandre Rolim Fernandes.
Coorientação: Prof. Dr. Alesandra de Araújo Benevides.

1. Evasão Escolar;. 2. Violência; . 3. Machine Learning; . 4. Predição.. I. Título.

CDD 621.3

IGOR ANTÔNIO GOMES TELES

APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DO ABANDONO E EVASÃO DOS ESTUDANTES DO ESTADO DO CEARÁ

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia Elétrica e Computação do Programa de Pós-Graduação em Engenharia Elétrica e Computação do *Campus* de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em . Área de Concentração: .

Aprovada em: 27 de Janeiro de 2023

BANCA EXAMINADORA

Prof. Dr. Carlos Alexandre Rolim Fernandes (Orientador)
Universidade Federal do Ceará (UFC)

Profª. Drª. Alesandra de Araújo Benevides (Coorientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Iális Cavalcante de Paula Júnior
Universidade Federal do Ceará (UFC)

Prof. Dr. João Cesar Moura Mota
Universidade Federal do Ceará (UFC)

Agradeço a Deus e à minha família pelo amor, apoio e esperança. Sem vocês, eu não teria chegado até aqui. Sou grato por tudo.

AGRADECIMENTOS

A Deus, que me deu forças para continuar minha jornada tanto acadêmica quanto profissional, me levando a conquistar tudo o que tenho hoje.

Aos meus pais, Maria de Jesus Gomes Teles e Raimundo Nonato de Araújo Teles, com todo meu amor e gratidão, por tudo o que fizeram por mim ao longo de minha vida. Desejo poder ter sido merecedor do esforço dedicado por vocês em todos os aspectos, especialmente quanto à minha formação.

Aos meus dois grandes amores que partiram dessa vida durante o desenvolvimento desse trabalho, minha vó, Josefa Maria de Jesus (*in memorian*) e minha querida tia Adriana Gomes dos Santos (*in memorian*), espero que possam contemplar de onde estiverem cada ação e láurea realizada por mim e meu irmão, pois somos fruto do carinho, educação e conhecimentos repassados por vocês.

Ao meu irmão, Ramon Handerson Gomes Teles, por todo incentivo e auxílio desde meu início escolar até os momentos atuais de minha vida, obrigado por ser modelo de pesquisador e cientista para mim.

A minha companheira de vida e noiva, Quiteria Larissa Teodoro Farias, minha gratidão e amor por você por ter aceitado compartilhar a vida comigo e dado apoio e forças durante os momentos difíceis enfrentados durante essa jornada.

Ao Prof. Dr. Carlos Alexandre Rolim Fernandes e a Profª Dra. Alessandra Araújo Benevides, pela excelente orientação, além das contribuições prestadas e paciência durante a construção desse trabalho.

À SEDUC, que por intermédio e sob supervisão da Profª Dra. Alessandra concedeu acesso aos dados do SPAECE para o estudo.

À Secretaria de Segurança Pública e Defesa Social – SSPDS/CE. Especificamente ao setor de Gerência de Estatísticas e Geoprocessamento – GEESP por ter cedido os dados para o presente estudo.

À CAPES, pelo apoio financeiro com a manutenção da bolsa de auxílio. Aos colegas da turma de mestrado, pelas reflexões, críticas e sugestões recebidas.

“Cenários não dizem respeito a predizer o futuro, e sim a perceber e analisar os futuros no presente.” (Peter Schwartz)

RESUMO

O abandono e a evasão escolar são temas frequentes na Educação. Os números dão ideia do tamanho do problema. Em 2018, cerca de quatro em cada dez brasileiros de 19 anos não concluíram o Ensino Médio com base na Pesquisa Nacional por Amostra de Domicílios Contínua (PnadC), do IBGE. O abandono ocorre quando o aluno deixa de frequentar as aulas durante o ano letivo. Já a evasão escolar diz respeito à situação do aluno que abandonou a escola ou reprovou em determinado ano letivo, e que no ano seguinte não efetuou a matrícula para dar continuidade aos estudos. Dito isso, o propósito deste projeto é propor modelos de predição de situações de evasão e abandono para alunos do estado do Ceará, usando bases de dados sociais, de desempenho escolar e e em registros das mães nas bases de dados CVLI e Maria da Penha. Outro propósito do trabalho é determinar quais fatores são os que mais impactam na evasão e abandono. Foram utilizados os dados longitudinais dos anos de 2012 a 2019 dos dados escolares obtidos do Censo Escolar para verificar a situação dos alunos que evadiram ou abandonaram. No total, foram usadas 4 bases de dados: Censo escolar, SPAECE, CVLI e Maria da Penha. Os procedimentos foram realizados através do sistema gerenciador de banco de dados Postgresql, Software SPSS e o Weka. Após o pré-processamento, limpeza e aplicação de filtros, os dados foram utilizados para treinamento da máquina e verificação de predição para tomadas de decisão acerca de possíveis situações de evasão e abandono. Foram utilizados os classificadores Multilayer Perceptron (MLP), Support Vector Machine (SVM) e Floresta aleatória, foi também aplicado Correlation based feature selection - CFS para encontrar os melhor atributos para o estudo, sendo selecionados como atributos o desempenho em português e matemática, etnia, etapa de ensino e o indicador da presença da mãe em bases de violência. Foram alcançadas as respectivas acuracias 83,9

Palavras-chave: Evasão Escolar; Violência; Machine Learning; Predição.

ABSTRACT

Leaving and dropping out of school are frequent themes in Education. The numbers give an idea of the size of the problem. In 2018, around four out of ten 19-year-old Brazilians did not finish high school based on the Continuous National Household Sample Survey (PnadC), by IBGE. Dropout occurs when the student stops attending classes during the school year. School dropout, on the other hand, concerns the situation of the student who dropped out of school or failed in a given school year, and who in the following year did not enroll to continue his/her studies. The purpose of this project is to propose models for predicting dropout and dropout situations for students in the state of Ceará, using social databases, school performance and mothers' records in the CVLI and Maria da Penha databases. Another purpose of the work is to determine which factors have the most impact on evasion and abandonment. Longitudinal data from the years 2012 to 2019 of school data obtained from the School Census were used to verify the situation of students who dropped out or dropped out. In total, 4 databases were used: School Census, SPAECE, CVLI and Maria da Penha. The procedures were carried out through the Postgresql database management system, SPSS Software and Weka. After pre-processing, cleaning and applying filters, the data were used for machine training and prediction verification for decision-making about possible situations of evasion and abandonment. The Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Random Forest classifiers were used, Correlation based feature selection - CFS was also applied to find the best attributes for the study, with performance in Portuguese and Mathematics, ethnicity being selected as attributes. , teaching stage and the indicator of the mother's presence in bases of violence. The respective accuracies of 83.9

Keywords: School Evasion; Violence; Machine Learning; Prediction.

LISTA DE FIGURAS

Figura 1 – Panorama da Ciência de Dados 24
Figura 2 – Etapas do KDD 25
Figura 3 – Exemplo de árvore de decisão prevendo o abandono escolar dos alunos . . . 29
Figura 4 – Topologia de uma MLP 31
Figura 5 – Fluxograma metodológico do KDD 39
Figura 6 – Percentual de dados duplicados por ano de Censo 42
Figura 7 – União de dados do censo escolar 44
Figura 8 – Instrução de união da base Censo com Spaece 45
Figura 9 – Instrução de verificação na base Maria da Penha 46
Figura 10 – Instrução de verificação na base CVLI 46
Figura 11 – Instrução de verificação do aluno na base CVLI 46
Figura 12 – Categorias da Etnia para situação de Evasão 52
Figura 13 – Categorias da Etnia para situação de Abandono 53
Figura 14 – Categorias da Etnia para situação de Concluiu 53
Figura 15 – Distribuição de sexo por etapa de ensino e situação dos alunos 53
Figura 16 – Alunos que possuem mães em base de violência 54
Figura 17 – Percentual de casos de violência por classe 55
Figura 18 – Percentual de casos de violência por classe 55
Figura 19 – Matriz de confusão MLP 59
Figura 20 – Matriz de confusão SVM 60
Figura 21 – Matriz de confusão Floresta aleatória 61

LISTA DE TABELAS

Tabela 1 – Algumas funções do Kernel 32
Tabela 2 – Padrões de desempenho SPAECE 35
Tabela 3 – Variáveis CENSO 41
Tabela 4 – Variáveis SPAECE 43
Tabela 5 – A Matriz de Confusão 49
Tabela 6 – Situação do desempenho 50
Tabela 7 – Média de idades por situação 51
Tabela 8 – Resultado da seleção do recurso CFS no conjunto de dados 58
Tabela 9 – Sensibilidade dos treinamentos 62

LISTA DE ABREVIATURAS E SIGLAS

AIR	Instituto Americano de Educação
CVLI	Crimes violentos letais e intencionais
IBGE	Instituto Brasileiro de Geografia e Estatística
ITS	Intelligent Tutoring System
KDD	Knowledge Discovery in Database
LMS	Learning Management System
MDE	Mineração de Dados Educacionais
MD	Mineração de dados
MEC	Ministério da Educação
SGBD	Sistema Gerenciador de Banco de Dados
SPAECE	Sistema Permanente de Avaliação da Educação Básica do Ceará
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO 13
1.1 OBJETIVOS 15
1.1.1 Objetivos gerais 15
1.1.2 Objetivos específicos 16
2 TRABALHOS RELACIONADOS 17
2.1 Evasão e Abandono escolar 17
2.2 Modelagem preditiva do abandono escolar na educação 19
2.3 A violência doméstica e o risco da evasão escolar 21
3 FUNDAMENTAÇÃO TEÓRICA 22
3.1 Evasão e abandono escolar 22
3.2 Ciências de Dados 23
3.3 Mineração de dados Educacionais 23
3.4 Knowledge Discovery in Database (KDD) 25
3.5 Aprendizado de Máquina 26
3.5.1 Modelagem Preditiva 26
3.5.2 Aprendizagem supervisionada, não supervisionada e por reforço 27
3.5.3 Tarefas de aprendizado de máquina 28
3.5.4 Classificadores 28
3.5.4.1 Florestas Aleatórias *29*
3.5.4.2 Perceptron multicamadas (Multilayer Perceptron - MLP) *30*
3.5.4.3 Máquina de Vetor de Suporte (Support Vector Machine - SVM) *31*
3.5.4.4 Seleção de atributos baseada em correlação (Correlation based feature selection - CFS) *32*
3.6 Censo Escolar 33
3.7 SPAECE 34
3.8 Base de Dados de Violência 37
4 MATERIAIS E MÉTODOS 39
4.1 Levantamento das bases de dados 39
4.2 Seleção das variáveis 40
4.2.0.1 Seleção de dados da base CENSO *40*

4.2.1 *Seleção de dados da base SPAECE* 42
4.2.2 *Seleção de dados das bases CVLI e Maria da Penha* 43
4.3 **União de dados** 44
4.3.1 *União dos dados do Censo* 44
4.3.2 *União dos dados SPAECE com CENSO* 45
4.3.3 *União dos dados Maria da Penha e CVLI* 45
4.4 **Treinamento dos dados** 47
4.5 **Métricas de desempenho** 48
5 **RESULTADOS OBTIDOS** 50
5.1 **Análises de desempenho dos alunos** 50
5.2 **Modelo de predição** 57
5.2.1 *Variáveis selecionadas para o treinamento* 57
5.2.2 *Resultados dos treinamentos e testes dos dados no modelo de predição* . . . 58
6 **CONCLUSÕES E TRABALHOS FUTUROS** 63
REFERÊNCIAS 65
APÊNDICE 68
ANEXOS 71

1 INTRODUÇÃO

O abandono e a evasão escolar são temas frequentes na educação. Os números relativos a estes tópicos dão ideia do tamanho do problema. De acordo com a pesquisa nacional por amostra de domicílios contínua (PNADC), do IBGE, em 2018, cerca de quatro em cada dez brasileiros de 19 anos não concluíram o ensino médio. Essas duas situações possuem várias formas de interpretação e essa diversidade de conceituação torna imprecisa a quantificação dos casos, dificultando o estudo das causas e dos princípios desse problema que perdura até hoje.

Sobre a evasão escolar, trata-se da fuga ou desistência da escola em função da realização de outra atividade. A diferença entre evasão e abandono escolar foi utilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira/Inep (1998). Nesse caso, o termo abandono significa a situação em que o aluno se desliga da escola, mas retorna no ano seguinte, enquanto que o termo evasão remete ao aluno que sai da escola e não volta mais para o sistema escolar.

Administradores e professores de escolas têm se esforçado para reduzir o abandono escolar há algum tempo (ELLIOTT; VOSS, 1974), mas este fenômeno continua a persistir nas escolas como um problema durante os dias atuais (Wiltz e Slate, 2016). O abandono e evasão não são considerados apenas um problema educacional sério, mas também um grave problema social, é algo que afeta até mesmo países com alto desenvolvimento econômico.

Os autores Wiltz e Slate (2016) citam dentre as características individuais dos alunos os seguintes pontos. O desempenho educacional, que consiste no desempenho acadêmico. A mobilidade no ensino fundamental, comportamento e atitudes do aluno, que consiste no envolvimento acadêmico em seu aspecto das atividades escolares. E as atividades sociais e as características demográficas, que envolvem e analisam o meio onde vivem.

Para Silva e Araújo (2017), fatores sociais, culturais, políticos e econômicos levam à situação de abandono e evasão. Pontua-se também a escola, onde os educadores têm colaborado a cada dia para o problema se agravar, mediante a utilização de um método didático superado ou de uma prática cristalizada como por inexperiência, acaba por desenvolver o conteúdo de forma descontextualizada e sem sentido para o aluno.

De acordo com Vecina e Ferrari (2002), a criança e o adolescente necessitam de uma relação afetiva e estável para ir, por intermédio dela, construindo sua identidade, ou seja, necessita de uma família equilibrada e protetora para que seu processo de identificação não seja conturbado. Corroborando essa ideia, Cardia (1997) pondera que crianças e adolescentes que

vivem em um ambiente marcado por violência aprendem a usá-la como forma de vida e têm grande possibilidade de reproduzi-la na vida adulta.

Ristum (2010) ressalta, ainda, que crianças e adolescentes que testemunham a violência doméstica dentro de casa ou que são agredidas pelos pais tendem a apresentar uma série de consequências na vida escolar, como dificuldade de concentração, de aprendizagem e de relacionamento com os colegas e professores, sendo esses, fatores que aumentam os riscos de evasão e abandono.

Crianças e jovens que crescem nesse meio de violência, muitas vezes, respondem aos conflitos cotidianos e à necessidade de autoafirmação, tão típicos da juventude, usando a linguagem aprendida, da violência. Quando tais incidentes ocasionam uma morte, uma espiral de agressões e de vinganças recíprocas envolvendo grupos de jovens gera inúmeras outras vítimas fatais, sendo que o rastro da origem de todos os problemas há muito foi apagado por uma sequência de eventos, tornando invisíveis para a sociedade as consequências do aprendizado da violência intrafamiliar.

Dito isso e observando o papel da família, em especial o da mãe, no processo educacional do aluno, dentro desse contexto, o presente projeto objetiva traçar o perfil dos alunos com caráter de evasão e abandono que possuem registros da mãe nas bases de violência doméstica (Lei Maria da Penha) e de Crimes violentos letais intencionais (CVLI). Através do uso de aprendizagem de máquina, objetiva-se prever o abandono escolar como base para uma intervenção direcionada.

Tendo em vista o problema, os agentes que influenciam e a definição do que se caracteriza como abandono e evasão, faz-se necessária a criação de um modelo de predição dos riscos de evasão e abandono desses estudantes para que medidas possam ser tomadas o quanto antes, impedindo ou diminuindo a ocorrência dessas situações que perduram ao longo dos anos.

A modelagem preditiva usando aprendizado de máquina com big data já vem sendo aplicada com sucesso em outras áreas, incluindo negócios e saúde pública, e tem um grande potencial para construir sistemas de alerta precoce para identificar possíveis evasões ao ensino médio.

O presente estudo utilizou aprendizagem de máquina nos dados relativos à presença do aluno ao longo dos anos no CENSO escolar, seu desempenho na prova SPAECE, e a verificação da presença da mãe em bases de violência, para prever a evasão e o abandono dos alunos na faixa de 5º e 9º ano no estado do Ceará. O modelo preditivo permite a identificação

de alunos em risco desde o início, ajudando no desenvolvimento de ações que visem diminuir a desistência da jornada estudantil.

Foi utilizada a descoberta de conhecimento em base de dados (Knowledge Discovery in Databases – KDD) para organização dos dados e pré-processamento. Os dados correspondem à união da base de dados CENSO, SPAECE, CVLI e Maria da penha nos anos de 2012 à 2019. Após a organização e mineração dos dados, onde foi realizada a limpeza dos dados como remoção de informações duplicadas ou inconsistentes e a seleção dos melhores atributos de entrada usando grid search, ocorreu o treinamento dos dados. Para essa etapa foi utilizado o software Weka e aplicados os classificadores de Machine Learning (ML), Multilayer Perceptron (MLP), Support Vector Machine (SVM) e Floresta aleatória para realização da predição da situação do aluno e para verificar os percentuais de chances de evasão e conclusão do aluno. Foram obtidos os seguintes resultados das acurácias dos respectivos classificadores, 83,9%, 81,24% e 71,3%, de onde se conclui que o classificador MLP obteve o melhor resultado.

Outro objetivo do presente trabalho foi analisar quais são as variáveis que mais contribuem para a evasão e abandono, através de uma análise do ganho de informações dos parâmetros de entrada envolvidos, para isso foi utilizado o Correlation based feature selection - CFS, onde foi possível obter as variáveis mais valiosas para o estudo e treinamento dos dados. Este algoritmo de seleção de variáveis selecionou como fatores mais relevantes as duas variáveis de desempenho, que correspondem às notas do aluno em matemática e português, a variável cor da raça, que classifica a etnia, a variável etapa, que pontua a série que o aluno pertence, 5º ou 9 ano, e as variáveis provindas das bases de dados de violência, clvi e maria da penha, que sinalizam para cada amostra se a mãe do aluno está contida em uma das duas bases.

1.1 OBJETIVOS

1.1.1 Objetivos gerais

Este estudo tem como objetivo geral a proposição de modelos de predição de situações de evasão e abandono de alunos do estado do Ceará, com base em dados sociais, de desempenho escolar e em registros das mães nas bases de dados CVLI e violência doméstica.

Outro objetivo é encontrar quais atributos que mais impactam nesse estudo. Foram utilizados vários algoritmos de aprendizado de máquina com o intuito de que possa ser analisado também qual apresenta melhor acurácia nessa situação com a referida base de dados.

1.1.2 Objetivos específicos

Para esse estudo os seguintes objetivos específicos foram considerados.

- Elaborar e executar a metodologia de coleta, tratamento e análise do banco de dados usado neste estudo.
- Avaliar e descrever as características mais relevantes para definição do perfil de alunos em situação de evasão e abandono;
- Determinar o modelo de previsão a ser utilizado junto a seus classificadores.
- Analisar quais variáveis e fatores foram mais importantes para o estudo e influenciaram no treinamento dos dados.

2 TRABALHOS RELACIONADOS

Nesta seção serão abordados alguns trabalhos relacionados sobre os estudos acerca de evasão, abandono escolar e modelagem preditiva.

2.1 Evasão e Abandono escolar

Muitos estudos retratam o porquê das causas de evasão e abandono e examinam indicadores de risco para evasão. Balfanz, Herzog e Mac Iver (2007) usaram dados longitudinais para 13.000 estudantes de 1996-2004 e descobriram que os estudantes que estavam frequentemente ausentes, marcados com pontuações de comportamento insatisfatório, reprovado em matemática e reprovado em inglês tiveram 68%, 56%, 54% e 42% menos probabilidade de se formar do que outros.

Burke (2015), usando amostras de 6.118 alunos que entraram na nona série no distrito escolar de Oregon em 2007-2008, descobriu que os indicativos de presença, que possuíam menos de 80% e do GPA (Grade Point Average) abaixo de 2,0, foram indicadores fundamentais de evasão. A National High School Center oferece orientação sobre o desenvolvimento de um sistema de alerta e também é recomendado o uso de atendimento, comportamento e desempenho do curso como indicadores-chave (Therriault et al., 2010). Para esse estudo foi utilizado o modelo de regressão GPA com análises estatísticas e cálculos probabilísticos com logit, onde p é uma probabilidade de sucesso em um determinado evento, então p / (1 p) correspondente a chance do mesmo. Logo o logit da probabilidade é o logaritmo das chances.

Na Coréia, Yoon, Ryu e Kim (2010) realizaram uma análise aprofundada das evasões com dados de 12.280 estudantes em situação de abandono no ensino fundamental e 14.572 estudantes em situação de abandono no ensino médio. Em resposta aos formulários das causas secundárias do abandono escolar do estudante, a maior parte dos entrevistados respondeu baixo desempenho acadêmico ou não gosto de estudar (34,8%), seguido de preparação para exame de qualificação escolar (8,8%); formação em línguas estrangeiras (8,1%); dificuldades financeiras, distúrbios emocionais, distúrbios físicos, e outras doenças (4,2%); problemas familiares (3,2%); regras escolares rígidas (2,5%); residencial instabilidade (2,4%); fugir de casa (1,5%); educação alternativa (1,3%); e pobre relacionamento com amigos e professores (0,7%). Rumberger e Lima (2008) afirmam que poucas evidências foram encontradas para a importância dos recursos escolares como causa da evasão e abandono escolar, mas existe uma evidência forte e favorável

quando existem pequenas classes de alunos, ou seja, o clima escolar e o clima acadêmico têm grande importância. Finalmente, características da comunidade na qual está inserido, como níveis de pobreza ou riqueza, para além das características familiares, parecem exercer um papel importante na decisão de abandono e sucesso escolar. Rumberger e Lim (2008) realizaram uma revisão de 203 artigos publicados nos Estados Unidos, nos últimos 25 anos, e dividiram os fatores que predizem se o estudante abandona ou consegue concluir o ensino médio em dois grupos: aqueles associados às características individuais dos alunos e os associados às características institucionais de suas famílias, escolas e comunidades.

