UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE SERRA TALHADA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Mineração de Dados Educacionais para a Classificação de Perfis de Evasão do Ensino Superior

Por

Rafael Gentil de Barros Santos

Serra Talhada,
Agosto/2022

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

RAFAEL GENTIL DE BARROS SANTOS

Mineração de Dados Educacionais para a Classificação de Perfis de Evasão do Ensino Superior

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

Orientadora: Ellen Polliana Ramos Souza

Serra Talhada,
Agosto/2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S237m Santos, Rafael
Mineração de Dados Educacionais para a Classificação de Perfis de Evasão do Ensino Superior / Rafael Santos. - 2022.
27 f. : il.

Orientadora: Ellen Polliana Ramos Souza.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Serra Talhada, 2022.

1. Mineração de dados. 2. Dados Educacionais. 3. Dados Abertos. I. Souza, Ellen Polliana Ramos, orient. II. Título

CDD 004

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

RAFAEL GENTIL DE BARROS SANTOS

Mineração de Dados Educacionais para a Classificação de Perfis de Evasão do Ensino Superior

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendida e aprovada por unanimidade em 19/08/2022 pela banca examinadora.

Banca Examinadora:

Ellen Polliana Ramos Souza
Orientador(a)
Universidade Federal Rural de Pernambuco

Prof. Paulo Mello da Silva
Universidade Federal Rural de Pernambuco

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

Dedico este trabalho aos meus Irmãos,

pois ergueram-me quando não conseguia alcançar o meu sonho.

"Um por todos, todos por um".

(Autor: Alexandre Dumas)

AGRADECIMENTOS

Quero agradecer primeiramente a Deus, por ter sido generoso comigo, permitindo que este trabalho pudesse ser concluído. Quero agradecer também as minhas irmãs, Ramiri Poliana e Rafaela Hosana, por acreditarem em mim e por terem se tornaram minhas grandes mães tão novas, pois sempre cuidaram e me incentivaram a buscar sempre o melhor e a nunca desistir. E sem elas eu sei que não teria sido capaz de levar adiante muitas coisas na minha vida. Obrigado por tudo.

Quero agradecer ainda aos meus amigos por terem me apoiado em muitos momentos da minha vida, principalmente meu amigo/cunhado que nunca desistiu de mim.

Por fim, gostaria de agradecer a minha orientadora Dra. Ellen Polliana Ramos Souza, pela paciência na orientação, por me incentivar a dar meu máximo e a não desistir apesar das dificuldades, por me dar uma direção em meio ao caos, por me ajudar a alcançar conhecimento necessário para resolver determinados problemas e a concluir este trabalho, a ela possuo uma eterna gratidão e admiração.

“Por mais inteligente que alguém possa ser, se não for humilde, o seu melhor se perde na arrogância. A humildade ainda é a parte mais bela da sabedoria”

(Autor: desconhecido.)

RESUMO

A evasão escolar é uma problemática que atinge diversas instituições e é considerada uma grande preocupação para empresários, diretores, pesquisadores, pais e alunos. As perdas ocasionadas pela evasão tanto no setor público como privado, causam ociosidade de professores, funcionários, equipamentos e espaço físico. Este trabalho tem como objetivo desenvolver de um comitê de classificador para realizar a predição dos discentes com possibilidade de evasão. O método CRISP-DM foi usado para entender, preparar e modelar os dados da solução. Para a preparação dos dados, foram utilizadas as ferramentas Pentaho e RapidMiner. A linguagem de programação Python foi utilizada para implementar o comitê de classificador. Como resultados, espera-se ajudar no entendimento do perfil dos discentes com a possibilidade de evasão e como esse fenômeno pode ser evitado pelos gestores das instituições.

Palavras-chave: Mineração de dados, Dados Educacionais, Dados Abertos.

ABSTRACT

School dropout is a problem that affects several institutions and is considered a major concern for entrepreneurs, directors, researchers, parents and students. The losses caused by evasion in both public and private sectors are a source of idleness for teachers, employees, equipment and physical space. This work aims to develop a classifier committee to perform the prediction of students with the possibility of dropping out. The CRISP-DM method will was used to understand, prepare and model the solution. For data preparation, the Pentaho and RapidMiner tools were used. The Python programming language was used to implement the classifier committee. As a result, it is expected to help in understanding the profile of students with the possibility of dropout and how this phenomenon can be avoided by the managers of the institutions.