Das características institucionais são mencionados os seguintes pontos: a estrutura familiar ao longo do processo escolar, a renda e recurso familiar e o capital social tais como expectativas educacionais e acompanhamento do progresso escolar dos filhos. Além disso, são apontadas algumas características escolares como a composição dos estudantes da escola; os recursos escolares; as políticas e práticas da escola; e outras características estruturais.

Sob o ponto de vista da participação dos pais, Burchinal et. al. (2002) afirmam que as crianças tendem a mostrar melhores habilidades acadêmicas se os pais tiverem maior envolvimento e maior grau de escolaridade. Descrevem também que, quando os pais são mais participativos, as crianças evidenciam maior competência para a leitura, diminuindo os riscos de insucesso escolar nesse aspecto.

Na questão da participação dos pais na vida escolar dos filhos, encontra-se também como fonte de conflitos nessa relação o nível de escolarização dos pais. Alguns pais pouco podem ajudar, pois demonstram carência de orientação, muitas vezes, por não serem instruídos para agir e orientar seus filhos ou por não terem recebido esse acompanhamento em fase estudantil também. Contudo, na visão da escola, quando a vida escolar apresenta problemas, tanto os pais como os filhos são responsáveis pelo desempenho insatisfatório, na visão da escola (Peisner, Ellen, Yazejian, 2002).

Ainda sobre a participação dos pais no universo escolar, Sígolo e Lollato (2001) enfatizam o grau de aproximação entre a escola e a família. Revelam que a mãe, com maior frequência, é quem acompanha as atividades escolares dos filhos e, a partir da realização de tarefas em casa, os pais podem perceber o desenvolvimento ou não de novos comportamentos.

Lopes (2010) ressalta que, para a amenização de alguns problemas referentes à evasão, é necessária uma ação firme dos poderes públicos, principalmente em relação aos gestores escolares, que precisam assegurar um bom ensino e aprendizagem. Desempenho

ruim também é um fator de evasão; oposto a isso, há alunos que evadem por não se sentirem "desafiados e estimulados".

Em um apanhado geral da literatura sobre abandono escolar, em 203 estudos no assunto, chegam-se a algumas conclusões relevantes: notas baixas no início do processo educativo é um forte aspecto de previsão de futuro abandono; desempenho inadequado frequente costuma implicar reprovação; faltas, atos delinquentes e abuso de substâncias ilegais são fortes preditores de abandono. Essa superação poderá acontecer em um ambiente familiar estável, e o acesso a recursos sociais e financeiros influencia de forma significante a probabilidade de o estudante completar seus estudos (RUMBERGER E LIMA, 2008).

A partir destes trabalhos, verifica-se que o governo e a sociedade precisam cumprir seus deveres educacionais, evitando os alunos de deixar as escolas e fornecer aos alunos em risco serviços educacionais adequados. Cada ano, na Coreia, país de primeiro mundo, cerca de 50.000 alunos abandonam as escolas (KEDI, 2018). Para fornecer-lhes assistência significativa, há uma necessidade urgente de esforços para desenvolver planos que impeçam a evasão escolar e ajudem jovens fora da escola (Battin-Pearson et al., 2000).

2.2 Modelagem preditiva do abandono escolar na educação

No que concerne às técnicas de predição no contexto do abandono escolar, as literaturas anteriores usavam principalmente modelos projetados para compreender as causas por trás dos comportamentos de abandono. Eles são muito úteis para revelar a estrutura de várias camadas das causas que explicam por que os jovens abandonam as escolas. No entanto, essas literaturas não são tão precisas à pronta aplicação em salas de aula reais porque sem dados sobre vários fatores que afetam comportamentos de evasão, é difícil identificar os níveis de risco individuais dos alunos.

Quando a Aprendizagem de Máquina é empregada como técnica de MDE, considera-se que ela está incluída na primeira possibilidade prevista por Baker (2000), com a finalidade de entender situações educacionais e com isso dar apoio no processo de tomada de decisão, nesse sentido há muitos estudos que foram e estão sendo desenvolvidos, entretanto alguns podem ser destacados, como: O estudo realizado pelo Departamento de Educação e Treinamento em Victoria, Instituto Americano de Educação (AIR), Dejaeger et al. (2011) Cortez e Silva (2008) e Zhang e Wu (2019).

O Departamento de Educação e Treinamento em Victoria, Austrália, desenvolveu

a Ferramenta de Mapeamento do Estudante (SMT) para identificar alunos que não estão no sistema educacional (Lamb e Rice, 2008). Nos Estados Unidos, quem abandonou o ensino médio tornou-se alvo da atenção dos legisladores quando seu número disparou no início dos anos 1960 (Rumberger e Larson, 1998).

O Instituto Americano de Educação (AIR) desenvolveu o Sistema de Alerta Precoce para diagnosticar alunos em risco no ensino médio (Therriault, Heppen, O'Cummings, Fryer, e Johnson, 2010). Knowles (2015) criou o *Wisconsin Dropout Early Warning System* (DEWS), um modelo preditivo do risco de abandono escolar para alunos do sexto ao nono ano usando o aprendizado de máquina e MD. Estes sistemas de diagnóstico para os alunos de risco determinam os níveis de risco usando os registros dos alunos em suas vidas escolares, em vez de utilizar variáveis pessoais identificadas pela pesquisa.

Quando os alunos do ensino fundamental e médio desistem de aprender, isso se traduz em custos sociais para toda a sociedade, bem como grandes perdas para a vida do indivíduo (Finn, 1989). Embora seja fundamental ajudar os jovens que não frequentam a escola e os que abandonaram, uma abordagem mais fundamental seria identificar possíveis desistências e impedi-los de deixar as escolas.

Ressalta-se, que assim como aplicado pelo AIR (2010), existem na literatura diversos trabalhos que usam diferentes técnicas de MD no contexto educacional. Singh e Kumar (2012), por exemplo, utilizaram a técnica de árvore de decisão para gerar conhecimento aos gestores da instituição para avaliar o desempenho de seus alunos. Dejaeger et al. (2011), por outro lado, utilizou a técnica chamada de clusterização de dados para identificar os principais fatores de satisfação dos alunos em duas instituições de ensino e, consequentemente, para a construção de modelos para apoiar os gestores no processo de tomada de decisão estratégica.

O objetivo de Cortez e Silva (2008) era analisar o desempenho dos alunos sob uma perspectiva de quais atributos mais influenciam na previsão do desempenho. Para isso, os autores utilizaram quatro algoritmos: Árvores de decisão, Random Forest, Redes Neurais Simples e SVM. Os resultados dos autores mostraram que uma boa precisão preditiva pode ser alcançada, desde que estejam disponíveis as primeiras e/ou segundas séries do período escolar. Cortez e Silva (2008) ressaltam ainda que o desempenho do aluno é altamente influenciado por avaliações anteriores e pelo número de faltas. Como resultado direto desta pesquisa, os autores relatam que ferramentas mais eficientes de previsão do aluno podem ser desenvolvidas, melhorando a qualidade da educação e aprimorando a gestão dos recursos escolares.

Por fim, Rodrigues et al. (2016) desenvolveram sua pesquisa no contexto de cursos e-leaning do tipo Massive Open Online Courses (MOOCs) para o desenvolvimento de sua investigação. Os autores contextualizam o problema de sua pesquisa destacando que, com o rápido desenvolvimento de cursos desse tipo, tornou-se uma questão importante na pesquisa educacional explorar as características de aprendizagem on-line e fornecer apoio à melhoria dos métodos de ensino e das atividades acadêmicas impedindo também a evasão dos alunos. Os autores utilizaram ID3, C4.5 e CART, todos baseados em Árvore de Decisão. As precisões alcançadas pelos modelos sobre a base de dados de teste foram respectivamente: 81%, 75%, 76%. Zhang e Wu (2019) afirmaram que os modelos baseados em árvores de decisão são consideravelmente simples de serem implementados, e têm precisão relativamente satisfatória.

2.3 A violência doméstica e o risco da evasão escolar

A violência doméstica é um dos acontecimentos mais recorrentes da contemporaneidade e desempenha um papel vital no atraso do desenvolvimento integral dos adolescentes. De acordo com Kefas (2016), a violência doméstica é o principal fator que impede significativamente o desempenho acadêmico e, em última análise, se destaca como o gargalo para o desempenho acadêmico dos alunos enquanto Adubi e Ashara (2018) relataram que a violência é um dreno em praticamente todos os setores da estrutura social com atendente consequências na educação, saúde, desenvolvimento econômico e humano em geral.

Os efeitos da exposição ao abuso de violência doméstica segundo CORA (2014) podem ser tão prejudiciais para crianças e adolescentes como os danos que uma pessoa abusiva inflige ao seu parceiro. Três a 10 milhões de crianças testemunham violência ou abuso doméstico. A extensão do risco e o trauma de testemunhar violência ou abuso doméstico depende da idade da criança e da duração da gravidade e frequência do abuso.

Um evento traumático pode interromper gravemente a rotina escolar e os processos de ensino e aprendizagem. Geralmente há altos níveis de transtorno emocional, potencial para comportamento perturbador ou perda de frequência do aluno, a menos que esforços sejam feitos para chegar aos alunos e funcionários com informações e serviços adicionais. Alunos traumatizados pela exposição à violência mostraram ter médias de notas mais baixas, comentários mais negativos em seus registros cumulativos, e mais faltas à escola relatadas do que outros alunos. Eles podem ter aumentado dificuldades de concentração e aprendizagem na escola e pode envolver-se de forma invulgarmente imprudente ou comportamento agressivo.

3 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos fundamentais que guiaram o desenvolvimento do trabalho, tais como o conceito do que se intitula evasão e abandono escolar, mineração de dados, aprendizado de máquina e a origem das bases dos dados, como CENSO e SPAECE.

3.1 Evasão e abandono escolar

Chung et al. (2013) apresentam definições separadas para jovens fora da escola e alunos em risco de abandono. "Jovens fora da escola" são definidos como jovens em idade escolar com menos de 19 anos que não frequentam as escolas. Mesmo que eles geralmente se sobreponham ao abandono escolar, essa categoria inclui os "pré-escolares" que não ingressaram em instituições de ensino obrigatório, os "meio do programa desistentes" que desistem após entrar nas escolas e os "desistentes pré-ingresso" que não avançam para instituições de nível superior.

Por outro lado, "jovens em risco" refere-se a jovens que estão expostos a riscos pessoais e ambientais, propensos a experimentar problemas comportamentais ou psicológicos e ter dificuldade para atingir o normal desenvolvimento sem intervenção educacional apropriada. Este grupo de jovens relatam altos riscos de fuga de casa, abandono, desemprego, violência, prostituição, drogas, abusos e outras condutas impróprias, crimes, bem como distúrbios psicológicos, como depressão, ansiedade e suicídio (Khu et al., 2005).

Embora nem todos os jovens fora da escola se tornem jovens em risco, a probabilidade de se tornar um é alta. O que também leva a fortes indícios e chances de evasão. Por este motivo, pode ser crucial que um jovem permaneça dentro da escola. Ressalta-se que a evasão escolar é a fuga ou desistência da escola em função da realização de outra atividade, conforme pontuado anteriormente por Khu (2005).

A diferença entre evasão e abandono escolar é utilizada pelo INEP, onde o termo abandono significa a situação em que o aluno se desliga da escola, mas retorna no ano seguinte, enquanto que o termo evasão remete ao aluno que sai da escola e não volta mais para o sistema escolar.

3.2 Ciências de Dados

Ciências de Dados engloba diversas áreas, como, estatística, métodos científicos, inteligência artificial (IA) e análise de dados. Basicamente, a ciências de dados desenvolve estratégias para preparar para análise, exploração e visualização de dados, e a partir desses processos é possível obter informações para tomar melhores decisões (ORACLE, 2021).

Um dado é o resultado de uma coleta de uma resposta a uma pergunta, uma situação ou problema, ou ainda em alguma mediação realizada. Quando atribui-se significado para esses dados, obtém-se as informações, e a partir do momento em que estão disponíveis para um determinado fim, será gerado o conhecimento. A ciência de dados tem o dado, a informação e o conhecimento como suas principais matérias primas. Muitas vezes confundida com uma simples análise estatística, a ciência de dados compreende desde a coleta até o descarte dos dados. Primeiro, ocorre a produção dos dados, que pode ocorrer por meio de sensoriamento, pesquisas ou coletas, esses dados podem ser mantidos em banco de dados ou planilhas. Posteriormente, eles são convertidos para formatos que sejam compatíveis com as ferramentas de análise, como o formato CSV utilizado para compatibilidade do SPSS e banco de dados PostGres.

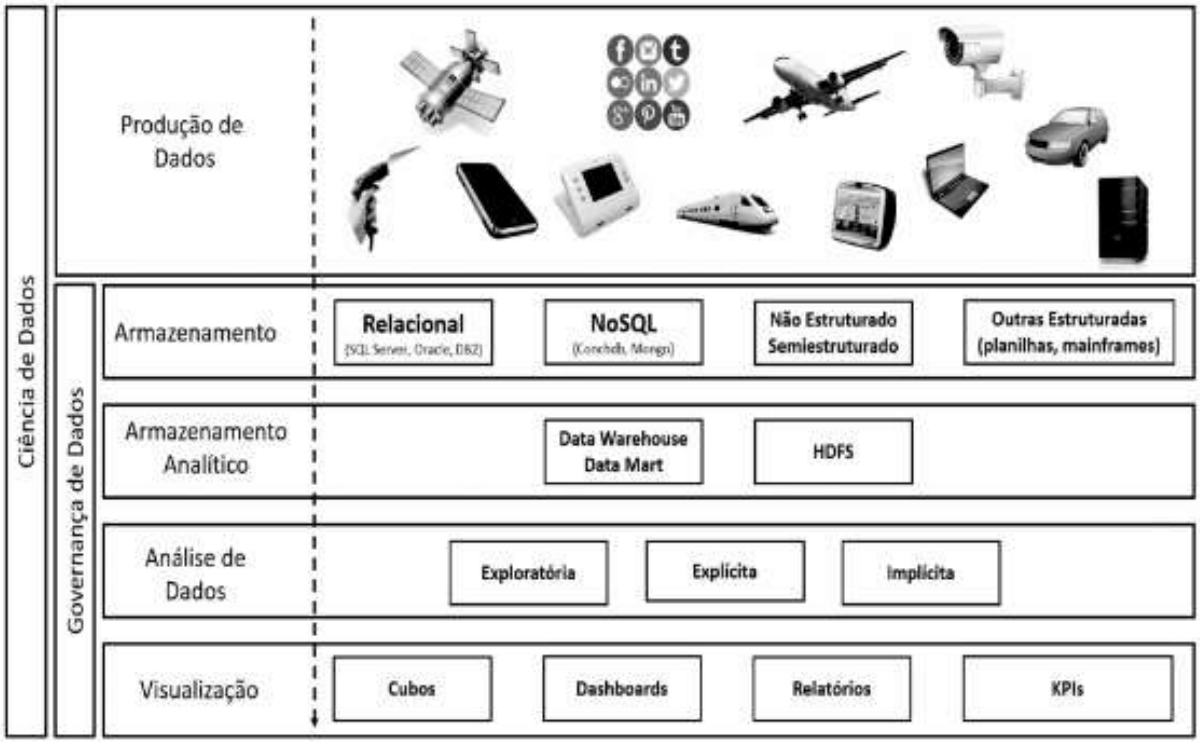
Depois de preparados os dados, direciona-se para a fase de extração de informações e conhecimento, onde, após tratamento dos dados, podem ser usadas técnicas estatísticas ou de aprendizado de máquina para este fim, neste trabalho utilizou-se os processos de KDD descritos na Subseção 4. Após essa fase, a informação é apresentada através de gráficos ou mesmo relatórios, onde o usuário pode enxergar a informação de maneira mais clara. Assim a ciência de dados acompanha todo o processo de vida do dado na busca de extrair algum conhecimento (AMARAL, 2016).

A utilização de algoritmos de mineração de dados fornece ao analisador mais informações sobre o conjunto de dados ao qual são aplicados, assim os padrões levantados podem ofertar aos pesquisadores novas perspectivas e cenários.

3.3 Mineração de dados Educacionais

Mineração de Dados (MD) é uma área de pesquisa multidisciplinar, envolvendo basicamente Ciência de Dados, estatística e aprendizado de máquina. A MD é parte principal de um processo que tem como entrada uma Base de Dados e como saída um Conhecimento (Fayyad et al., 1996). Ela é dividida em tarefas como predição, clusterização e associação que devem ser

Figura 1 – Panorama da Ciência de Dados



Fonte: (AMARAL, 2016).

escolhidas de acordo com análises exploratórias inicialmente feitas sobre os dados (Han et al., 2006).

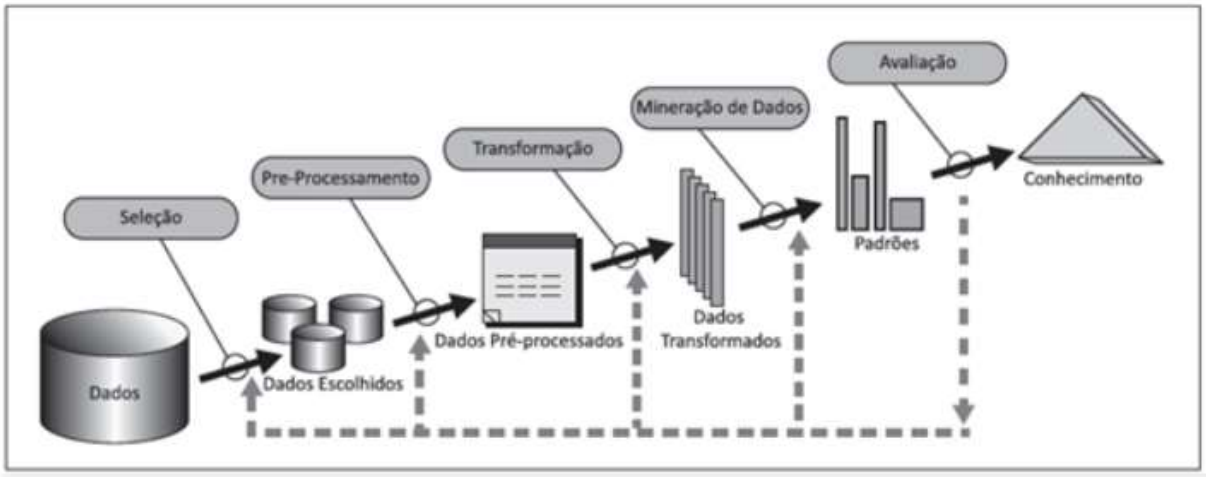
A MD tem sido amplamente utilizada em diferentes áreas, principalmente medicina, indústria, marketing, agronegócios e educação (Berry e Linoff, 2004). Na Educação, a Mineração de Dados Educacionais (MDE) é uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados oriundos de contextos educacionais (Romero e Ventura,2010; Paiva et al., 2012).

Os tipos de estudos desta área são classificados, segundo Romero e Ventura (2010) em: i) educação offline, para análises em dados de desempenho do aluno, comportamento, currículo etc, ou seja, gerados em ambientes de sala de aula; ii) aprendizado eletrônico (e-learning) e Sistema de Gestão da Aprendizagem ou LMS (do inglês, Learning Management System), para análise de dados armazenados em sistemas LMS no formato de log e bases de dados; e, iii) Sistemas Tutores Inteligentes ou ITS (do inglês *Intelligent Tutoring System*) e Sistemas Hipermídias Adaptativos Educacionais ou AEHS (do inglês, *Adaptive Educational Hypermedia System*), os quais são aplicados sobre dados de sistemas que se adaptam a cada estudante em particular, aos cursos oferecidos que estão em forma de log, aos modelos de usuários etc.

3.4 Knowledge Discovery in Database (KDD)

Conforme descrito por Fayyad, Piatetsky-Shapiro e Smyth (1996), KDD consiste em um processo não trivial que almeja a identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, que estão embutidos nos dados. As cinco etapas que constituem este processo são apresentadas abaixo.

Figura 2 – Etapas do KDD



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

A descoberta de conhecimento em base de dados (*Knowledge Discovery in Databases – KDD*) constitui-se de um processo que se inicia pela escolha dos dados que documentam de alguma maneira a pergunta que o especialista deseja responder. Os dados são integrados e pré-processados para que sejam entregues estruturados, higienizados, selecionados e padronizados à tarefa de mineração de dados.

Na tarefa de mineração, aplica-se alguma técnica inteligente capaz de encontrar soluções que auxiliam o especialista na descoberta de uma resposta. O resultado desta tarefa deve ser pós-processado para que se apresentem análises qualitativa e/ou quantitativa dos elementos encontrados e, quando possível, apresentadas de maneira que possam ser interpretadas de maneira a facilitar a tomada de decisão.

Podemos verificar que no processo do KDD há uma dupla seta na integração de todas as etapas. Isto significa que o processo pode ser sequencial, ou seja, seguindo cada passo por vez, ou iterativo, o processo pode ser executado em passos arbitrários ou mesmo repetidos caso haja necessidade.

3.5 Aprendizado de Máquina

Dentro da mineração de dados existem tarefas que podem ser realizadas para extrair informações dos conjuntos de dados. Segundo Tan et al. (2018) essas tarefas são geralmente divididas em duas grandes categorias: preditivas e descritivas.

De acordo com Tan et al. (2018) as duas tarefas tem por objetivo minimizar o erro entre a predição e os verdadeiros valores de uma variável objetivo. Ambas possuem um conjunto de técnicas que podem ser aplicadas aos conjuntos de dados. Na próxima subseção, será apresentada uma breve história sobre os tipos de aprendizado de máquina e a modelagem preditiva

3.5.1 Modelagem Preditiva

A predição, em aprendizagem de máquina, trata-se de um conjunto de técnicas de construção e uso de modelos para fazer previsões com base em padrões extraídos de dados (Kelleher, Namee, D'Arcy, 2015). Com o aumento de poder de computação e *big data* disponível, a modelagem preditiva foi aplicada com sucesso em vários campos. Por exemplo, a modelagem preditiva ajuda os médicos a diagnosticar uma doença com base em dados de pacientes anteriores e ajuda a empresa a prever as preferências dos clientes com base no passado de itens comprados.

A modelagem preditiva também tem um grande potencial na educação, prevendo evasão dos alunos, padrões de curso ideais e assim por diante. A chave da análise preditiva é construir (ou treinar) modelos para fazer previsões com base em dados anteriores, e o aprendizado de máquina é usado para treinar os modelos.

O poder do aprendizado de máquina na modelagem preditiva vem de sua ênfase na generalização de um modelo. Ou seja, um bom modelo preditivo deve ser generalizável para dados não vistos anteriormente. Deve-se levar em consideração algumas ameaças, como a generalização de um modelo, o *overfitting* , que ocorre quando um modelo ajusta muito ruído nos dados. Para se avaliar o efeito do *overfitting*, o aprendizado de máquina divide um determinado conjunto de dados em conjunto de dados de treinamento e teste, e também usa validação cruzada, que será abordada nas próximas seções. Além disso, o fato de a técnica utilizada não conseguir modelar os dados também deve ser levada em consideração.

3.5.2 Aprendizagem supervisionada, não supervisionada e por reforço

O aprendizado de máquina é um conjunto de algoritmos capazes de aprender a partir de um conjunto de dados (Goodfellow, Bengio e Courville, 2016). Normalmente, o aprendizado de máquina se enquadra em duas categorias: aprendizagem supervisionada e não supervisionada.

No aprendizado supervisionado, o aprendizado de máquina utiliza algoritmos que aprendem a relação entre recursos descritivos, os preditores ou dados de entrada (*input*) e recurso de destino, dados de saída (Output) em um conjunto de dados. Usando o modelo de treinamento da aprendizagem supervisionada, deseja-se prever com precisão o resultado de observações futuras ou para melhor compreender a relação entre o resultado e os preditores (*accuracy*). Por exemplo, o modelo para a previsão de evasão de alunos pode ser treinado usando um aprendizado supervisionado em que uma máquina usa algoritmos de aprendizagem que irão aprender a relação entre as evasões dos alunos e vários preditores.

Para a aprendizagem supervisionada, um conjunto de dados deve conter o recurso de destino (ou resultado) e características descritivas (ou preditores). Nesse sentido, o conjunto de dados para a aprendizagem supervisionada é muitas vezes chamado de conjunto de dados rotulados, o que significa que o conjunto de dados contém um rótulo(ou destino) que supervisiona o processo de aprendizado. Muitos modelos tradicionais como máquina de vetor de suporte (SVM), MLP e Floresta Aleatória são usados para aprendizagem supervisionada (James, Witten, Hastie, Tibshirani, 2013).

Usando o modelo de treinamento da aprendizagem não supervisionada, deseja-se descobrir a subjacente estrutura dos dados em vez de prever o recurso de destino. Por exemplo, em um mercado de segmentação, os pesquisadores estão interessados em identificar grupos distintos de clientes com base em múltiplas características dos clientes com a esperança de marketing direcionado. O conjunto de dados usado para a aprendizagem não supervisionada é muitas vezes chamado de dados não rotulados porque o conjunto de dados não contém o recurso de destino. Destaca-se alguns algoritmos que realizam tarefas nos conjuntos de dados como Clustering K-means, modelos de mistura, clustering hierárquico e redes neurais são exemplos usados para aprendizagem não supervisionada.

A aprendizagem por reforço (*Reinforcement learning* - RL) é uma estrutura computacional para modelar e automatizar a aprendizagem direcionada a objetivos e a tomada de decisão sequencial (RICHARD, 2018). Ao contrário de outras abordagens de aprendizado, ou seja, aprendizado supervisionado e aprendizado não supervisionado, o RL enfatiza o aprendizado

por um agente a partir da interação direta com seu ambiente.

A RL é particularmente adequada para configurações que envolvem um agente que precisa aprender uma política sobre o que fazer em diferentes situações, como mapear estados para ações, para maximizar uma utilidade de longo prazo. O agente deve explorar diferentes ações para descobrir ações de alta recompensa; crucialmente, as ações afetam não apenas a recompensa recebida imediatamente, mas também o próximo estado e, por meio disso, todas as recompensas futuras. Essas características, ações com consequências de longo prazo, recompensa atrasada e tomada de decisão sequencial sob incerteza, são as principais características da RL.

Neste estudo, será utilizada aprendizagem supervisionada para prever a evasão e o abandono dos alunos. Serão verificadas as técnicas mais utilizadas e verificada qual possui melhor acurácia. Usando o conjunto de dados do CENSO escolar, SPACE, violência doméstica e CLVI, após a realização dos filtros e limpeza dos dados, os mesmos serão treinados com os modelos, SVM, MLP e florestas aleatórias, que são um conjunto popular de algoritmos de aprendizagem supervisionada para construir o modelo preditivo de evasão de alunos.

3.5.3 Tarefas de aprendizado de máquina

As tarefas mais comuns no aprendizado de máquina são a regressão, classificação, agrupamento e redução de dimensão. A tarefa na regressão é a previsão do valor numérico (ou contínuo). Os exemplos incluem a previsão do preço das ações, preço do produto e assim por diante. A tarefa de classificação lida com a previsão de um valor categórico. Os exemplos incluem a previsão de e-mails que são ou não spam, assim como pacientes que estejam ou não com uma enfermidade, e nesse estudo com a possibilidade de evasão ou abandono, ou permanência.

A tarefa de agrupamento está relacionada à localização do agrupamento natural de dados. Exemplos incluem segmentação de clientes, agrupamento de genes e assim por diante. A redução da dimensão está relacionada à redução do número de variáveis na análise para melhorar a interpretabilidade dos resultados e a eficiência do algoritmo. O foco deste estudo, que é a previsão de evasão e abandono, corresponde ao problema de classificação binária.

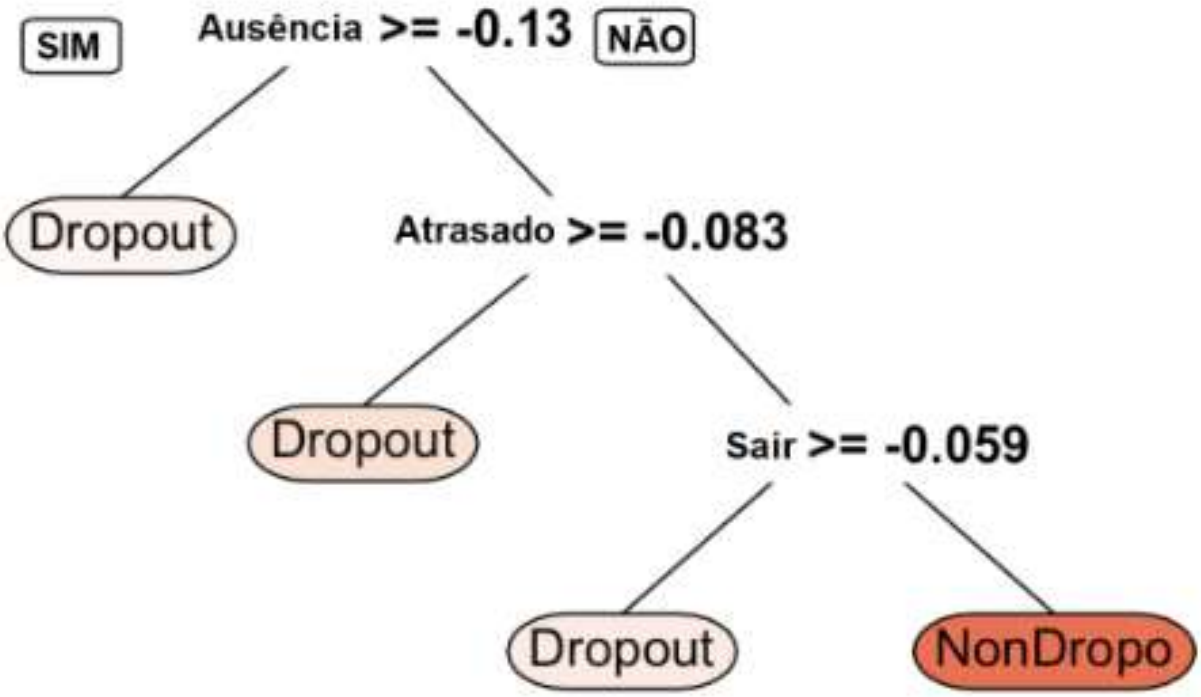
3.5.4 Classificadores

Nesta seção serão apresentados os modelos utilizados para o treinamento dos dados do presente estudo.

3.5.4.1 Florestas Aleatórias

O modelo de florestas aleatórias (Flach, 2012) é, tradicionalmente, utilizado para tarefas de classificação e regressão. No aprendizado de máquina, os ensemble methods (métodos de conjunto) produzem previsões finais combinando previsões de vários algoritmos, o que evita o sobreajuste e, portanto, melhora a precisão de previsão. Além disso, o modelo de floresta aleatória tem capacidade de seleção de recursos implícita e é menos sensível à seleção de hiperparâmetros (Couronne, Probst e Boulesteix, 2018).

Figura 3 – Exemplo de árvore de decisão prevendo o abandono escolar dos alunos



Fonte: Autor

O modelo de floresta aleatória é uma coleção de árvores de decisão geradas com base em um subconjunto aleatório do conjunto de dados original. A decisão final no modelo de florestas aleatórias pode ser feita combinando as decisões de classificação dessas árvores de decisão. A árvore de decisão, que são os blocos de construção do modelo de floresta aleatória, é um gráfico semelhante a uma árvore para classificação.

No exemplo da Figura 3, na árvore de decisão, existem dois tipos de nós: nós de decisão e nós folha. O nó de decisão (por exemplo, Ausência> = -0,13, Atrasado> = - 0,083, Sair> = -0,059) tem dois ou mais ramos para dividir e a folha nó (por exemplo, Dropout, Nondropout) representa uma classificação ou decisão final.

Nesta árvore de exemplo (Figura 3), os alunos com parâmetro de Ausência maior que -0,13 são classificados como abandono, já os alunos com Ausência menor que -0,13, Atrasado menor que -0,083 e Sair menor que -0,059 são classificados como não evasão. Como uma árvore de decisão escolhe suas divisões (por exemplo, -0,13, -0,083, -0,059 no exemplo acima)?

A árvore de decisão usa a medida de impureza (ou heterogeneidade) para selecionar as melhores divisões em cada filial. No exemplo específico de evasão de alunos, se um limite específico divide nossa amostra em puramente uma única classe (por exemplo, 10 desistentes vs 0 não desistentes), dizemos que as classes são puras ou homogêneas. No extremo oposto, se o limite específico divide a amostra em classes iguais (por exemplo, 5 desistências vs 5 não desistentes), dizemos que as classes são impuras ou heterogêneas.

Entropia ou o índice de Gini são as medidas populares de impureza que têm os valores 0 para classes puras e 1 para classes impuras. Em cada ramo, os algoritmos da árvore de decisão encontram um limite que minimizam medidas de impureza.

O modelo de floresta aleatória usa um grande número de árvores para tomar uma decisão. O modelo de floresta aleatória gera um conjunto de amostras usando bagging (*bootstrap aggregating*), o que significa dizer que as amostras e atributos são sorteados com repetição. Isso aumenta a probabilidade de o modelo ser mais confiável ao processar dados não vistos ou novos. O conjunto de dados de treinamento original por reamostragem com substituição treina as árvores de decisão usando cada uma das amostras e, em seguida, combina as previsões dessas múltiplas árvores de decisão para fazer uma previsão final. A estratégia típica para combinar várias decisões é uma votação em que a decisão da maioria torna-se a decisão final.

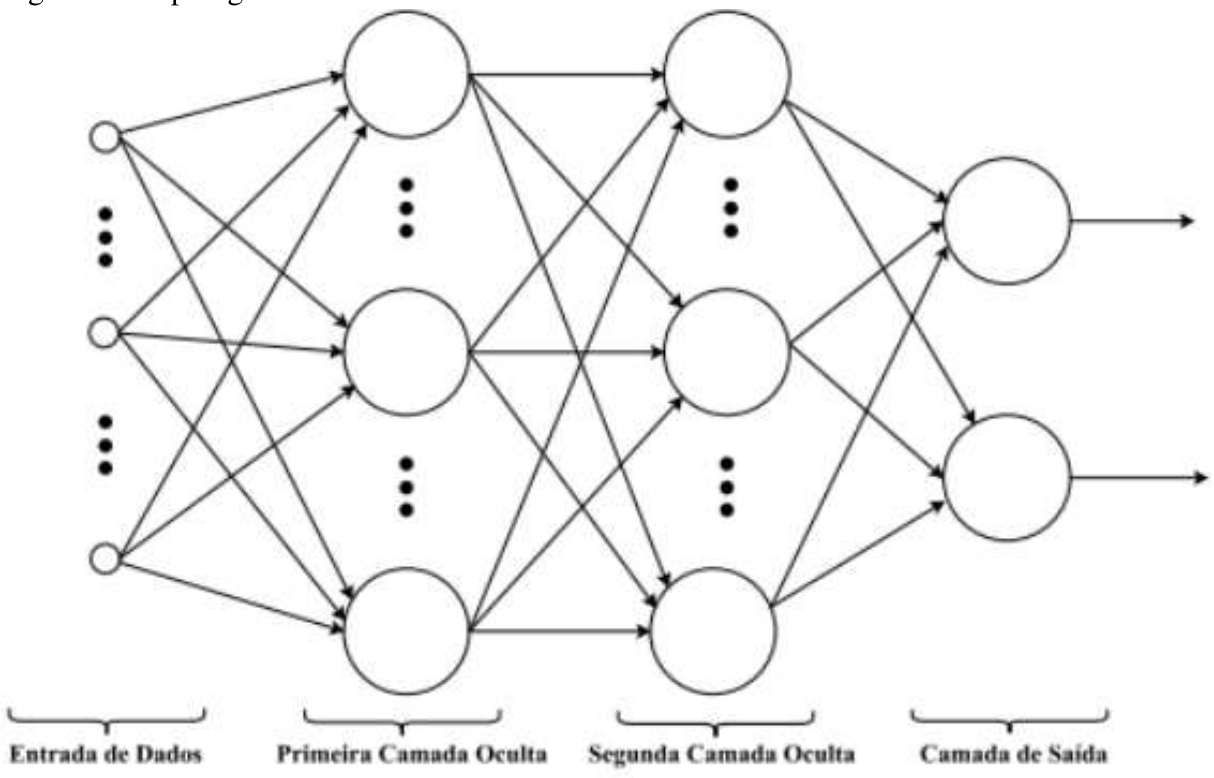
3.5.4.2 Perceptron multicamadas (Multilayer Perceptron - MLP)

As redes neurais artificiais do tipo *Multilayer Perceptron* (MLP) são compostas por um conjunto de unidades de entrada de dados que constituem uma camada de entrada, uma ou mais camadas intermediárias (ocultas) de neurônios, e uma camada de saída, também composta por neurônios. A motivação para utilização dessa rede é pelo fato de que elas aprendem tarefas complexas através das camadas intermediárias, sendo isso possível graças à extração progressiva de características significativas de padrões de entrada. (HAYKIN, 1999)

Redes como a percetron, que contém apenas uma camada, não conseguem resolver problemas que não são linearmente separáveis. Diferentemente, as MLP possuem uma estrutura composta por diversas camadas intermediárias de neurônios, sendo essa uma característica

importante para que possa ser feito a separação de elementos não separáveis por um hiperplano, ou seja, uma separação não-linear dos elementos. A Figura 4 mostra os processos realizados por uma MLP e suas camadas.

Figura 4 – Topologia de uma MLP



Fonte: Autor

A MLP utiliza o algoritmo *backpropragation* para treinamento da rede (LEUNG; HAYKIN, 1991). Este algoritmo é supervisionado e utiliza pares de entrada e saída (x, yd) para, por meio dos mecanismos de correção de erros, ajustar os pesos da rede. O treinamento ocorre em duas fases, cada uma percorre um sentido diferente da rede. Estas fases são forward e backward. A fase *forward* é utilizada para definir a saída da rede para um dado padrão de entrada. A fase *backward* utiliza a saída desejada e a saída fornecida pela rede para atualizar os pesos de suas conexões (BRAGA et al., 2000).

3.5.4.3 Máquina de Vetor de Suporte (Support Vector Machine - SVM)

O SVM é um método baseado na teoria de aprendizagem estatística e otimização matemática (VAPNIK, 1995). Deste modo, constitui um algoritmo supervisionado utilizado para a tarefa de classificação que utiliza um hiperplano como separador de classes (TAN et al. 2005). Esse hiperplano é obtido usando os vetores de suporte, ou conjunto de treinamento, e funciona

como um suporte para o limite da decisão ao classificar.

Com o SVM resolve-se tanto problemas de classificação quanto de regressão, envolvendo duas classes, mas pode ser estendido para problemas multi-classes. Foi proposto por Vapnick e é uma abordagem de aprendizagem supervisionada baseada na noção de kernel, ou mais especificamente, de funções denominadas *kernels*.

Entre as funções kernel mais utilizadas, incluem-se: polinomial (incluindo o kernel linear), os de função de base radial (*radial basis function* – RBF) e os sigmoidais. Cada um deles tem parâmetros em suas respectivas funções, tal como ilustrados na Tabela 3. Tais parâmetros precisam ser determinados pelos usuários. Em se tratando de kernels polinomiais, quando o parâmetro d assumir valor igual 1, tem-se um kernel linear.

Tabela 1 – Algumas funções do Kernel

Tipo de Kernel	**Função $K(x_i,x_j)$**	**Parâmetros**
Polinomial	$(\delta(x_i \cdot x_j)+\kappa)^d$	δ, κ e d
Gaussiano (RBF)	$\exp(-\sigma\|x_i - x_j\|^2)$	σ
Sigmoidal	$\tanh(\delta(x_i \cdot x_j)+\kappa)$	δ e κ

Fonte: Adaptado de (FACELI et al. 2011)

3.5.4.4 Seleção de atributos baseada em correlação (Correlation based feature selection - CFS)

Dentre as métricas mais utilizadas para avaliar um subconjunto de atributos, pode-se destacar as medidas de dependência e consistência, usadas pelo algoritmo *Correlation-based Feature Selection* (CFS) (LIU; YU, 2005).

O algoritmo CFS classifica os subconjuntos gerados de acordo com uma função de correlação com base em uma recompensa heurística de avaliação (HALL, 1999). Esse algoritmo avalia a importância de um subconjunto de atributos em função da predição individual de cada atributo e o grau de correlação entre eles.

O CFS seleciona um subconjunto de recursos de forma explícita. O algoritmo avalia cada recurso individualmente e atribui uma pontuação com base em sua importância (KOPRINSKA; RANA; AGELIDIS, 2015).

Depois de calcular uma matriz de correlação, o CFS aplica uma estratégia heurística de busca para encontrar um bom subconjunto de atributos de acordo com a seguinte eq.:

$$M(S) = \frac{k \times r_{ac}}{\sqrt{k + k(k-1)\, r_{aa}}} \qquad (1)$$

Em que $M(S)$ é o mérito de um subconjunto de atributos S contendo k atributos, r_{ac} é a média da correlação entre atributo-classe e r_{aa} é a média da correlação entre atributo-atributo.

3.6 Censo Escolar

A análise neste trabalho tem como objetivo alcançar estatísticas educacionais construídas com base nos microdados do Censo Escolar da Educação Básica longitudinais de 2012 a 2019, divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (BRASIL, 2014d). Desde as primeiras tentativas de construção de indicadores educacionais, os relacionados com o acesso, a permanência, a progressão e conclusão de crianças e jovens do sistema escolar têm ocupado um lugar central para o conhecimento e compreensão da dinâmica de funcionamento dos sistemas educativos, assim como para a detecção de problemas e dificuldades em seu interior (Córdova, 2008).

O próprio Inep define o Censo Escolar, o mais importante levantamento estatístico sobre o sistema educacional brasileiro, como “uma pesquisa que tem por objetivo realizar um amplo levantamento sobre as escolas de educação básica no País” (INEP, s.d.). Esse levantamento é feito anualmente com a finalidade de constituir-se em instrumento de planejamento, gestão e avaliação da política educacional brasileira, sendo realizado a partir de questionários respondidos pelas escolas. O banco de dados dos censos escolares é constituído pelo conjunto dos questionários preenchidos pelas escolas, com base nos registros escolares.

Há que se registrar ainda que o preenchimento destes questionários dependerá da compreensão subjetiva das pessoas encarregadas de, com base nos registros, prestar as informações solicitadas, havendo a possibilidade de inconsistências, como erros em idades ou demais respostas. Nesse sentido, de acordo com Ramos (2014), os estudos estatísticos são formas representativas de um dado contexto ou realidade, sendo impossível ao cientista dar conta de todas as possíveis causas e variáveis que compõem a realidade social como um fenômeno

multicausal.

Com todas as possíveis limitações, o Censo Escolar é a fonte estatística oficial do sistema escolar brasileiro, o que justifica o investimento em sua utilização neste estudo. Desse modo, a análise sobre a referida base de dados busca compreender um fenômeno específico, a ocorrência da permanência ou não do aluno ao longo dos anos, onde será verificado se ocorreu abandono, que corresponde à ausência do aluno no ano seguinte, ou evasão, que corresponde à ausência por dois anos seguidos.

3.7 SPAECE

A partir da obra de Vianna (2003) e Pequeno (2000), compreende-se que a constituição do processo de avaliação educacional no Ceará esteve em sintonia com uma preocupação nacional – na esfera federal pelo Ministério da Educação, e na esfera estadual pela SEDUC. Os autores supracitados destacam diversas pesquisas nessa área, na qual se observa que, muito lentamente, delinearam o que viria a ser o Sistema de Avaliação da Educação Básica (SAEB) em 1990 e, anos após, o Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE)

O SPAECE caracteriza-se como avaliação externa de larga escala e foi aplicado pela 1ª vez em 1994 em todas as escolas estaduais de Fortaleza com adesão voluntária das mesmas. As primeiras aplicações tiveram como base de definição dos conteúdos os Referenciais Curriculares Básicos (CRB/SEDUC), os Parâmetros Curriculares Nacionais (PCN's) e os manuais de apoio de ensino de jovens e adultos, livros do telensino. O objetivo principal era avaliar a qualidade do ensino oferecido nas escolas públicas do Estado do Ceará, mediante o desempenho dos alunos nas disciplinas Língua Portuguesa e Matemática, nos anos finais do ensino fundamental e médio, ressalta Pequeno (2000).

Por meio do SPAECE, são avaliados, anualmente, de forma censitária, o nível de leitura dos alunos do 2º ano do ensino fundamental e o domínio das competências e habilidades nas disciplinas de Português e Matemática, no 5º e 9º anos do ensino fundamental e no 3º ano do ensino médio. Esse sistema de avaliação tem por objetivo fornecer subsídios para a formulação, reformulação e monitoramento das políticas educacionais, e possibilita, aos professores, gestores escolares e dirigentes governamentais, um diagnóstico da educação básica da rede pública de ensino cearense (CEARÁ, 2012).

O desempenho escolar de qualidade implica, necessariamente, a realização dos objetivos curriculares de ensino propostos. Os padrões de desempenho estudantil, nesse sentido,

são balizadores dos diferentes graus de realização educacional alcançados pela escola. Por meio deles, é possível analisar a distância de aprendizagem entre o percentual de estudantes que se encontra nos níveis mais altos de desempenho e aqueles que estão nos níveis mais baixos.

Tendo em vista o objetivo dessa avaliação, os dados serão utilizados para apurar o desempenho dos alunos nas provas de português e matemática, e o nome da mãe do aluno para que possa ser feita a conexão com as outras bases de violência. Atualmente, os dados administrativos já foram cedidos pela Secretaria de Educação Básica do Ceará e complementarão a base de dados do CENSO.

O desempenho na prova do SPAECE será o indicador de desempenho do aluno em situação de evasão ou abandono. Para compreender esse indicador e os padrões de desempenho utilizados, o Tabela 2 mostra os valores correspondentes a cada faixa de performance do estudante.

Tabela 2 – Padrões de desempenho SPAECE

	5º ano		9º ano	
	Português	Matemática	Português	Matemática
Muito Crítico	<= 125	<= 150	<= 200	<= 225
Crítico	175	200	250	275
Intermediário	225	250	300	325
Adequado	> 225	> 250	> 300	> 325

Fonte: Autor

Segue a explanação, segundo o Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora (CAEd/UFJF), para cada classificação do desempenho de português e matemática do 5º ano.

Para o desempenho em língua portuguesa, classifica-se o estudante com nota abaixo de 125, como muito crítico, isso mostra que ainda não podem ser considerados leitores autônomos, pois necessitam, para isso, desenvolver habilidades que lhes permitam interagir de modo mais

eficaz com textos. Para o desempenho em matemática, são considerados, para essa mesma categoria, aqueles que obtiveram nota até 150, ou seja, evidencia-se que possuem conhecimentos elementares para este período de escolarização. O desafio que se coloca nesta fase é o de viabilizar condições para que os alunos possam encontrar significado para cada objeto matemático de seu estudo.

Estudantes com nota entre 125 e 175, são classificados como críticos. Nessa condição, o estudante começa a desenvolver um leque de habilidades que lhe permitirá avançar para um nível mais complexo de leitura. Para matemática, o aluno possui essa classificação, se a sua nota estiver entre 150 e 200. As habilidades matemáticas que mais se evidenciam são as relativas aos significados atribuídos aos números naturais, seja em um contexto social ou escolar.

Para notas entre 175 e 225, o estudante é classificado como intermediário. Aqui, o estudante já compreende o sentido de palavras ou expressões, o efeito do uso de pontuação e de situações de humor. Além disso, reconhecem o efeito de sentido de notações em um texto de linguagem mista. Em matemática, essa classificação é dada ao aluno que possui nota entre 200 e 250. Aqui há maior expansão do conhecimento matemático necessário à série, tanto no que tange à ampliação do leque de habilidades relativas à resolução de problemas quanto na complexidade que exige dos alunos melhor desempenho ao lidar com o Sistema de Numeração Decimal.

Para notas acima de 225, o estudante é classificado como adequado. Os estudantes que se localizam neste Padrão de Desempenho já desenvolveram habilidades para uma leitura autônoma. Em matemática, para notas acima de 250, os estudantes provam que desenvolveram as habilidades relativas ao campo Tratamento da informação nos padrões anteriores a este, demonstrando serem capazes de fazer leituras e interpretação de tabelas de até dupla entrada e gráficos de barras e setores.

Para alunos do 9º ano, as classificações dos padrões de desempenho são as seguintes. Para o desempenho em língua portuguesa, classifica-se o estudante com nota abaixo de 200, como muito crítico. Neste Padrão de Desempenho, os alunos se limitam a realizar operações básicas de leitura, interagindo apenas com textos do cotidiano, de estrutura simples e de temáticas que lhes são familiares. Para o desempenho em matemática, são considerados, para essa mesma categoria, aqueles que obtiveram nota até 225, evidenciam que possuem conhecimentos elementares para este período de escolarização. O desafio que se coloca nesta fase é o de viabilizar condições para que os alunos possam encontrar significado para cada objeto matemático de seu estudo.

Estudantes com nota entre 200 e 250 são classificados como críticos. Os estudantes

cujas médias de proficiência estão situadas neste Padrão de Desempenho ampliam suas habilidades de leitura, sendo capazes de interagir com textos de temática menos familiar e de estrutura um pouco mais complexa. Para matemática, o aluno possui essa classificação, se a sua nota estiver entre 225 e 275. Neste padrão, amplia-se o leque de habilidades relativas ao campo Numérico e o Algébrico começa a se desenvolver.

Para notas entre 250 e 300 o estudante é classificado como intermediário. As habilidades características deste Padrão de Desempenho revelam um avanço no desenvolvimento da competência leitora, pois os alunos demonstram ser capazes de realizar inferência de sentido de palavras/expressões em textos literários em prosa e verso, interpretando textos de linguagem mista. Em matemática, essa classificação é dada ao aluno que possui nota entre 275 e 325. As habilidades características deste Padrão de Desempenho evidenciam uma maior expansão dos campos Numérico e Geométrico. Os estudantes neste Padrão de Desempenho demonstram compreender o significado de números racionais em situações mais complexas, que exigem deles uma maior abstração em relação a esse conhecimento.

Para notas acima de 300, o estudante é classificado como adequado. Os estudantes nessa classificação podem ser considerados leitores proficientes, conseguem selecionar informações, levantar hipóteses, realizar inferências, autorregular sua leitura, corrigindo sua trajetória de leitura quando suas hipóteses não são confirmadas pelo texto. Em matemática, para notas acima de 325. Neste padrão, os estudantes demonstram resolver problemas envolvendo equação do 2° grau e sistema de equações do 1° grau. No nível avançado da escala, os estudantes utilizam o raciocínio matemático de forma mais complexa, conseguindo identificar e relacionar os dados apresentados em diferentes gráficos e tabelas para resolver problemas ou fazer inferências.

3.8 Base de Dados de Violência

CVLI, nomenclatura criada em 2006 pela Secretaria Nacional de Segurança Pública (SENASP), tem o objetivo de agregar todos os atos delituosos praticados com o emprego da violência vitimando o indivíduo sem proporcionar-lhe qualquer meio de resistência para salvaguardar sua própria vida.

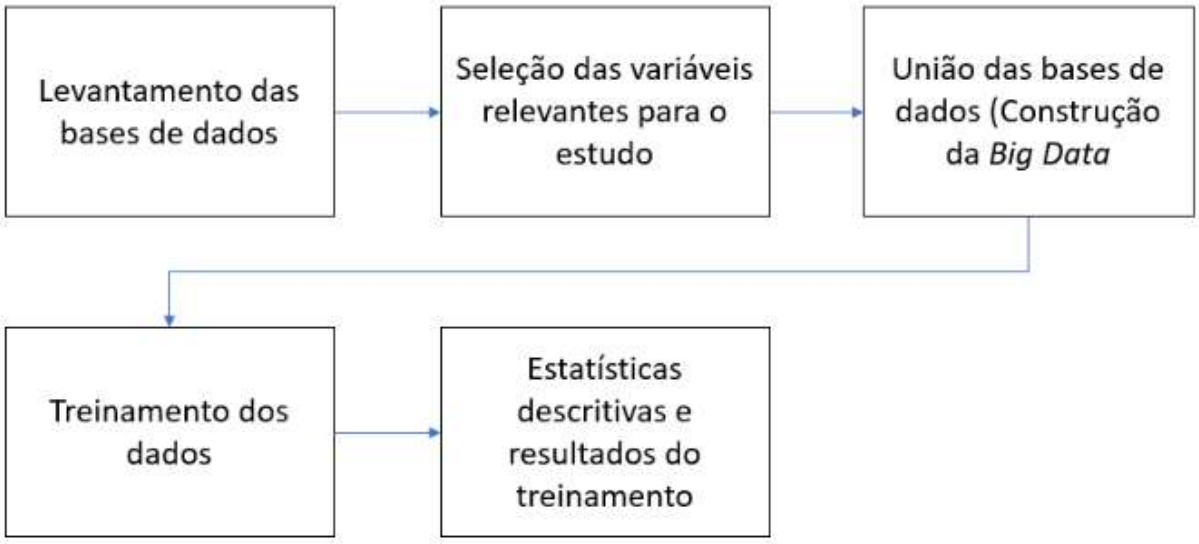
Os crimes que agregam a nomenclatura citada acima são descritos como os crimes de homicídio doloso art.121 (matar alguém), Latrocínio (roubo seguido de morte art.127), e a Lesão Corporal Seguida de Morte art.159, todos descritos no código penal brasileiro. Esses crimes também são denominados de crimes hediondos, tipificados na lei 8.072/90, conhecida como a

Lei dos Crimes Hediondos, modalidade criminosa praticada contra o ser humano que significa algo causador de sensação de impunidade, pavor social ou repúdio nos órgãos responsáveis pela segurança pública tanto do país quanto do estado.

4 MATERIAIS E MÉTODOS

Nesta seção serão apresentados os dados que foram utilizados no estudo, bem como as métricas, métodos e ferramentas. Para melhor visualizar os processos realizados, descritos na sequência do texto, o fluxograma do método KDD usado na presente pesquisa é apresentado na figura 5.

Figura 5 – Fluxograma metodológico do KDD



Fonte: Autor

Diversas ferramentas para minerar dados e análises encontram-se disponíveis na literatura. Citam-se algumas delas: Weka, Microsoft Excel, SPSS®, SQL, SAS OnDemand. Para esse estudo, foi utilizado o software SPSS para realizar a leitura e aplicação dos filtros nos dados, e o software e sistema gerenciador de banco de dados PostgreSQL para uma melhor manipulação desses.

4.1 Levantamento das bases de dados

Tendo como foco do estudo o público alvo da educação fundamental de 5º e 9º ano, investigou-se o fluxo escolar desses alunos no estado do CE. Em particular, analisou-se a permanência do aluno, verificando os índices de abandono e evasão. O abandono ocorre quando o aluno deixa de frequentar as aulas durante o ano letivo. Já a evasão escolar diz respeito à situação do aluno que abandonou a escola ou reprovou em determinado ano letivo, e que no ano seguinte não efetuou a matrícula para dar continuidade aos estudos.

Dentro deste contexto, os dados do estudo foram obtidos através do censo escolar,

disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), da análise dos dados de desempenho na prova do SPAECE e da verificação do nome da mãe desse estudante nas bases de dados de violência doméstica e de CVLI.

Os dados obtidos do INEP para verificação da presença do aluno ao longo dos anos foram apurados através dos dados do censo escolar. A base de microdados do censo escolar é composta por quatro tabelas: escola, turma, matrícula e docente.

Cada tabela é composta por uma série de variáveis com intuitos de obter informações específicas sobre seu eixo. Dessas tabelas a que possui os dados dos alunos é a tabela matricula que é composta por 85 variáveis distribuídas em três blocos de variáveis: Dados do aluno, dados da turma, dados da escola. A variável utilizada para verificar a presença dos alunos ao longo dos anos foi a variável PK_COD_ALUNO, que corresponde ao código do aluno na base de dados do INEP. A lista completa das variáveis pode ser conferida na seção Anexos.

4.2 Seleção das variáveis

Seguindo os processos do KDD dispostos na Figura 5 e por meio de um entendimento bem definido do objetivo, inicialmente foi necessário selecionar as variáveis, bem como os dados, que serão usados nesse processo. Os procedimentos serão explicitados nas subseções seguintes.

4.2.0.1 Seleção de dados da base CENSO

Após o download das bases de dados do censo dos anos de 2012 a 2020 e identificação do arquivo MATRICULA_NORDESTE, foram aplicados filtros para selecionar os alunos do estado do Ceará, esses procedimentos foram realizados através do software SPSS em cada um dos anos de aplicação. Após aplicação dos filtros foram selecionadas as variáveis úteis ao trabalho, mostradas na tabela 3.

A Tabela 3 mostra as variáveis utilizadas para o estudo que foram selecionadas de uma base com 113 variáveis (ANEXO A), essas variáveis foram filtradas de acordo com os propósitos do estudo, no que concerne em unir demais bases de dados através do código do aluno e da verificação das variáveis que possuam informações sobre seu desempenho em avaliações e demais dados das mães dos alunos provindas das outras bases. Foram eliminadas as variáveis que correspondem a localidade e demais códigos de localização que não agregam no resultado da pesquisa.

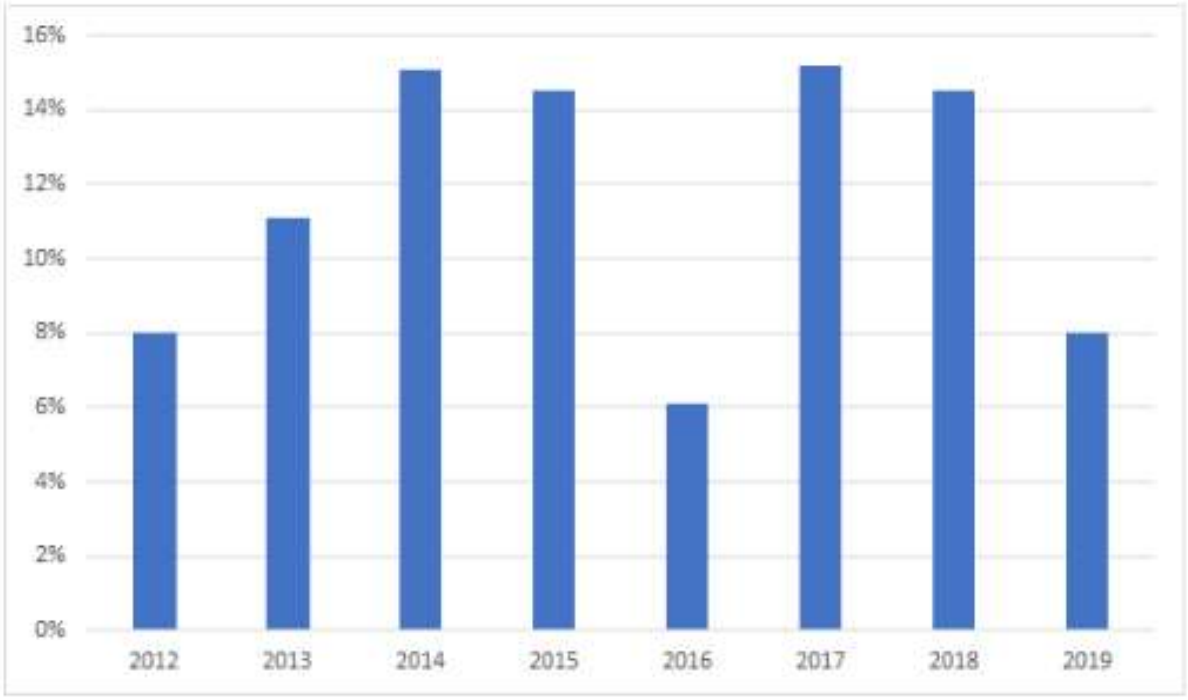
Tabela 3 – Variáveis CENSO

Nome da Variável	Descrição da Variável	Tipo	Categoria
ANO_CENSO	Ano do Censo	Num	
FK_COD_ALUNO	Código do aluno (ID_INEP)	Num	
TP_SEXO	Sexo	Char	M - Masculino
			F - Feminino
TP_COR_RACA	Cor/raça	Char	0 - Não declarada
			1 - Branca
			2 - Preta
			3 - Parda
			4 - Amarela
			5 - Indígena
PK_COD_ENTIDADE	Código da escola	Num	
COD_MUNICIPIO ESCOLA	Código do município da escola	Num	

Fonte: Autor

Em seguida, foi realizado todo o pré-processamento, uma vez que, frequentemente, os dados são encontrados com inúmeras inconsistências. Essas tarefas são fundamentais, pois o objetivo é eliminar incongruências como dados duplicados. Em particular, idades inconsistentes (mais de 100 anos), variáveis decimais fora dos padrões (decimais inconsistentes) e campos nulos que poderão gerar inconsistência em análises estatísticas e aplicações de algoritmos de treinamento de máquina. Nesta fase foram retirados os dados duplicados.

Figura 6 – Percentual de dados duplicados por ano de Censo



Fonte: Autor.

A Figura 6 mostra os percentuais de dados duplicados encontrados ao longo dos anos, vale ressaltar que muitos registros constam como duplicados pelo fato do aluno ser transferido de uma escola para outra, com isso gera uma duplicata, um registro em cada escola em um ano.

Posteriormente, foi realizada a transformação que consiste na aplicação de normalização, agregação, criação de novos atributos ou seleção de um conjunto específico de dados. Nessa etapa os microdados de cada ano foram repassados do software SPSS para o sistema gerenciador de banco de dados (SGBD) software PostgreSQL, sistema de código aberto, disponibilizado gratuitamente e amplamente utilizado por desenvolvedores de sistemas (POSTGRESQL, 2007).

4.2.1 Seleção de dados da base SPAECE

A base de dados do SPAECE foi disponibilizada pela Secretaria da Educação Básica do Estado do Ceará. O intuito da utilização da base de dados é para apurar o desempenho dos alunos na prova SPAECE (seção 4.5), identificar a etapa de ensino (5º ou 9º), o nome do aluno e de sua mãe. A base de dados é composta por microdados dos desempenhos de várias etapas de ensino, para este estudo serão considerados dois arquivos de cada etapa de ensino (5º e 9º ano), cada arquivo com informações sobre o desempenho em língua portuguesa e de matemática para as respectivas etapas de ensino dos anos de 2012 a 2018.

Cada arquivo contém 36 variáveis e após selecionar as variáveis de interesse restaram as variáveis contidas na Tabela 4.

Tabela 4 – Variáveis SPAECE

Nome da Variável	Descrição da Variável	Tipo	Categoria
CD_ALUNO	Código do aluno (ID_INEP)	Num	—
CD_ESCOLA	Código da escola	Num	—
CD_ETAPA	Etapa de ensino	Num	4 e 15: 2º Ano do EF 7 e 18: 5º Ano do EF 11 e 41: 9º Ano do EF 25, 30 e 35: 1ª Série do EM 26, 31 e 36: 2ª Série do EM 27, 32 e 37: 3ª Série do EM
NO_ALUNO	Nome do aluno	Char	—
NO_MAE	Nome da mãe	Char	—
VL_PROFICIENCIA	Desempenho do aluno em Português ou Matemática	Double	000.00 - 500.00

Fonte: Autor

Para um melhor manuseio das funções do banco de dados, os arquivos foram renomeados para SPAECE_MAT e SPAECE_POR. A variável VL_PROFICIENCIA foi separada em duas variáveis; d_mat e d_por, para o recebimento do desempenho em matemática e português, respectivamente.

4.2.2 Seleção de dados das bases CVLI e Maria da Penha

As bases de dados de violência foram disponibilizadas pela Secretaria de Segurança Pública e Defesa Social – SSPDS/CE. A utilização dessas bases de dados finda o conjunto dos dados utilizados para o estudo. O objetivo é a verificação da situação da mãe do aluno, se ela está contida na base de dados CVLI ou Maria da Penha, ou o próprio aluno.

Ressalta-se o objetivo inicial que é a verificação do perfil dos alunos que estão em situação de evasão ou abandono e que possuem a mãe em bases de dados de violência. As bases de dados de violência são compostas por variáveis como idade, nome da vítima, sexo e informações sobre como o crime ocorreu e que tipo de arma foi utilizada. Para o estudo foi considerada a variável nome da vítima uma vez que o intuito é a verificação do nome da mãe do aluno na base da violência. O conjunto de dados foi repassado também para SGBD *Postgres*.

4.3 União de dados

Nesta subseção será apresentado como foi realizada a união entre as bases de dados o que culmina na base de dados para a realização do treinamento dos dados.

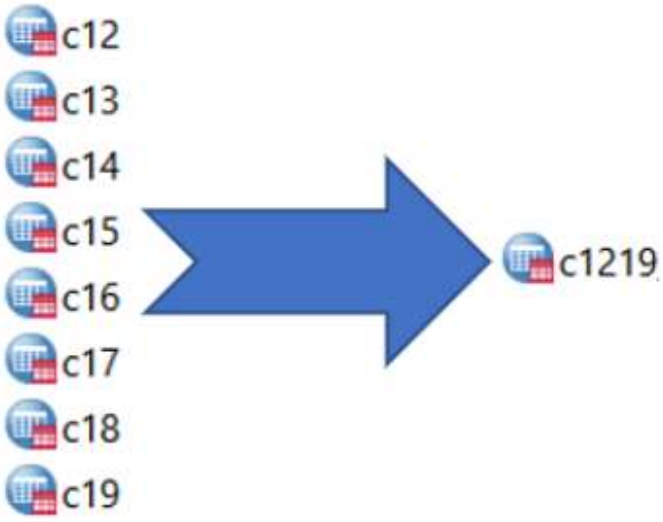
Após a verificação da evasão ou abandono do aluno, e a mescla dos dados de seu desempenho na prova do SPAECE, foi realizada a adição das informações da presença da mãe em bases de dados de violência. A união das bases de dados gerou um conjunto de informações possíveis de serem analisadas e verificadas os padrões dos alunos que se encontram em situação de evasão ou abandono através dos algoritmos aplicados.

4.3.1 União dos dados do Censo

Após a execução dos passos de seleção descritos na seção 5.3.2, os arquivos foram unificados para o arquivo c1219. A quantidade de registros de 2012 a 2019 corresponde a 18.171.454 registros, após exclusão de duplicatas e incongruências a base de dados foi concluída com o total de 4.236.352 de registros.

Após adição dos microdados dos anos de 2012 a 2019 no SGBD foi adicionada uma nova variável chamada situação, onde ela tem por objetivo classificar o aluno em 1 que corresponde à situação de abandono, 2 para a situação de evasão e 0 caso o aluno tenha concluído os estudos sem passar por nenhuma situação. A variável FK_COD_ALUNO foi utilizada como parâmetro para verificar a presença ou não do aluno ao longo dos anos tornando possível a classificação do aluno em abandono ou evasão.

Figura 7 – União de dados do censo escolar



Fonte: Autor.

A Figura 7 explicita a união das bases de dados longitudinais dos anos de 2012 a 2019 (c12 - c19) após os procedimentos de pré-processamento e seleção de variáveis culminando em uma única base c1219. Desta forma, conclui-se os procedimentos com a base de dados censo, onde o intuito foi a verificação da situação estudantil do aluno ao longo dos anos e posteriormente a criação de uma base única que facilitará à adição de novas variáveis provindas da base de dados SPAECE, Maria da penha e CVLI.

4.3.2 União dos dados SPAECE com CENSO

Os registros da base SPAECE foram adicionados na base de dados c1219 utilizando como chave a variável FK_COD_ALUNO da base c1219 e CD_ALUNO do SPAECE. A seguinte instrução presente na figura 8 foi aplicada.

Figura 8 – Instrução de união da base Censo com Spaece

```
Atualize c1219
Defina d_mat = a.VL_PROFICIENCIA
a partir (
     Selecione cd_aluno, VL_PROFICIENCIA a partir de
     SPAECE_MAT
) é a
onde a.cd_aluno = c1219.fk_cod_aluno;
```

Fonte: Autor.

A instrução acima realiza o incremento do desempenho do aluno encontrado nos registros da base SPAECE_MAT e o adiciona na variável d_mat em c1219. A mesma instrução foi utilizada para adicionar as demais variáveis da Tabela 5 na base c1219.

Ressalta-se que a variável NO_MAE é a variável chave para união com a base de dados de violência, CVLI e Maria da Penha descritas na próxima seção.

4.3.3 União dos dados Maria da Penha e CVLI

Na tabela de banco de dados correspondente ao Censo com as variáveis do SPAECE, c1219, foram adicionadas mais 3 variáveis:

- s_mpenha
- s_cvli
- s_aluno

As variáveis foram preenchidas com valores 0 caso o registro do nome da mãe do

aluno não seja encontrado em uma das bases de violência e 1, caso o registro seja encontrado. A instrução na Figura 9 foi executada para verificação do pertencimento do nome da mãe na base de dados Maria da Penha.

Figura 9 – Instrução de verificação na base Maria da Penha

```
Atualize c1219 a
Defina s_mpenha = 1
onde existe (
     selecione 1 a partir de maria_da_penha b
     onde a.no_mae = b.nome_da_vitima
     );
```

Fonte: Autor.

Para verificação do não pertencimento do nome da mãe é adicionado o parâmetro not na condição where, conforme mostra a Figura 10.

Figura 10 – Instrução de verificação na base CVLI

```
Atualize c1219 a
Defina s_cvli = 0
Onde não existe (
     selecione 1 a partir de cvli b
     onde a.no_mae = b.nome_da_vitima
     );
```

Fonte: Autor.

Além da verificação do nome da mãe do aluno, o motivo da evasão ou abandono do aluno pode ter ocorrido pela presença do próprio aluno em uma das bases de dados de violência, com isso foi realizada a busca pelo nome do aluno conforme instrução presente na Figura 11.

Figura 11 – Instrução de verificação do aluno na base CVLI

```
Atualize c1219 a
Defina s_aluno = 1
Onde existe (
     Seleccione 1 a partir de cvli b
     Onde a.no_aluno = b.nome_da_vitima
     );
```

Fonte: Autor.

Após a conclusão desse passo a base de dados c1219 está completa. A tabela possui

as seguintes informações após união de informações da base de dados censo, SPAECE, Maria da Penha e CVLI.

- fk_cod_aluno: Variável correspondente ao número do aluno no INEP, identificador único que foi utilizado para verificar a permanência do aluno ao longo dos anos.
- num_idade: Idade do aluno no momento que foi aplicado o censo.
- tp_sexo: Sexo do aluno (Masculino ou Feminino).
- tp_cor_raca: Raça do aluno.
- situacao: Variável que designa se o aluno está em situação de abandono (preenchido com 1), evasão (preenchido com 2) ou concluiu (preenchida com 0, indica que o aluno concluiu os estudos) etapa: Corresponde à etapa de ensino, 5º ou 9º ano.
- nm_mae: Nome da mãe do aluno, variável que foi adicionada pela base de dados SPAECE e que foi utilizada para buscar o nome da mesma em bases de violência.
- d_mat: Desempenho do aluno em matemática na prova do SPAECE
- d_por: Desempenho do aluno em língua portuguesa na prova do SPAECE
- nm_aluno: Nome do aluno, variável que foi utilizada para buscar a ocorrência do nome do aluno na base de dados de violência.
- m_penha: Situação da mãe do aluno na referida base de dados. Preenchida com 0 caso tenha encontrado o registro, 1 caso não.
- cvli: Situação da mãe do aluno na referida base de dados. Preenchida com 0 caso tenha encontrado o registro, 1 caso não.
- s_aluno: Situação do aluno na base de dados cvli. Preenchida com 0 caso tenha encontrado o registro, 1 caso não.

4.4 Treinamento dos dados

Para o treinamento dos dados foi utilizado classificação para predição dos riscos e evasão dos estudantes que possuem as mães com registros em bases de violência e as estratégias aplicadas para validar os processos de classificação.

Para o treinamento dos dados utilizou-se a ferramenta Weka. Weka é um software de código aberto que fornece uma coleção de algoritmos de aprendizagem de máquina para diversas tarefas de mineração de dados, além de ferramentas para pré-processamento dos dados, classificação, regressão, agrupamentos, regras de associação e visualização de informações. Os experimentos foram realizados com os algoritmos de classificação da árvore aleatória, multilayer

perceptron (MPL) e support vector machine (SVM) pontuados em 4.5.3.

Sobre a seleção de variáveis, a primeira seleção utilizada ocorreu através do processo do KDD com o intuito de excluir de forma subjetiva os atributos que não são relevantes para o estudo, como endereço, códigos de escola, município etc.

Após exclusão dessas variáveis foi possível à realização de melhores análises com um tamanho reduzido de informações, porém, para a etapa de testes e treinamento dos dados, além da exclusão de variáveis que já ocorreu no KDD é necessária a escolha das melhores variáveis para a realização dessas ações. O algoritmo de avaliação de atributos utilizado foi o CFS, que foi aplicado dentro do software Weka.

4.5 Métricas de desempenho

A última etapa do método KDD consiste na obtenção das estatísticas descritivas e resultados do treinamento, que serão mostrados na próxima seção. Nesta seção, apresentamos algumas das métricas de desempenho usadas neste trabalho.

O desempenho de um modelo preditivo precisa ser avaliado com base em vários desempenhos métricos. O problema de prever a evasão de alunos pode ser classificado , como já visto na seção anterior, como uma classificação no aprendizado de máquina. Neste estudo, serão utilizadas as seguintes métricas de desempenho para classificação para avaliar o modelo treinado: precisão, sensibilidade, especificidade. A precisão, sensibilidade e especificidade são baseadas na matriz de confusão.

O problema de classificação é voltado em prever uma classe usando vários preditores. No aprendizado de máquina, os preditores são normalmente chamados de features ou características. A matriz de confusão visualiza o desempenho dos modelos para a classificação usando uma tabela cruzada (*cross-table*).

A matriz de confusão nos permite realizar uma análise mais detalhada da situação do nosso classificador uma vez que ela distingue nossos resultados em quatro classes. A seguir um exemplo com um classificador binário.

A matriz de confusão na classificação binária é a tabela cruzada 2 por 2 em que o positivo e negativo reais estão na coluna, e o positivo e o negativo previstos estão na linha, conforme representado na Tabela 5.

Na matriz de confusão, o VP, VN, FP, FN são, respectivamente, os casos em que as classes reais e previstas são ambas positivas, as classes reais e previstas são negativas, as classes

Tabela 5 – A Matriz de Confusão

		Real	
		Positivo	Negativo
Previsão	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: Autor

reais são negativas, mas as classes previstas são positivas, e os casos em que as classes reais são positivas, mas as classes previstas são negativas.

A matriz de confusão nos permite avaliar inicialmente o desempenho de um modelo preditivo, observando a tabela e, além disso, é a base para o cálculo de outras métricas de desempenho para classificação binária. A acurácia é a proporção correta da acurácia sobre o número total de acurácias feitas pelo modelo, e é definido como (VP + VN) / (VP + FP + FN + VN) na matriz de confusão. A sensibilidade (ou taxa positiva verdadeira) é a proporção daqueles previstos como positivos (Taxa de Abandono) entre os verdadeiros positivos, e é definido como VP / (VP + FN) na matriz de confusão. A especificidade (ou taxa negativa verdadeira) é a proporção daqueles previstos como negativos (ou seja, sem abandono no nosso caso) entre os verdadeiros negativos, e é definido como VN / (VN + FP) na matriz de confusão.

5 RESULTADOS OBTIDOS

Este trabalho abrange a união de 4 bases de dados: CENSO, SPAECE, violência doméstica e CVLI. A união dessas bases resultou em uma única base de dados de Evasão e Abandono contendo variáveis de desempenho e dados das mães dos alunos, sendo possível fazer uma análise das características do aluno com caráter de evasão e abandono através de estatísticas do seu desempenho.

Neste capítulo serão apresentadas análises e resultados provenientes da metodologia proposta. O capítulo é dividido em duas subseções que correspondem às estatísticas descritivas, onde são apresentados resultados de análises estatísticas e gráficos sobre o perfil dos alunos em determinada situação, e aos resultados dos treinamentos dos dados do sistema de predição, onde são apresentados os resultados dos testes realizados, hiperparâmetros utilizados e melhores acurácias dos treinamentos dos dados de cada um dos modelos preditivos.

5.1 Análises de desempenho dos alunos

Após a conclusão do processo de união das bases de dados utilizadas no presente estudo, algumas análises foram possíveis de serem realizadas. Nas tabelas e figuras a seguir são mostrados os resultados das análises estatísticas realizadas.

Tabela 6 – Situação do desempenho

Situação	Etapa	Matemática	Português
Abandono	5° ano	187.41	154.04
	9° ano	200.30	188.35
Evasão	5° ano	171.96	154.74
	9° ano	195.57	189.38
Concluiu	5° ano	229.90	168.94
	9° ano	222.19	213.64

Fonte: Autor

Na Tabela 6 são apresentadas as médias dos alunos por etapa e situação de ensino. Pode-se observar o nível de proficiência de cada categoria dos alunos de acordo com os padrões

de desempenho apresentados na Tabela 4.7. Observa-se que nas etapas de ensino de 5º e 9º ano, os melhores índices estão na categoria dos alunos que concluíram os estudos. Dentre aquelas que estão em situações de evasão e abandono, os alunos do 5º e 9º ano que evadiram obtiveram melhores resultados na categoria matemática, e em português, os alunos de 5º e 9º ano que abandonaram obtiveram melhores resultados.

Comparando as classes, é observada uma grande diferença dos alunos que concluíram para os alunos em situação de evasão e abandono, em especial na etapa de ensino 5º ano. De fato, a diferença das notas dos alunos que concluíram para os que evadiram ou abandonaram é maior para os alunos do 5º ano do que para os alunos do 9º ano. Entre as classes de abandono e evasão, observa-se uma diferença na etapa de ensino 5º ano na disciplina matemática, onde abandono possui maior média, já para português a média é praticamente igual. Conclui-se, a partir da Tabela 6, o maior rendimento dos alunos, sobretudo do 5º ano, que não enfrentaram problemas a ponto de evadir ou abandonar os estudos, tanto em português quanto em matemática. E para os os alunos que deixaram, em algum momento ou de forma definitiva os estudos, é observado que as médias da categoria abandono, é melhor que as dos alunos em situação de evasão em 3 dos 4 casos.

Tabela 7 – Média de idades por situação

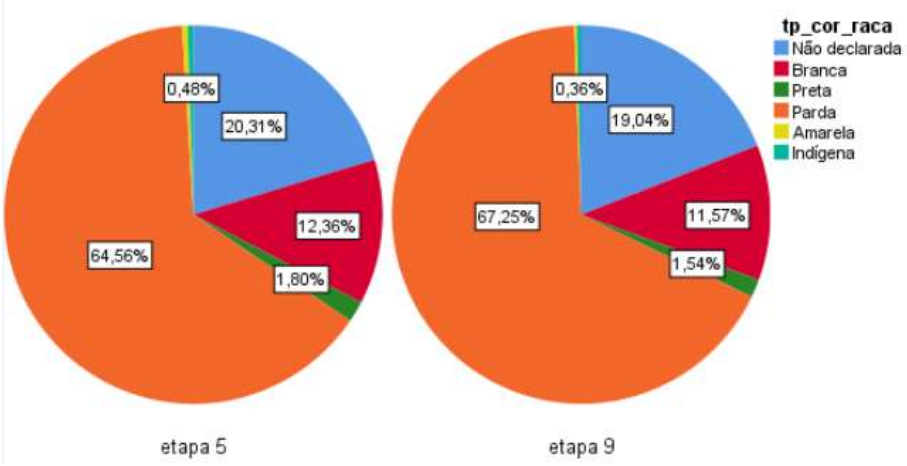
Situação / Sexo	5º Ano			9º Ano		
	Mas	Fem	Geral	Mas	Fem	Geral
Evasão	11.2	12.2	11.7	16.3	16.5	16,4
Abandono	13.2	12.8	15	17.3	16.9	17,1
Concluiu	17.3	18.1	17,7	18.1	18.2	18,1

Fonte: Autor

Na Tabela 7 são apresentadas as médias de idades dos alunos em cada situação, por etapa de ensino para o sexo feminino e masculino. Pode-se observar que a maior média de idade corresponde aos alunos que são categorizados como concludentes, o que faz total sentido uma vez que correspondem a alunos que foram até o final de sua jornada estudantil. Para os demais alunos, ou seja, em situações de evasão e abandono, as categorias possuem respectivamente 11.7 e 15 anos de média de idade. Essa informação complementa o perfil do aluno que será descrito logo mais. Fazendo um comparativo das idades das categorias de evasão e abandono, pode-se

observar que existe uma maior média de idade na situação abandono, para ambos os sexos, tanto no 5º quanto no 9º ano. Isto faz sentido uma vez que são alunos que mais se aproximam das características de alunos concludentes, pois saíram mas retornaram após um ano para o sistema de ensino, diferente dos alunos em situação de evasão que não retornaram. Verificando as categorias de evasão e abandono, pelo sexo, o masculino possuí maior média na categoria abandono, e o sexo feminino possuí maior média na situação evasão. Conclui-se a partir da Tabela 7 a predominância na média das idades dos alunos concludentes, e observa-se dentre as 3 classes, que o sexo feminino possuí maiores médias de idade dentre as 3 categorias.

Figura 12 – Categorias da Etnia para situação de Evasão

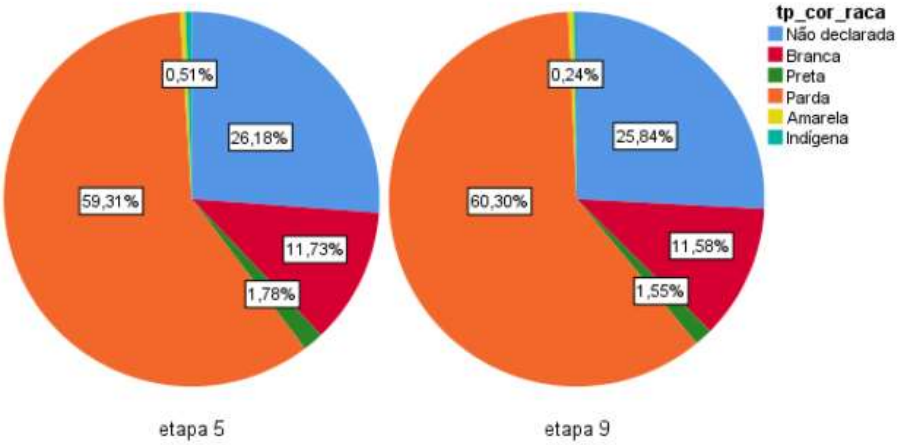


Fonte: Autor.

Na Figuras 12, 13 e 14 são apresentadas as categorias das etnias para os alunos em situação de evasão, abandono e conclusão, respectivamente. Essas informações são importantes para complementar o perfil dos alunos que se encontram em cada situação. Conclui-se a predominância da etnia parda nas 3 situações e nas duas etapas, 5º e 9º ano, respectivamente, seguidos de uma quantidade considerável de cor não declarada, fato esse que é observado nas 3 classes. Entre as classes de abandono e evasão é observado um aumento da não declaração de cor na categoria abandono, onde chega a ocupar 25% do total das amostras. Dentre as classes, os alunos concludentes possuem as maiores taxas de alunos de etnia parda, com 67,15% e 68% para 5º e 9º ano, respectivamente. Na situação de conclusão observa-se também um aumento na autodeclaração de cor, onde a etnia branca tem quantidade superior ao de cor não declarada.

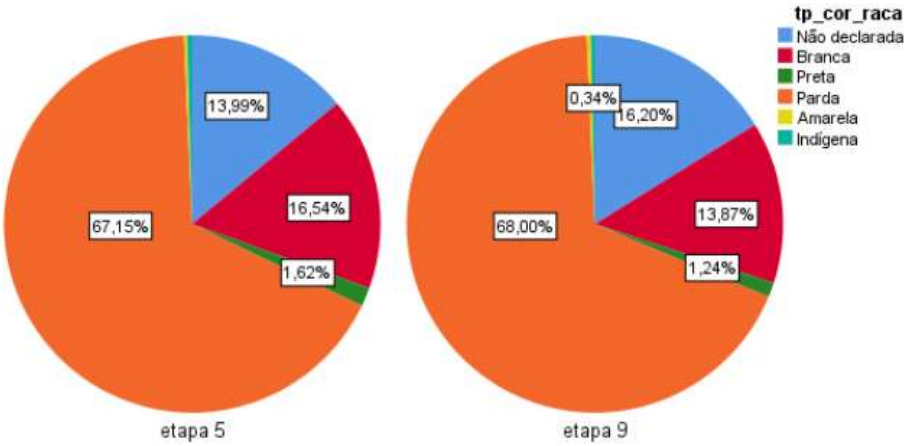
Na Figura 15 é apresentada à distribuição do quantitativo de alunos, classificados pelo sexo, etapa de ensino e situação. Observa-se no 5º ano, para os alunos em situação de abandono e evasão, que a predominância é do sexo masculino, e para os alunos concludentes, predomina-se os alunos do sexo feminino. Para a etapa de ensino do 9º ano, os alunos em

Figura 13 – Categorias da Etnia para situação de Abandono



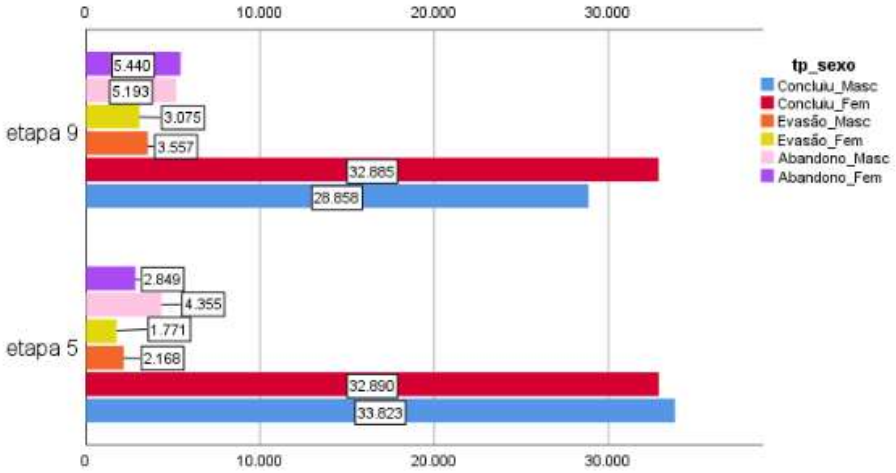
Fonte: Autor.

Figura 14 – Categorias da Etnia para situação de Concluiu



Fonte: Autor.

Figura 15 – Distribuição de sexo por etapa de ensino e situação dos alunos

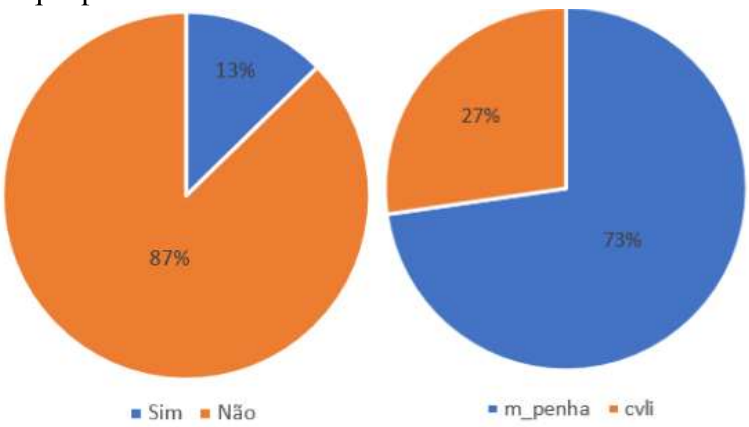


Fonte: Autor.

situação de abandono, predomina-se o sexo feminino. Para os alunos em situação de evasão, o sexo masculino possuí maior número. Para aqueles que estão categorizados como concludentes

o sexo feminino predomina. Verificando as duas etapas de ensino, conclui-se que no 9 ano existe uma ligeira diferença na quantidade do sexo feminino que abandonou ou evadiu, na para a etapa do 5 ano a maior taxa de abandono e evasão em todos os casos é feminina.

Figura 16 – Alunos que possuem mães em base de violência



Fonte: Autor.

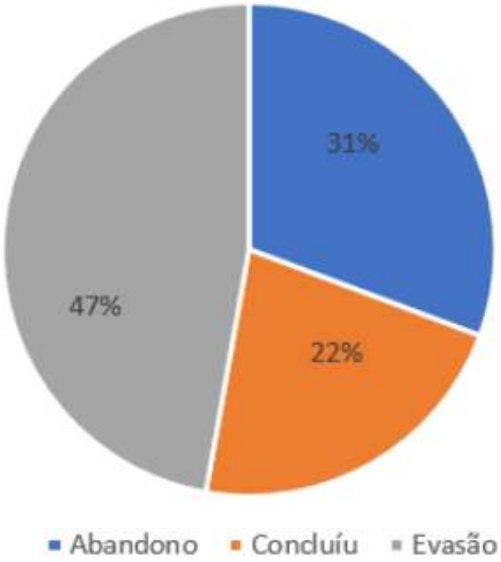
A Figura 16 mostra, dentre os alunos cujas mães estão nas bases de violência, a divisão percentual do quantitativo de alunos com mães contidas em bases de violência, observa-se que em 13% dos dados foram localizados nas bases de dados de violência mães acometidas de alguma violência, enquanto que 87% dos demais alunos não foram encontrados registros nas bases de violência.

A Figura mostra ainda a divisão desses 13% de registros de mães à base de violência categorizados em cvli e maria da penha. Observa-se que deste percentual, 73% dos registros de violência estão centrados na base de dados maria da penha, representado pela variável m_penha e 27% correspondem a registros na base de dados de cvli.

Complementando a Figura 16, a Figura 17 mostra a divisão desse percentual para as classes de evasão, abandono e dos alunos concludentes. Observa-se que a grande quantidade de registros estão localizados na base de evasão e abandono, com maior predominância nos alunos em situação de evasão. Pode-se concluir que os alunos em situação de evasão, que deixaram a escola por mais de um ano e não retornaram mais para o sistema estudantil, são os que mais possuem registros de mães em bases de violência, quase metade dos registros da base de dados, com 47%. Na classe Abandono, 31% dos casos possuem mães em registros de violência, e dos alunos que concluíram os estudos, 22% possuem mães em bases de violência.

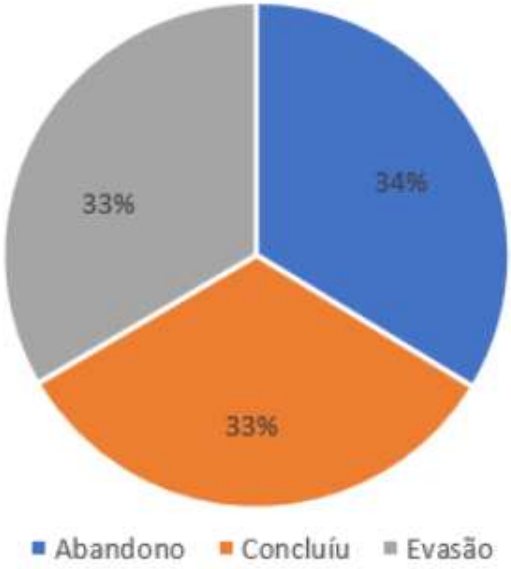
A Figura 18 mostra, dentre os alunos que não estão com mães em registros em bases de violência, a divisão entre as categorias abandono, evasão e conclusão. Pode-se ver que existem

Figura 17 – Percentual de casos de violência por classe



Fonte: Autor.

Figura 18 – Percentual de casos de violência por classe



Fonte: Autor.

33% dos alunos em situação de abandono e que não possuem mães nos registros de violência, seguidos de 33% em situação de evasão e 33% para os alunos que concluíram. Conclui-se que os outros 2/3 de cada base estão em uma das categorias de alunos que abandonaram, evadiram ou concluíram que estão presentes na base de violência mostradas nas figuras anteriores. Além disso, comparando a informação da Figura 18 com a Figura 17 pode-se observar, em especial na situação dos alunos que concluíram, um menor percentual na taxa de alunos na base de dados de violência, presentes na Figura 17, já a Figura 18 apresenta um maior número de alunos que concluíram, com 33%, e que não possuem registros em base de violência, o que faz sentido.

Para situação de abandono, a taxa é de 34%, ligeiramente superior à quantidade de alunos que possuem registros em base de violência. Na situação de evasão, ocorre um decréscimo, o que faz sentido por conta que o grande índice de alunos em situação de evasão está presente na base de dados de violência, na Figura 17 com 47%, já para os alunos que estão em situação de evasão mas sem registros em base de violência a taxa é de 33%.

Verificando as informações dos alunos de acordo com os padrões de desempenho apresentados e as informações apresentadas sobre etnia, sexo e idade dos alunos, sobre podemos concluir o seguinte:

- Os alunos do 5º ano em situação de abandono demonstram conhecimento de nível crítico, em matemática e português. Prevalece a etnia pardo, sexo masculino e com média de idade de 15 anos. Já os alunos do 9º ano apresentam para matemática e português os níveis muito críticos, caracterizados em grande maioria como pardo, do sexo masculino e com média de idade 16 anos
- Para os alunos do 5º ano em situação de evasão, eles apresentam níveis críticos em português e matemática. São em grande maioria pardos e do sexo masculino com média de idade de 11.7 anos. Os alunos do 9º ano, em ambas disciplinas, estão categorizados como muito críticos. O público, em grande parte, é masculino, pardo e com média de 16 anos.
- Para os alunos do 5º ano que concluíram os estudos sem passar por situações de evasão ou abandono, em matemática apresentam nível de proficiência intermediário e em português, crítico. Grande parte é do sexo masculino de etnia parda e com média de idade de 17.2 anos. Para os alunos de 9º ano, em ambas disciplinas, estão categorizados como críticos. O sexo feminino é classificado como maior, a etnia parda e a média de idade para os concludentes é de 18 anos.
- Os alunos cujas mães estão registradas nas bases de violência possuem uma maior tendência a evadirem ou abandonarem.

Observa-se que, embora a situação dos alunos que concluíram os estudos sem evadir ou abandonar estejam em situações críticas, ainda são ligeiramente maiores que as médias dos alunos que se encontram em situação de abandono e evasão, onde apontam como críticas e muito críticas.

5.2 Modelo de predição

Nesta subseção serão apresentados os resultados do treinamento e testes dos dados do modelo de predição, bem como características e variáveis utilizadas.

5.2.1 Variáveis selecionadas para o treinamento

Durante a utilização do KDD, algumas variáveis foram excluídas para que os procedimentos de análises pudessem ser realizados de maneira mais eficiente, descartando de forma subjetiva variáveis que não agregam ao estudo, conforme apresentado na subseção 5.2, onde ocorreu a seleção das variáveis.

Após a primeira análise realizada, os dados seguem para os testes e treinamentos, porém, para essa etapa é necessário um novo refinamento das variáveis, para identificar e remover os atributos estatisticamente irrelevantes para a melhor utilização e desempenho dos algoritmos impostos.

Para isso, foi aplicado o algoritmo de Seleção de Característica Baseada em Correlação (CFS). O CSF, descrito na Seção 4.5, é aplicado para decidir os atributos mais importantes na previsão do desempenho do aluno. CFS é um método em que um conjunto de atributos é considerado bom se ele contém atributos altamente correlacionados com a classe, e contém atributos não correlacionados entre si. O coração do método CFS é uma heurística de avaliação de subconjuntos que considera não somente a utilidade de atributos individuais, mas também o nível de correlação entre eles. CFS primeiro calcula uma matriz de correlação de atributo-classe e atributo-atributo e um peso (score) de um conjunto de atributos é associado, pode-se observar a utilização em (1) na Seção 4.5.5. Os atributos selecionados através do algoritmo CSF são apresentados na Tabela 8. Após o refinamento através do algoritmo CSF são apresentados os seguintes resultados das características selecionadas:

Conforme mostrado na Tabela 8, 6 atributos foram selecionados como mais importantes. Além dos 6 atributos, a variável s_aluno também é selecionada pois é utilizada como saída do classificador, com 3 classes: o número 0 indica que o aluno concluiu os estudos, 1 que ele abandonou em algum momento escolar, e 2 em situação de evasão escolar. Os atributos d_mat e d_por correspondem às notas de proficiência dos alunos em matemática e português, respectivamente. O atributo tp_cor_raca possui a informação da etnia do aluno, e o atributo etapa corresponde à etapa de ensino do aluno, se ele está no 5º ou 9º ano.

Tabela 8 – Resultado da seleção do recurso CFS no conjunto de dados

Rank	Atributos Selecionados
1	d_mat
2	d_por
3	tp_cor_raca
4	etapa
5	cvli
6	m_penha

Fonte: Autor

As variáveis escolhidas pelo algoritmo CFS vão ao encontro de fatores que predizem se o estudante abandona ou consegue concluir os estudos segundo Rumberger e Lim (2008), confirmando assim a eficácia da seleção dos atributos. Segundo os autores, dentre as características individuais, os autores citam: o desempenho educacional (desempenho acadêmico e mobilidade no ensino fundamental, desempenho acadêmico ao longo do ensino médio e retenção no ensino fundamental); o comportamento e atitudes do aluno (tais como o envolvimento acadêmico no aspecto das atividades escolares e das atividades sociais, o absenteísmo e as expectativas educacionais); as características demográficas (cor/raça, gênero); e experiências prévias (como cursar a pré-escola). Ressalta-se também que as variáveis clvi e m_penha são atributos de grande importância para o estudo, através delas é possível verificar os alunos que possuem suas mães listadas em bases de dados de violência, sendo essa uma informação importante para o treinamento dos dados. Observa-se os dados na Figura 17, onde é possível verificar o percentual de ocorrências de violência em cada classe. Assim, esses atributos retornados por este método de seleção de características são referidos como características essenciais dos alunos neste estudo.

5.2.2 Resultados dos treinamentos e testes dos dados no modelo de predição

Os dados foram aplicados em três algoritmos: MLP, floresta aleatória e SVM. Em todos os experimentos foram realizados ajustes nos hiperparâmetros dos classificadores a fim de se obter os melhores resultados. Para o treinamento dos dados em todos os classificadores, foi utilizado o método de validação cruzada de dados K-Fold com K = 10. Para o classificador

MLP, foi utilizada a função de ativação sigmoidal tangente hiperbólica. As configurações dos hiperparâmetros foram realizadas com base em tentativa e erro, foram testadas várias configurações de neurônios de entrada e saída até chegar ao melhor resultado. O batch_size com valor de 100 foi escolhido, a função de perda (LossFunction) utilizada foi a de aproximação por erro quadrático com 80 épocas (number of epochs). O melhor resultado foi obtido com 15 neurônios na camada oculta e 3 neurônios de saída. Essa configuração dos hiperparâmetros garantiu o melhor resultado para o classificador MLP, que foi de 85.8%. A Figura 19 mostra a matriz de confusão resultante com MLP.

Figura 19 – Matriz de confusão MLP

	Estimado		
Real	Concluiu	Abandono	Evasão
Concluiu	99,9%	0,1%	0,0%
Abandono	1,4%	86,3%	16,2%
Evasão	15,2%	16,5%	71,2%

Fonte: Autor.

Para os casos relacionados à conclusão dos estudos, o modelo classificou corretamente 10165 amostras, o que corresponde a 99,9% dos casos. Seguidos do Abandono, onde foram classificados 86,3% dos casos, e a situação de evasão 71,2%. Um fator chave para esse índice está nos atributos correspondentes ao desempenho do aluno em português e matemática. De fato, observa-se na Tabela 6 a diferença significativa entre as médias em português e matemática dos alunos concluídos e dos alunos em situações de evasão e abandono. Essa característica tem grande importância e peso para a correta classificação das amostras classe 0. Para as classes 1 e 2 algumas amostras são erroneamente classificadas devido à proximidade de algumas características entre os alunos nas duas situações, em aspectos de raça e desempenhos.

Os erros durante o treinamento nas classificações das classes é resultado da proximidade de alguns dados, como as variáveis d_mat e d_por responsáveis pelo desempenho de um aluno. Existem amostras de alunos em situação de evasão e abandono que possuem notas de desempenho idênticas, assim como idade e sexo, nesse ponto as variáveis de violência (CVLI) e Maria da penha (m_penha) cumprem um papel importante na melhora da classificação, pois adicionam detalhes específicos para a pesquisa, que é a presença da mãe em bases de violência e

ajudam a melhor caracterizar o aluno.

Para o treinamento dos dados com SVM, os ajustes nos hiperparâmetros ocorreram através da aplicação da função GridSearch, onde são verificados os melhores valores para os hiperparâmetros e realizada a normalização dos dados, a função é disponibilizada dentro do WEKA. Para os hiperparâmetros do kernel, foi utilizado o de base radial e variável de relaxamento gamma com 0.1, o melhor resultado obtido para o classificador foi de 78.24%. A Figura 20 mostra a matriz de confusão resultante com SVM.

Figura 20 – Matriz de confusão SVM

	Estimado		
Real	Concluiu	Abandono	Evasão
Concluiu	87,3%	5,6%	7,0%
Abandono	9,4%	75,3%	15,3%
Evasão	14,1%	10,6%	72,7%

Fonte: Autor.

Para os casos relacionados à conclusão dos estudos, o modelo classificou corretamente 9076 amostras, o que corresponde a 87,3% dos casos. Foram classificados 5,6% das amostras erroneamente como sendo da categoria abandono e 7% da classe evasão. Para os casos de abandono, o modelo categorizou corretamente 7583 amostras, correspondente a 75,3% dos casos, e para os casos de abandono, foram classificadas 7502 amostras corretamente, o que corresponde a 72,7% dos casos. Para os casos de evasão, 72,7% dos casos foram classificados corretamente, 14,1% foram classificados como alunos concludentes pela aproximação das características dos alunos que saíram do sistema escolar há dois anos, definição essa que corresponde ao aluno que faz parte da classe evasão. Observa-se que o resultado não foi tão bom quanto as acurácias com MLP em decorrência da natureza dos dados e do ajuste ao modelo proposto. Ressalta-se que os resultados apresentados passaram por otimizações após aplicações de funções como GridSearch e vários reajuste nos parâmetros no ambiente Weka.

Para o treinamento da floresta aleatória, ao qual consiste em um conjunto de árvores de decisão, um dos parâmetros a ser modificado é a quantidade de árvores que será utilizada, para cada árvore os atributos serão selecionados randomicamente. No WEKA essa opção está definida como num_iterations, ao qual remete-se ao número de interações, de árvores geradas. A

definição da quantidade de interações foi realizada através de tentativa e erro, o que resultou na escolha de 70 iterações. Outro artifício utilizado dentro do ambiente WEKA foi a possibilidade da utilização de entropia. Essa função foi de grande importância devido ao tamanho da árvore que estava se criando por conta da natureza dos dados e quantidades de amostras. Os erros de redução das podas foram ativados, ou seja, a partir do cálculo da entropia dos valores, se o valor da entropia for maior que o anterior, é cortado o galho. Com essas configurações, que foram as melhores dentre os experimentos realizados, foi alcançada a acurácia de 71.4%. A Figura 21 mostra a matriz de confusão resultante com floresta aleatória.

Figura 21 – Matriz de confusão Floresta aleatória

	Estimado		
Real	Concluiu	Abandono	Evasão
Concluiu	87,3%	5,6%	7,0%
Abandono	9,4%	75,3%	15,3%
Evasão	14,1%	10,6%	72,7%

Fonte: Autor.

Para os casos relacionados à conclusão dos estudos, o modelo classificou corretamente 7023 amostras, o que corresponde a 67,9% dos casos, 15,3% foram classificados como abandono e 16,8% como evasão. Para os casos de abandono, o modelo categorizou corretamente 7583 amostras, o que corresponde a 71,8% dos casos, 13,2% foram categorizados como concludentes e 14,9% como evasão. E para os casos de evasão, foram classificadas corretamente 7502 amostras, que correspondem a 74,6% dos casos, 15,1% foram categorizados como concludentes e 10,3% como abandono. Observa-se também, comparando com os demais resultados que a Floresta aleatória foi o que obteve menor acurácia, mesmo após ajustes nos parâmetros e aplicação de poda através da entropia. O treinamento com esse classificador foi realizado em menores quantidade por conta do tempo que se consumia e a vasta quantidade de memória utilizada para os testes, de toda forma, acredita-se que o resultado seria ligeiramente maior com mais testes, uma vez a o ambiente WEKA já realizada todos os ajustes para o ganho de eficácia.

Algumas informações complementares para as medições dos índices de desempenho são os índices de sensibilidade e especificidade. A sensibilidade avalia a capacidade do método de detectar com sucesso resultados classificados como positivos, já a especificidade avalia a

capacidade do método de detectar resultados negativos. Verifica-se que os maiores índices estão de fato nos resultados obtidos com o MLP. As informações podem ser observadas na Tabela 9.

Tabela 9 – Sensibilidade dos treinamentos

Sensibilidade				
	Conclusão	**Abandono**	**Evasão**	**Média**
MLP	0,858	0,839	0,815	**0,837**
FLORESTA A.	0,709	0,731	0,705	**0,715**
SVM	0,788	0,823	0,765	**0,792**
Especificidade				
MLP	0,999	0,863	0,712	**0,858**
FLORESTA A.	0,679	0,696	0,746	**0,707**
SVM	0,873	0,753	0,727	**0,784**

Fonte: Autor

Outras variáveis podem ser utilizadas para caracterizar mais o aluno e ajudar no treinamento dos dados, como os índices de violência onde ele mora, assim como demais condições socioemocionais que venham também à agregar nas possibilidades de risco de evasão ou abandono desse aluno. Tais variáveis não são contempladas nas bases de dados do presente estudo, entretanto ficam como propostas para trabalhos futuros à adesão de mais características para melhores resultados e análises.

6 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho realizou a proposição de modelos de predição de situações de evasão e abandono para alunos do estado do Ceará, usando bases de dados relativas aos desempenho dos alunos, à situação social e relacionadas à violência contra a mulher. No total, foram 4 bases de dados: Censo escolar, SPAECE, CVLI e Maria da Penha. A ideia principal é gerar modelos de MD e aprendizado de máquina que possam medir o risco de evasão e abandono dos alunos, bem como determinar quais fatores são os que mais impactam nesse estudo.

Baseado nos resultados encontrados, pode-se afirmar que as técnicas de MD, quando aplicadas corretamente, podem trazer informações úteis para a gestão escolar e aos dirigentes no que concerne ao trabalho e acompanhamento dos estudantes que passam por essas situações, além do tracejar de estratégias que possam evitar a evasão e o abandono do aluno.

A utilização do KDD, no que concerne na limpeza e agregação de dados de várias bases de dados para gerar o conhecimento utilizado para MDE, foi de grande importância para esse trabalho, tendo em vista as grandes quantidades de dados manipuladas nas bases de dados utilizadas (ENEM, SPAECE, CLVI e Maria da penha) para que pudessem ser realizadas as análises e treinamentos da máquina.

Um ponto observado neste trabalho foi que após união das bases e análise sobre os dados de desempenho, situação da mãe em bases de violência, informações sobre idade e raça, foi possível verificar a grande diferença do desempenho estudantil entre os alunos concludentes e os caracterizados em situação de evasão e abandono. É observado, inclusive, na Figura 17, o percentual de registros de mães presentes em bases de dados de violência para cada classe, o que complementa a justificativa dos altos índices de evasão e abandono, junto de características como as médias das avaliações e na situação em que os alunos se encontram, conforme mostra tabela 6, onde são categorizados como críticos ou muito críticos.

Partindo para etapa de treinamento dos dados para o processo de classificação foram utilizados os algoritmos MLP, SVM e floresta aleatória através do software Weka. Foram utilizadas técnicas de pré-processamento de dados e mineração de dados para criar modelos preditivos capazes de classificar, com boas acurácias, estudantes que se encontram em situação de evasão.

Foram verificadas as acurácias, grau de sensibilidade e especificidade dos modelos preditivos a fim de informar quais deles obtêm melhores resultados e podem contribuir com informações úteis para melhorias no processo de identificação da evasão desses alunos. Os

resultados com a rede neural de tipo MLP obtiveram melhor eficácia atingindo 85,8% de acerto nas classificações dos alunos em suas chances de evasão ou abandono, seguido do SVM com 78,24% e da floresta aleatória com 71,4%. Os classificadores, em especial o MLP, se mostraram eficientes para predição do desempenho acadêmico a partir de informações providas das bases de dados.

Pode-se afirmar que, além da informação do registro das mães em bases de dados de violência, as demais variáveis, como informações socioemocionais, socioeconômicas e dados de registros de crimes e violência de onde o aluno habita, são complementos importantes para uma melhor caracterização das amostras para o treinamento, pois são aspectos que poderão levar a evasão ou abandono do aluno. Além disso, quanto maior e melhor caracterizadas as informações dos alunos na base de dados de treinamento, melhor será a precisão na classificação da informação do aluno estar ou não em situação de evasão ou abandono.

Para trabalhos futuros, além de agregar outras bases de dados com características que sinalizam possibilidades e risco do aluno estar em situações de evasão e abandono, a meta é abranger para os outros níveis de ensino analisando também outros fatores e variáveis que podem influenciar no desempenho preditivo e integrar os métodos propostos a sistemas utilizados pela gestão educacional através do desenvolvimento de ferramentas que permitam a educadores e administradores informar novos dados e analisar os resultados de forma fácil.

REFERÊNCIAS

ABUDI, E. C.; ASHARA, O. C. Domestic violence and adolescents' psychological adjustment. **Journal of Professional Counselling (JPC)**, v. 1, n. 2, p. 229–235, 2018.

BALFANZ, R.; HERZOG, L.; IVER, D. M. Prevenindo o desengajamento e mantendo os alunos no caminho da graduação em escolas urbanas de nível médio: identificação precoce e intervenções eficazes. **Psicólogo educacional**, v. 42, n. 4, p. 223–235, 2007.

BATTIN-PEARSON, S.; NEWOOMB, M.; ABBOTT, R.; HILL, K.; GATALANO, R.; HAWKINS, J. Preditores do abandono escolar precoce: um teste de cinco teorias. **Diário de Psicologia Educacional**, v. 92, n. 3, p. 568–582, 2000.

BERRY, M. J.; LINOFF, G. S. **Data mining techniques: for marketing, sales, and customer relationship management**. [*S. l.*]: John Wiley Sons, 2004.

BURKE, A. **Identificação precoce de resultados de formatura do ensino médio em Oregon em escolas da Rede de Liderança**. [*S. l.*], 2005.

CARDIA, N. A violência urbana e a escola. **Contemporaneidade e Educação**, v. 2, n. 2, p. –, 1997.

CHUNG, J.; KANG, T.; KIM, S.; RYOO, J.; LEE, D.; LEE, J.; HWANG, J. **Estudo de políticano Sistema de Apoio a Jovens Fora da Escola**. 2013. Jeollanamdo Office of Education.

CORTEZ, P.; SILVA, A. Usando data mining para prever o desempenho dos alunos do ensino médio. In: BRITO, A.; TEIXEIRA, J. (Ed.). **Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)**. Porto, Portugal: EUROSIS, 2008. p. 5–12. ISBN 978-9077381-39-7.

CóRDOVA, R. M. C. **Indicadores educativos: hacia um estado del arte**. México, DF: Universidad Nacional Autónoma de México: Instituto de Investigaciones Sociales, 2008.

DEJAEGER, K.; GOETHALS, F.; GIANGRECO, A.; MOLA, L.; BAESENS, B. Gaining insight into student satisfaction using comprehensible data mining techniques. **European Journal of Operational Research**, v. 218, n. 2, p. 548–562, 2012.

EDUCAçãO, C. S. da. **SPAECE**. 2012. [Online; accessed 11-Mar-2021]. Disponível em: http://www.seduc.ce.gov.br/index.php/avaliacao-educacional/62-avaliacao-educacional/spaece/5171-informacoes.

ELLIOT, D.; VOSS, H. **Delinquency and Dropout**. Lexington, MA: D.C. Heath and Company, 1974.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. In: **Advances in Knowledge Discovery and Data Mining**. AAAIPress: The Mit Press, 1996.

FERRARO, A. R.; VARGAS, E. L. B.; MACHADO, N. C. F. Qualidade das estatísticas originadas dos registros escolares: um estudo exploratório no bairro fragata, na cidade de pelotas/rs. **Sociedade em Debate**, v. 7, n. 3, p. 47–76, 2001. Disponível em: http://revistas.ucpel.tche.br/index.php/rsd/article/viewFile/564/504.

FINN, J. D. Dropout from school. **Review of Educational Research**, v. 59, n. 2, p. 117–142, 1989.

FLACH, P. **Aprendizado de máquina: a arte e a ciência dos algoritmos que dão sentido aos dados**. [*S. l.*]: Cambridge University Press, 2012.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Aprendizado profundo**. [*S. l.*]: MIT Press, 2016.

HALL, M. A. **Correlation-based feature selection for machine learning**. Tese (Doctoral Dissertation) – University of Waikato, Hamilton, New Zealand, 1999.

HAN, H.; WANG, W. Y.; MAO, B. H. Borderline-smote: um novo over-sampling método na aprendizagem de conjuntos de dados desequilibrados. In: **Conferência Internacional de Inteligência Computing**. [*S. l.*]: Springer, 2005. p. 878–887.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **Uma introdução à estatística e aprendizagem**. Nova York: Springer, 2013. v. 112.

KEDI. **Anuário estatístico da educação**. Seul: Desenvolvimento Educacional da Coréia Instituto, 2018.

KEFAS, D. P. Domestic violence: Impediment to adolescent psycho-social development. **Journal of Community Psychology**, 2016.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies**. Cambridge, MA, USA: MIT Press, 2015.

KHU, B.-Y. *et al.* **A study on the development of a support model for youth at risk**. Seoul: National Youth Commission, 2005.

KNOWLES, J. E. Of needles and haystacks: Building an accurate statewide early warning system in wisconsin. **Journal of Educational Data Mining**, v. 7, n. 3, p. 18–67, 2015.

KOPRINSKA, I.; RANA, M.; AGELIDIS, V. Correlation and instance based feature selection for electricity load forecasting. **Knowledge-Based Systems**, Elsevier, v. 82, p. 29–40, 2015.

LAMB, S.; RICE, S. **Effective strategies for increasing school completion**. Melbourne: Department of Education and Early Childhood Development, 2008.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 17, n. 4, p. 491–502, 2005.

LOPES, N. Como combater o abandono e a evasão escolar. **Revista Nova Escola**, 2015. Disponível em: http://gestaoescolar.abril.com.br/aprendizagem/como-combater-abandono-evasao-escolar-falta-alunos-abandono-acompanhamento-frequencia-551821.shtml.

ORACLE. **What is data science**. 2021. Disponível em: https://www.oracle.com/br/data-science/what-is-data-science/.

PEISNER, F.; ELLEN, S.; YAZEJIAN, N. Predicting parental perceptions of children's longitudinal school success from early child care experiences. In: AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. **Annual Meeting of the American Educational Research Association**. New Orleans, 2012. p. 197.

PEQUENO, M. I. C. Sistema permanente de avaliação da educação básica do ceará (spaece) na vertente da avaliação do rendimento escolar. **R.brás. Est. Pedag.**, v. 81, n. 197, p. 128–134, Jan./Apr. 2000.

PEQUENO, M. I. C.; COELHO, S. M. d. A. A construção do processo de avaliação educacional no ceará. In: **Avaliação institucional**. Fortaleza: Ed.UECE, 2003.

PUTERMAN, M. L. **Markov Decision Processes: Discrete Stochastic Dynamic Programming**. 1st. ed. [*S. l.*]: John Wiley & Sons, Inc., 1994.

RISTUM, M. P. A violência doméstica contra crianças e as implicações da escola. **Temas em Psicologia**, v. 18, n. 1, p. 231–232, 2010. Accessed on March 20, 2021. Disponível em: http://www.sbponline.org.br/revista2.

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

RUMBERGER, R.; LARSON, K. Toward explaining the differences in educational achievement among mexican american language minority students. **Sociology of Education**, v. 71, p. 69–93, 1998.

RUMBERGER, R.; LIMA, S. A. **Why Students Drop Out: A Review of 25 Years of Research**. 2008. California Dropout Research Project, Policy Brief 15, University of California.

SIGOLO, S. R. L.; LOLLATO, S. O. Aproximações entre escola e família: um desafio para educadores. In: CHAKUR, C. (Ed.). **Problemas da Educação sob o Olhar da Psicologia**. São Paulo: Cultura Acadêmica Editora, 2001. p. 37–65.

SILVA, R. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis conseqüências. **Educação Por Escrito**, Universidade Federal do Rio Grande do Sul, v. 8, n. 1, p. 35–48, 2017.

SINGH, S.; KUMAR, V. Classification of student's data using data mining techniques for training, placement department in technical education. **International Journal of Computer Science and Network**, Academy Industry Research Collaboration Center (AIRCC), v. 1, n. 4, p. 121–126, 2012.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. [*S. l.*]: MIT press, 2018.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 2. ed. [*S. l.*]: Pearson, 2018. ISBN 9780133128901 and 0133128903. ISBN 9780133128901.

TEIXEIRA, I. N. de Estudos e P. E. A. **Censo escolar da educação básica**. Brasília, DF: INEP/MEC, 2013.

THERRIAULT, S.; HEPPEN, J.; O'CUMMINGS, M.; FRYER, L.; JOHNSON, A. **Aviso prévio guia de implementação do sistema**. 2010. Obtido no site do National High School Center: http://www.melhoresescolasdeensinomedio.org/documents/NHSCEWSImplementationGuide.pdf.

VECINA, T.; FERRARI, D. **O fim do silêncio na violência familiar. Teoria e prática**. São Paulo: Agora, 2002.

VIANA, H. M. Avaliações nacionais em larga escala: análises e propostas. **Estudos em Avaliação Educacional**, n. 27, p. 1–26, 2003. Disponível em: http//educa.fcc.org.br/pdf/eae/n27/n27a02.pdf.

WILTZ, J.; SLATE, J. R. Differences in dropout rates by ethnicity/race of middle school students: A multi-year analysis. **Global Journal of Human-Social Science Research**, v. 16, n. 8, p. 30–34, 2016.

YOON, C.; RYU, B.; KIM, S. **Uma análise aprofundada do abandono escolar e um estudo sobre estratégias personalizadas**. 2010. Instituto Nacional de Políticas Juvenis da Coréia.

ZHANG, Y.; WU, B. Research and application of grade prediction model based on decision tree algorithm. In: ACM. **Turing Celebration Conference (ACM TURC 2019)**. Chengdu, China, 2019. p. 1–6. Disponível em: https://doi.org/10.1145/3321408.3322857.

APÊNDICE A - VARIÁVEIS MARIA DA PENHA

1	**AIS**	**Áreas Integradas de Segurança (AIS)**	**INT**
2	**DATA**	**Data do ocorrido**	**DD/MM/AAAA**
3	**HORA**	**Hora do ocorrido**	**HH:MM**
4	**DIA DA SEMANA**	**Dia do ocorrido**	**VARCHAR**
5	**IDADE**	**Idade da vítima**	**INT**
6	**SEXO DA VÍTIMA**	**Sexo**	**CHAR**
7	**NOME DA VÍTIMA**	**Nome da vítima**	**VARCHAR**
8	**RAÇA**	**Raça do ocorrido**	**VARCHAR**
9	**PROFISSÃO**	**Profissão do ocorrido**	**VARCHAR**
10	**ESTADO CIVIL**	**Estado Civil da vítima**	**VARCHAR**
11	**ESCOLARIDADE**	**Escolaridade da vítima**	**VARCHAR**

APÊNDICE B - VARIÁVEIS CLVI

1	**AIS**	**Áreas Integradas de Segurança (AIS)**	**INT**
2	**TIPO_ARMA**	**Arma Branca, Arma de Fogo, Outros Meios**	**VARCHAR**
3	**MUNICÍPIO**	**Município do ocorrido**	**VARCHAR**
4	**DATA**	**Data do ocorrido**	**DD/MM/AAAA**
5	**SEXO DA VÍTIMA**	**Sexo da Vítima**	**VARCHAR**
6	**NOME DA VÍTIMA**	**Nome da Vítima**	**VARCHAR**
7	**IDADE**	**Idade da vítima**	**INT**
8	**RAÇA**	**Raça da vítima**	**VARCHAR**
9	**PROFISSÃO**	**Profissão da vítima**	**VARCHAR**
10	**ESTADO CIVIL**	**Estado Civil da vítima**	**VARCHAR**
11	**ESCOLARIDADE**	**Escolaridade da vítima**	**VARCHAR**

ANEXO A - Variáveis CENSO

N	Nome da Variável	Descrição da Variável	Tipo	Tam.(1)	Categoria
1	NU_ANO_CENSO	Ano do Censo	Num	4	
DADOS DO ALUNO					
2	ID_ALUNO	Código do aluno (ID_INEP)	CHAR	32	
3	ID_MATRICULA	Código único da matrícula	Num	8	
4	NU_DIA	Data de nascimento do aluno - dia	Num	2	DD
5	NU_MES	Data de nascimento do aluno - mês	Num	2	MM
6	NU_ANO	Data de nascimento do aluno - ano	Num	4	YYYY
7	NU_IDADE_REFERENCIA	Idade do aluno no mês de referência do Censo Escolar (31	Num	3	

		de maio)			
8	NU_IDADE	Idade calculada pelo ano de nascimento do aluno	Num	3	
9	TP_SEXO	Sexo	Num	1	1 - Masculino 2 - Feminino
10	TP_COR_RACA	Cor/raça	Num	1	0 - Não declarada 1 - Branca 2 - Preta 3 - Parda 4 - Amarela 5 - Indígena
11	TP_NACIONALIDADE	Nacionalidade	Num	1	1 - Brasileira 2 - Brasileira - nascido no exterior ou naturalizado 3 - Estrangeira
12	CO_PAIS_ORIGEM	Código País da nacionalidade	Num	3	Ver Anexo 4 - Não aplicável para alunos de nacionalidade brasileira ou brasileira nascido no exterior ou naturalizado

13	**CO_UF_NASC**	**Código UF de nascimento**	**Num**	**2**	**- Não aplicável para alunos de nacionalidade estrangeira**
14	**CO_MUNICIPIO_NASC**	**Código Município de nascimento**	**Num**	**7**	**- Não aplicável para alunos de nacionalidade estrangeira**
15	**CO_PAIS_RESIDENCIA**	**Código País de residência[2]**	**Num**	**3**	**Ver Anexo 4**
16	**CO_UF_END**	**Código UF de residência**	**Num**	**2**	**- Não aplicável para alunos residentes no exterior**
17	**CO_MUNICIPIO_END**	**Código Município de residência**	**Num**	**7**	**- Não aplicável para alunos residentes no exterior**
18	**TP_ZONA_RESIDENCIAL**	**Localização/Zona de residência**	**Num**	**1**	**1 - Urbana 2 - Rural - Não aplicável para alunos residentes no exterior**
19	**TP_LOCAL_RESI**	**Localização**	**Num**	**1**	**0 - Não reside em**

	D_DIFERENCIADA	**diferenciada da residência**			**área de localização diferenciada 1 - Área onde se localiza comunidade remanescente de quilombos 2 - Terra indígena 3 - Área de assentamento**
20	**IN_NECESSIDADE_ESPECIAL**	**Aluno (a) com deficiência, transtorno do espectro autista ou altas habilidades/superdotação**	**Num**	**1**	**0 - Não 1 - Sim**
21	**IN_BAIXA_VISAO**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Baixa visão**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
22	**IN_CEGUEIRA**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Cegueira**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**

23	**IN_DEF_AUDITIVA**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação (deficiência auditiva)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
24	**IN_DEF_FISICA**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Deficiência física**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
25	**IN_DEF_INTELECTUAL**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Deficiência intelectual**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
26	**IN_SURDEZ**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Surdez**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**

27	**IN_SURDOCEGUEIRA**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Surdocegueira**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
28	**IN_DEF_MULTIPLA**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Deficiência múltipla**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
29	**IN_AUTISMO**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Transtorno do Espectro Autista**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
30	**IN_SINDROME_ASPERGER**	**Tipo de deficiência, transtorno global do desenvolvimento ou altas habilidades/superdotação (Síndrome de Asperger)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**

31	**IN_SINDROME_RETT**	**Tipo de deficiência, transtorno global do desenvolvimento ou altas habilidades/superdotação (Síndrome de Rett)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
32	**IN_TRANSTORNO_DI**	**Tipo de deficiência, transtorno global do desenvolvimento ou altas habilidades/superdotação (Transtorno desintegrativo da infância)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
33	**IN_SUPERDOTACAO**	**Tipo de deficiência, transtorno do espectro autista ou altas habilidades/superdotação - Altas habilidades/superdotação**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
34	**IN_RECURSO_LEDOR**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem**

		Inep (Saeb) - Auxílio Ledor			**deficiência**
35	**IN_RECURSO_TRANSCRICAO**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Auxílio Transcrição**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
36	**IN_RECURSO_INTERPRETE**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Guia-Intérprete**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
37	**IN_RECURSO_LIBRAS**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Tradutor e Intérprete de Libras**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
38	**IN_RECURSO_LABIAL**	**Recursos necessários para uso do(a) aluno(a) e**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado**

		para a participação em avaliações do Inep (Saeb) - Leitura Labial			**- Não aplicável para alunos sem deficiência**
39	**IN_RECURSO_AMPLIADA_18**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Prova Ampliada (Fonte tamanho 18)**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
40	**IN_RECURSO_AMPLIADA_16**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Prova Ampliada (Fonte tamanho 16)**	**Num**	**1**	**0 - Não 1 - Sim**
41	**IN_RECURSO_AMPLIADA_20**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Prova Ampliada (Fonte tamanho 20)**	**Num**	**1**	**0 - Não 1 - Sim**

42	IN_RECURSO_AMPLIADA_24	Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Prova superampliada (Fonte tamanho 24)	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência
43	IN_RECURSO_CD_AUDIO	Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - CD com áudio para deficiente visual	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência
44	IN_RECURSO_PROVA_PORTUGUES	Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Prova de Língua Portuguesa como segunda língua para surdos e deficientes auditivos	Num	1	0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência
45	IN_RECURSO_VI	Recursos	Num	1	0 - Não

	DEO_LIBRAS	**necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Prova em vídeo Libras**			**1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
46	**IN_RECURSO_BRAILLE**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Material didático e Prova em Braille**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
47	**IN_RECURSO_NENHUM**	**Recursos necessários para uso do(a) aluno(a) e para a participação em avaliações do Inep (Saeb) - Nenhum**	**Num**	**1**	**0 - Não 1 - Sim 9 - Não informado - Não aplicável para alunos sem deficiência**
48	**IN_AEE_LIBRAS**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino da Língua Brasileira de Sinais - LIBRAS**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**

49	**IN_AEE_LINGUA_PORTUGUESA**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino da Língua Portuguesa como segunda língua**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
50	**IN_AEE_INFORMATICA_ACESSIVEL**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino da informática acessível**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
51	**IN_AEE_BRAILLE**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino do Sistema Braille**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
52	**IN_AEE_CAA**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino do uso da Comunicação Alternativa e Aumentativa (CAA)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**

53	**IN_AEE_SOROBAN**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino das técnicas do cálculo no Soroban**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
54	**IN_AEE_VIDA_AUTONOMA**	**Tipo de Atendimento Educacional Especializado (AEE) - Desenvolvimento de vida autônoma**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
55	**IN_AEE_OPTICOS_NAO_OPTICOS**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino do uso de recursos ópticos e não ópticos**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
56	**IN_AEE_ENRIQ_CURRICULAR**	**Tipo de Atendimento Educacional Especializado (AEE) - Enriquecimento curricular**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
57	**IN_AEE_DESEN_C**	**Tipo de**	**Num**	**1**	**0 - Não**

	OGNITIVO	**Atendimento Educacional Especializado (AEE) - Desenvolvimento de funções cognitivas**			**1 - Sim - Não aplicável para alunos sem deficiência**
58	**IN_AEE_MOBILIDADE**	**Tipo de Atendimento Educacional Especializado (AEE) - Ensino de técnicas para orientação e mobilidade**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para alunos sem deficiência**
59	**TP_OUTRO_LOCAL_AULA**	**Recebe escolarização em outro local (diferente da escola)**	**Num**	**1**	**1 - Em hospital 2 - Em domicílio 3 - Não recebe escolarização fora da escola**
60	**IN_TRANSPORTE_PUBLICO**	**Transporte escolar público**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos em turmas de EAD**
61	**TP_RESPONSAVEL_TRANSPORTE**	**Poder público responsável pelo transporte escolar**	**Num**	**1**	**1 - Estadual 2 - Municipal - Não aplicável para alunos em turmas**

					de EAD
62	**IN_TRANSP_BICICLETA**	**Tipo de veículo utilizado no transporte escolar público - Rodoviário (Bicicleta)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
63	**IN_TRANSP_MICRO_ONIBUS**	**Tipo de veículo utilizado no transporte escolar público - Rodoviário (Micro-ônibus)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
64	**IN_TRANSP_ONIBUS**	**Tipo de veículo utilizado no transporte escolar público - Rodoviário (Ônibus)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
65	**IN_TRANSP_TR_ANIMAL**	**Tipo de veículo utilizado no transporte escolar público - Rodoviário (Tração Animal)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
66	**IN_TRANSP_VANS**	**Tipo de veículo**	**Num**	**1**	**0 - Não utiliza**

	_KOMBI	**utilizado no transporte escolar público - Rodoviário (Vans/Kombi)**			**1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
67	**IN_TRANSP_OUTRO_VEICULO**	**Tipo de veículo utilizado no transporte escolar público - Rodoviário (Outro tipo de veículo rodoviário)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
68	**IN_TRANSP_EMBAR_ATE5**	**Tipo de veículo utilizado no transporte escolar público - Aquaviário/Embarcação (Capacidade de até 5 alunos)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
69	**IN_TRANSP_EMBAR_5A15**	**Tipo de veículo utilizado no transporte escolar público - Aquaviário/Embarcação (Capacidade de 5 a 15 alunos)**	**Num**	**1**	**0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público**
70	**IN_TRANSP_EMB**	**Tipo de veículo**	**Num**	**1**	**0 - Não utiliza**

	AR_15A35	utilizado no transporte escolar público - Aquaviário/Embarcação (Capacidade de 15 a 35 alunos)			1 - Utiliza - Não aplicável para alunos que não utilizam transporte público
71	IN_TRANSP_EMBAR_35	Tipo de veículo utilizado no transporte escolar público - Aquaviário/Embarcação (Capacidade acima de 35 alunos)	Num	1	0 - Não utiliza 1 - Utiliza - Não aplicável para alunos que não utilizam transporte público
72	IN_TRANSP_TREM_METRO	Tipo de veículo utilizado no transporte escolar Ferroviário (Trem/Metrô)	Num	1	0 - Não utiliza 1 - Utiliza
73	TP_INGRESSO_FEDERAIS	Forma de ingresso do aluno (apenas alunos de escolas federais das Etapas: Educação Profissional (39 e 40), Ensino Médio Integrado (30, 31, 32, 33 e 34) e EJA integrada à	Num	1	1 - Sem processo seletivo 2 - Sorteio 3 - Transferência 4 - Exame de seleção sem reserva de vaga 5 - Exame de seleção, vaga reservada para alunos da rede

		Educação Profissional de Nível Médio (73)			**pública de ensino 6 - Exame de seleção, vaga reservada para alunos da rede pública de ensino, com baixa renda e autodeclarado preto, pardo ou indígena 7 - Exame de seleção, vaga reservada para outros programas de ação afirmativa 8 - Outra forma de ingresso 9 - Exame de seleção, vaga reservada para alunos da rede pública de ensino, com baixa renda**
74	**TP_ETAPA_ENSINO**	**Etapa de ensino da matrícula**	**Num**	**2**	**1 - Educação Infantil - Creche**
					2 - Educação Infantil - Pré-escola
					4 - Ensino

					Fundamental de 8 anos - 1ª Série
					5 - Ensino Fundamental de 8 anos - 2ª Série
					6 - Ensino Fundamental de 8 anos - 3ª Série
					7 - Ensino Fundamental de 8 anos - 4ª Série
					8 - Ensino Fundamental de 8 anos - 5ª Série
					9 - Ensino Fundamental de 8 anos - 6ª Série
					10 - Ensino Fundamental de 8 anos - 7ª Série
					11 - Ensino Fundamental de 8 anos - 8ª Série

					14 - Ensino Fundamental de 9 anos - 1º Ano
					15 - Ensino Fundamental de 9 anos - 2º Ano
					16 - Ensino Fundamental de 9 anos - 3º Ano
					17 - Ensino Fundamental de 9 anos - 4º Ano
					18 - Ensino Fundamental de 9 anos - 5º Ano
					19 - Ensino Fundamental de 9 anos - 6º Ano
					20 - Ensino Fundamental de 9 anos - 7º Ano
					21 - Ensino Fundamental de 9 anos - 8º Ano

					41 - Ensino Fundamental de 9 anos - 9º Ano
					25 - Ensino Médio - 1º ano/1ª Série
					26 - Ensino Médio - 2º ano/2ª Série
					27 - Ensino Médio - 3ºano/3ª Série
					28 - Ensino Médio - 4º ano/4ª Série
					29 - Ensino Médio - Não Seriada
					30 - Curso Técnico Integrado (Ensino Médio Integrado) 1ª Série
					31 - Curso Técnico Integrado (Ensino Médio Integrado) 2ª Série
					32 - Curso Técnico Integrado (Ensino

					Médio Integrado) 3ª Série
					33 - Curso Técnico Integrado (Ensino Médio Integrado) 4ª Série
					34 - Curso Técnico Integrado (Ensino Médio Integrado) Não Seriada
					35 - Ensino Médio - Modalidade Normal/Magistério 1ª Série
					36 - Ensino Médio - Modalidade Normal/Magistério 2ª Série
					37 - Ensino Médio - Modalidade Normal/Magistério 3ª Série
					38 - Ensino Médio - Modalidade Normal/Magistério

					4ª Série
					39 - Curso Técnico - Concomitante
					40 - Curso Técnico - Subsequente
					65 - EJA - Ensino Fundamental - Projovem Urbano
					67 - Curso FIC integrado na modalidade EJA - Nível Médio
					68 - Curso FIC Concomitante
					69 - EJA - Ensino Fundamental - Anos Iniciais
					70 - EJA - Ensino Fundamental - Anos Finais
					71 - EJA - Ensino Médio
					72 - EJA - Ensino

					Fundamental - Anos iniciais e Anos finais[5]
					73 - Curso FIC integrado na modalidade EJA - Nível Fundamental (EJA integrada à Educação Profissional de Nível Fundamental)
					74 - Curso Técnico Integrado na Modalidade EJA (EJA integrada à Educação Profissional de Nível Médio) **- Não aplicável para turmas de atendimento educacional especializado (AEE) e atividade complementar**
75	**IN_ESPECIAL_EXCLUSIVA**	**Aluno de turma exclusiva de alunos com deficiência,**	**Num**	**1**	**0 - Não** **1 - Sim** **- Não aplicável para**

		transtorno do espectro autista ou altas habilidades/superdotação (Classes Especiais)			**turmas de atendimento educacional especializado (AEE) e atividade complementar**
76	**IN_REGULAR**	**Modo, maneira ou metodologia de ensino correspondente às turmas com etapas de escolarização consecutivas, Creche ao Ensino Médio. Etapas consideradas (nas antigas modalidades 1 ou 2): TP_ETAPA_ENSINO igual a 1,2,4,5,6,7,8,9,10,11, 14, 15,16,17,18,19,20,21, 41,25,26,27,28,29,30, 31, 32,33,34,35,36,37 ou 38.**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para turmas de atendimento educacional especializado (AEE) e atividade complementar**
77	**IN_EJA**	**Modo, maneira ou metodologia de**	**Num**	**1**	**0 - Não 1 - Sim**

		ensino correspondente às turmas destinadas a pessoas que não cursaram o ensino fundamental e/ou médio em idade própria. Etapas consideradas (nas antigas modalidades 2 ou 3): TP_ETAPA_ENSINO igual a 65,67,69,70,71,73 ou 74.			**- Não aplicável para turmas de atendimento educacional especializado (AEE) e atividade complementar**
78	**IN_PROFISSIONALIZANTE**	**Modo profissionalizante de ensino correspondente às turmas de cursos de formação inicial e continuada ou de qualificação profissional (Cursos FIC) articulados à EJA ou concomitantes; ou de cursos técnicos de nível médio nas**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para turmas de atendimento educacional especializado (AEE) e atividade complementar**

		formas articulada (integrada ou concomitante) ou subsequente ao ensino médio e de normal/magistério. Etapas consideradas (nas antigas modalidades 1, 2 ou 3): TP_ETAPA_ENSINO igual a 30,31,32,33,34, 35,36,37,38,39,40,65, 67,68,73 ou 74.			
DADOS DA TURMA					
79	**ID_TURMA**	**Código único da Turma**	**Num**	**8**	
80	**CO_CURSO_EDUC_PROFISSIONAL**	**Curso da Educação Profissional Técnica (Apenas etapas: 30,31,32,33,34,74,39, e 40**	**Num**	**5**	**Ver Anexo 2 - Cursos da Educação Profissional Técnica**
81	**TP_MEDIACAO_DIDATICO_PEDAGO**	**Tipo de mediação didático-pedagógica**	**Num**	**1**	**1 - Presencial 2 - Semipresencial 3 - Educação a Distância - EAD**

82	**NU_DURACAO_TURMA**	**Tempo de permanência na turma da matrícula do aluno - minutos**	**Num**	**4**	**- Não aplicável para alunos em turmas de EAD**
83	**NU_DUR_ATIV_COMP_MESMA_REDE**	**Tempo de permanência (em minutos) na turma de atividade complementar na mesma rede da turma de escolarização [3,4]**	**Num**	**4**	**- Não aplicável para alunos em turmas de EAD**
84	**NU_DUR_ATIV_COMP_OUTRAS_REDES**	**Tempo de permanência (em minutos) na turma de atividade complementar em outras redes [3,4]**	**Num**	**4**	**- Não aplicável para alunos em turmas de EAD**
85	**NU_DUR_AEE_MESMA_REDE**	**Tempo de permanência (em minutos) na turma de atendimento educacional especializado (AEE) na mesma rede da turma de escolarização [3,4]**	**Num**	**4**	**- Não aplicável para alunos em turmas de EAD**

86	**NU_DUR_AEE_OUTRAS_REDES**	**Tempo de permanência (em minutos) na turma de atendimento educacional especializado (AEE) em outras redes [3.4]**	**Num**	**4**	**- Não aplicável para alunos em turmas de EAD**
87	**NU_DIAS_ATIVIDADE**	**Número de dias por semana em que são realizadas as atividades da turma**	**Num**	**1**	**1 - Uma vez por semana 2 - Duas vezes por semana 3 - Três vezes por semana 4 - Quatro vezes por semana 5 - Cinco vezes por semana 6 - Seis vezes por semana 7 - Sete vezes por semana - Não aplicável para alunos em turmas de EAD**
88	**TP_UNIFICADA**	**Unificada, multietapa, multi ou correção de fluxo**	**Num**	**1**	**0 - Não 1 - Unificada 2 - Multietapa 3 - Multi**

					4 - Correção de fluxo **5 - Mista (Concomitante e Subsequente)**
89	**TP_TIPO_TURMA**	**Tipo de atendimento**	**Num**	**1**	**0 - Não se aplica** **1 - Classe hospitalar** **2 - Unidade de atendimento socioeducativo** **3 - Unidade prisional** **4 - Atividade complementar** **5 - Atendimento Educacional Especializado (AEE)**
90	**TP_TIPO_ATENDIMENTO_TURMA**	**Tipo de atendimento**	**Num**	**1**	**1 - Exclusivo Escolarização** **2 - Atividade complementar e escolarização** **3 - Atividade complementar** **4 - Atendimento Educacional Especializado (AEE)**
91	**TP_TIPO_LOCAL_**	**Local de**	**Num**	**1**	**0 - A turma não está**

	TURMA	**funcionamento diferenciado da turma**			**em local de funcionamento diferenciado 1 - Sala anexa 2 - Unidade de atendimento socioeducativo 3 - Unidade prisional**
	DADOS DA ESCOLA				
92	**CO_ENTIDADE**	**Código da Escola**	**Num**	**8**	
93	**CO_REGIAO**	**Código da região geográfica**	**Num**	**1**	
94	**CO_MESORREGIAO**	**Código da mesorregião**	**Num**	**4**	
95	**CO_MICRORREGIAO**	**Código da microrregião**	**Num**	**5**	
96	**CO_UF**	**Código UF da escola**	**Num**	**2**	
97	**CO_MUNICIPIO**	**Código Município da escola**	**Num**	**7**	
98	**CO_DISTRITO**	**Código completo do Distrito da escola**	**Num**	**9**	
99	**TP_DEPENDENCI**	**Dependência**	**Num**	**1**	**1 - Federal**

	A	Administrativa (Escola)			2 - Estadual 3 - Municipal 4 - Privada
100	TP_LOCALIZACAO	Localização (Escola)	Num	1	1 - Urbana 2 - Rural
101	TP_CATEGORIA_ESCOLA_PRIVADA	Categoria da escola privada	Num	1	1 - Particular 2 - Comunitária 3 - Confessional 4 - Filantrópica - Não aplicável para escolas públicas
102	IN_CONVENIADA_PP	Conveniada com o poder público (Escola)	Num	1	0 - Não tem convênio 1 - Sim - Não aplicável para escolas públicas
103	TP_CONVENIO_PODER_PUBLICO	Dependência do convênio com o poder público	Num	1	1 - Municipal 2 - Estadual 3 - Estadual e Municipal - Não aplicável para escolas públicas ou sem convênio

104	**IN_MANT_ESCOLA_PRIVADA_EMP**	**Mantenedora da escola privada - Empresa ou grupo empresarial do setor privado ou pessoa física**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**
105	**IN_MANT_ESCOLA_PRIVADA_ONG**	**Mantenedora da escola privada - Organização não governamental (ONG) - internacional ou nacional.**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**
106	**IN_MANT_ESCOLA_PRIVADA_OSCIP**	**Mantenedora da escola privada - Organização da Sociedade Civil de Interesse Público (Oscip)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**
107	**IN_MANT_ESCOLA_PRIV_ONG_OSCIP**	**Mantenedora da escola privada - Organização não governamental (ONG) - internacional ou nacional. Organização da Sociedade Civil de**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**

		Interesse Público (Oscip)			
108	**IN_MANT_ESCOLA_PRIVADA_SIND**	**Mantenedora da escola privada - Sindicatos de trabalhadores ou patronais, associações e cooperativas**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**
109	**IN_MANT_ESCOLA_PRIVADA_SIST_S**	**Mantenedora da escola privada - Sistema S (Sesi, Senai, Sesc, Outros)**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**
110	**IN_MANT_ESCOLA_PRIVADA_S_FINS**	**Mantenedora da escola privada - Instituições sem fins lucrativos**	**Num**	**1**	**0 - Não 1 - Sim - Não aplicável para escolas públicas**
111	**TP_REGULAMENTACAO**	**Regulamentação/Autorização no conselho ou órgão municipal, estadual ou federal de educação**	**Num**	**1**	**0 - Não 1 - Sim 2 - Em tramitação**
112	**TP_LOCALIZACA**	**Localização**	**Num**	**1**	**0 - A escola não está**

	O_DIFERENCIADA	**diferenciada da escola**			**em área de localização diferenciada 1 - Área de assentamento 2 - Terra indígena 3 - Área onde se localiza comunidade remanescente de quilombos**
113	**IN_EDUCACAO_INDIGENA**	**Educação Escolar Indígena**	**Num**	**1**	**0 - Não 1 - Sim**

ANEXO B - VARIÁVEIS SPAECE

	NOME DA VARIÁVEL	DESCRIÇÃO	TIPO
1	CD_ALUNO	Aluno	BIGINT
2	CD_CADERNO	Caderno	BIGINT
3	CD_CURSO	Curso	BIGINT
4	CD_DEFICIENCIA	Deficiência	BIGINT
5	CD_DISCIPLINA	Disciplina	BIGINT
6	CD_DISTRITO	Distrito	BIGINT
7	CD_ENSINO	Ensino	BIGINT
8	CD_ESCOLA	Escola	BIGINT

9	**CD_ESTADO**	**Estado**	**BIGINT**
10	**CD_ETAPA**	**Etapa de Escolaridade**	**BIGINT**
11	**CD_GRUPO_REFERENCIA**	**Grupos de escolas semelhantes, porém essa semelhança é defina por cada projeto**	**BIGINT**
12	**CD_IDENTIFICADOR**	**Identificador único do aluno**	**BIGINT**
13	**CD_LOCAL_ENTREGA**	**Local de entrega – Controle logística**	**BIGINT**
14	**CD_LOCALIZACAO**	**Localização**	**BIGINT**
15	**CD_MUNICIPIO**	**Município**	**BIGINT**
16	**CD_OBSERVACAO**	**Observação**	**BIGINT**
17	**CD_POLO**	**Pólo**	**BIGINT**

18	**CD_PROFESSOR**	**Professor**	**BIGINT**
19	**CD_PROJETO**	**Projeto**	**BIGINT**
20	**CD_REDE_ENSINO**	**Rede de ensino**	**BIGINT**
21	**CD_REGIAO**	**Região**	**BIGINT**
22	**CD_REGIONAL**	**Regional**	**BIGINT**
23	**CD_ROTA**	**Rota – Controle logística**	**BIGINT**
24	**CD_TIPO_ESCOLA**	**Tipo de escola**	**BIGINT**
25	**CD_TURMA**	**Turma**	**BIGINT**
26	**CD_TURNO**	**Turno**	**BIGINT**
27	**CD_UF**	**UF**	**BIGINT**
28	**DC_CADERNO**	**Caderno**	**VARCHAR**

29	**DC_CAMINHO_IMAGEM_XX**	**Caminho da imagem**	**VARCHAR**
30	**DC_CAP**	**Flag para casos validos para divulgação**	**VARCHAR**
31	**DC_DEFICIENCIA**	**Deficiência**	**VARCHAR**
32	**DC_ENSINO**	**Ensino**	**VARCHAR**
33	**DC_ETAPA**	**Etapa de Escolaridade**	**VARCHAR**
34	**DC_GRUPO_REFERENCIA**	**Grupo de referência**	**VARCHAR**
35	**DC_LAUDO**	**Flag para laudo**	**VARCHAR**
36	**DC_LOCAL_ENTREGA**	**Local de entrega**	**VARCHAR**
37	**DC_LOCALIZACAO**	**Localização**	**VARCHAR**
38	**DC_OBSERVACAO**	**Observação**	**VARCHAR**

39	**DC_REDE_ENSINO**	**Rede**	**VARCHAR**
40	**DC_SEXO**	**Sexo**	**VARCHAR**
41	**DC_SITUACAO**	**Situação do aluno no dia da avaliação**	**VARCHAR**
42	**DC_TIPO_ESCOLA**	**Tipo de escola**	**VARCHAR**
43	**DC_TRANSFERIDO**	**Transferido**	**VARCHAR**
44	**DC_TURMA**	**Turma**	**VARCHAR**
45	**DC_TURNO**	**Turno**	**VARCHAR**
46	**DT_NASCIMENTO**	**Nascimento**	**DATE**
47	**FL_ADICIONAL_REDACAO**	**Adicional de redação**	**INT**
48	**FL_ANEXO**	**Anexo**	**INT**

49	**FL_AVALIADO**	**Informar se o aluno é ou não Avaliado**	**INT**
50	**FL_BLOCO**	**Bloco**	**INT**
51	**FL_CAP**	**Cap, registro válido**	**INT**
52	**FL_EM_INOVADOR**	**Ensino médio inovador**	**INT**
53	**FL_ESC_EM_GESTAR**	**Ensino médio gestar**	**INT**
54	**FL_ESC_EM_INTEGRAL**	**Ensino médio integral**	**INT**
55	**FL_ESC_ESPECIAIS**	**Escolas especiais**	**INT**
56	**FL_EXTRA**	**Cartão extra ou remanejamento realizado internamente pelo CAEd.**	**INT**
57	**FL_FAETEC**	**Escolas Faetec**	**INT**

58	**FL_INDIGENA**	**Indicar escolas Indígenas**	**INT**
59	**FL_LAUDO**	**Indicar alunos com Laudo médico**	**INT**
60	**FL_MULT**	**Turmas múltiplas**	**INT**
61	**FL_MULTSERIADO**	**Turmas multiseriadas**	**INT**
62	**FL_OLIMPIADA_ETAPA_ESCOLAR**	**Olimpíada de etapa escolar**	**INT**
63	**FL_PRISIONAL**	**Aluno prisional**	**INT**
64	**FL_QUEST_MODELO**	**Modelo de questionário**	**INT**
65	**FL_QUEST_NBRANCO**	**Questionário em branco**	**INT**
66	**FL_QUEST_REPETIDO**	**Questionário repetido**	**INT**
67	**FL_RETIRAR**	**Retirar registro (s) de alguma**	**INT**

		especificidade do(s) resultado(s).	
68	FL_RETORNOU	Indicar se o cartão retornou ao CAEd e foi lido no scanner	INT
69	FL_TRANSFERIDO	Indicar Aluno transferido	INT
70	ID_ALUNO	Aluno	BIGINT
71	ID_PAC_FIM	Pacote de provas inicial	BIGINT
72	ID_PAC_INI	Pacote de provas final	BIGINT
73	ID_SEXO	Sexo do aluno	BIGINT
74	NM_ALUNO	Aluno	VARCHAR
75	NM_ANEXO_TURMA	Turma anexa	VARCHAR
76	NM_CURSO	Curso	VARCHAR

77	**NM_DISCIPLINA**	**Disciplina**	**VARCHAR**
78	**NM_DISTRITO**	**Distrito**	**VARCHAR**
79	**NM_ESCOLA**	**Escola**	**VARCHAR**
80	**NM_ESTADO**	**Estado**	**VARCHAR**
81	**NM_LOCAL_DE_ENTREGA**	**Local de entrega das provas**	**VARCHAR**
82	**NM_MAE**	**Mãe do aluno**	**VARCHAR**
83	**NM_MUNICIPIO**	**Município**	**VARCHAR**
84	**NM_POLO**	**Pólo**	**VARCHAR**
85	**NM_PROFESSOR**	**Professor**	**VARCHAR**
86	**NM_PROJETO**	**Projeto**	**VARCHAR**
87	**NM_REGIAO**	**Região**	**VARCHAR**

88	**NM_REGIONAL**	**Regional**	**VARCHAR**
89	**NM_TURMA**	**Turma**	**VARCHAR**
90	**NM_UNIDADE_OPERACIONAL**	**Unidade operacional**	**VARCHAR**
91	**NU_ACERTADO**	**Acertos**	**INT**
92	**NU_CADERNO**	**Caderno**	**BIGINT**
93	**NU_CPF**	**CPF**	**BIGINT**
94	**NU_ESTRATO**	**Estrato**	**INT**
95	**NU_FCA**	**Formulário de Controle de Aplicação**	**BIGINT**
96	**NU_GRUPO_COMPARACAO**	**Grupo de comparação**	**INT**
97	**NU_GRUPO_REFERENCIA**	**Grupo de referência**	**BIGINT**

98	**NU_LISTA_PRESENCA**	**Lista de presença**	**BIGINT**
99	**NU_ORDEM_MODELO**	**Ordem de modelo**	**BIGINT**
100	**NU_PACOTE**	**Pacote**	**BIGINT**
101	**NU_PNT_ALN_XX**	**Pontos do aluno para uma edição ou ano “XX”**	**BIGINT**
102	**NU_PONTOS**	**Pontos**	**BIGINT**
103	**NU_SEQUENCIAL**	**Seqüencial, identificação do aluno na prova**	**BIGINT**
104	**NU_SEQUENCIAL_MODELO**	**Seqüencial modelo**	**BIGINT**
105	**NU_TESTE**	**Teste**	**BIGINT**
106	**VL_D_PCT_ALN_XX**	**Acerto por Aluno no ano ou Edição “XX”**	

107	**REDACAO_CCX**	**Critério de avaliação para redação por competência X**	**INT**
108	**REDACAO_DC_CORRECAO_SITUACAO**	**Descrição da correção da redação**	**VARCHAR**
109	**REDACAO_FL_AVALIADO**	**Flag para determinar aluno avaliado em redação**	**INT**
110	**REDACAO_NOTA**	**Nota média de todas as competências obtidas em redação**	**FLOAT**
111	**REGIONAL_FORTALEZA**	**Divisão de regionais no município de Fortaleza**	**VARCHAR**
112	**REGIONAL_MACEIO**	**Divisão de regionais no município de Maceió**	**VARCHAR**
113	**RF_BBB_PPP**	**Onde "BBB" identifica o bloco e "PPP" a posição da questão no questionário**	**VARCHAR**

114	**RP_BBB_PPP**	**Onde "BBB" identifica o bloco e "PPP" a posição do item dentro do bloco no caderno**	**NVARCHAR**
115	**RPA_BBB_PPP**	**Onde "BBB" identifica o bloco e "PPP" a posição do item dentro do bloco no caderno**	**BIGINT**
116	**RQ_BBB_PPP**	**Onde "BBB" identifica o bloco e "PPP" a posição da questão**	**INT**
117	**SG_UF**	**UF/estado**	**VARCHAR**
118	**TP_DEFICIENCIA**	**Deficiência**	**VARCHAR**
119	**VL_ERRO_PRF_ALN_EE**	**Erro de proficiência por aluno em determinado ano ou edição "EE"**	**FLOAT**
120	**VL_ISE_ERRO_EE**	**Erro para índice sócio econômico por aluno em determinado ano ou edição**	**FLOAT**

		"EE"	
121	**VL_ISE_EE**	**Índice sócio econômico por aluno em determinado ano ou edição "EE"**	**FLOAT**
122	**VL_PERC_ACERTOS**	**Percentual de acertos**	**FLOAT**
123	**VL_PRF_ALN_YYY_EE**	**Proficiência do aluno em um determinado ano ou edição (EE). O dado “YYY” pode ser de leitura(LTR) ou escrita(ECT)**	**FLOAT**
124	**VL_PRF_ALN_EE**	**Proficiência do aluno em um determinado ano ou edição (EE)**	**FLOAT**