Keywords: Data Mining, Educational Data, Open Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BDs	Bases de Dados
LAI	Lei de Acesso a Informação
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira
MEC	Ministério da Educação
MDE	Mineração de Dados Educacionais
CRISP-DM	Cross-Industry Standard Process for Data Mining
DCBD	Descoberta de Conhecimento em Base de Dados
NB	Naive Bayes
ML	Machine Learning

SUMÁRIO

1 Introdução

A evasão escolar é uma problemática que atinge diversas instituições, estando presente nos mais variados níveis de ensino, e é considerada uma grande preocupação para empresários, diretores, pesquisadores, pais e alunos (COLPANI, 2018). Acrescenta-se também que a evasão pode acarretar vários prejuízos às instituições de ensino, sendo para o setor público, os recursos investidos sem o devido retorno; para o setor privado, importante perda de receita; para ambos os setores, fonte de ociosidade de professores, funcionários, equipamentos e espaço físico, representando assim perda social, de recursos e de tempo de todos os envolvidos no processo de ensino (LOBO, 2012).

Com disponibilização de dados abertos, dados legíveis disponíveis passíveis de uso, reuso e redistribuição por qualquer pessoa, atendendo, apenas, à exigência da atribuição da fonte dos mesmos e o compartilhamento pelas mesmas regras, abrem-se inúmeras possibilidades, como a análise aprofundadas das informações públicas, porém existe poucas pesquisas no contexto de dados abertos e que estão avançando ao poucos. Segundo (SANTOS; FERREIRA; MIRANDA, 2017), ao realizar uma pesquisa em base de artigos publicados e revistas acadêmicas, percebeu que, apesar da ampla abertura existente para obter dados abertos e da grande diversidade de tecnologias disponíveis para a utilização de tais dados, ainda há poucas iniciativas brasileiras no campo referente aos dados educacionais. Além disso, apesar da busca ter sido realizada para artigos no intervalo do ano de 2008 a 2017, apenas em 2011 foi encontrado o primeiro artigo e que em 2015 a quantidade de publicações no tema atingiu o maior número, com 11 artigos.

No Brasil, o direito de cada cidadão ter acesso aos dados esta previsto na Lei Federal 12.527/2011, conhecida como Lei de Acesso a Informação (LAI) (BRASIL, 2011), tornou-se possível a verificação de dados públicos atreves da transparência ativa na qual o governo publica dados dos seus órgãos que sejam de interesse do cidadão, diante disso, a quantidade de dados disponibilizados tem crescido consideravelmente. O Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira (INEP) é uma autarquia federal vinculada ao Ministério da Educação (MEC) responsável por gerar dados abertos referentes ao desempenho dos estudantes de instituições de ensino fundamental, médio e superior, públicas e privadas.

Neste sentindo, este trabalho teve como objetivo geral desenvolver um comitê classi-

ficador para a predição de perfis dos discentes com possibilidade de evasão utilizando dados históricos do Censo da Educação Superior. A justificativa para o uso de Mineração de Dados Educacionais (MDE) se dar por uma ampla quantidade dados educacionais abertos, no entanto, a forma bruta em que esses dados ainda são publicados e sua quantidade, torna-se humanamente impossíveis de serem analisados de forma manual. Deste modo fazer-se necessário o uso de técnicas para a descoberta de conhecimento automático, como é o caso de Mineração de Dados, adequando algoritmos existentes para área de conhecimento, beneficiando-se ainda mais da utilização apropriada de análise de dados (COSTA et al., 2013).

Este trabalho está organizado da seguinte maneira: o referencial teórico é apresentado na Seção 2. O método esta descrito na Seção 3. Os resultados esperados são relatados e discutidos na Seção 4.

2 Referencial Teórico

2.1 Dados Governamentais Abertos

É um termo usado para informações governamentais e dados de domínio público disponíveis na Internet para uso gratuito da sociedade. Este conceito refere-se à proteção dos dados públicos pertencentes aos cidadãos, que devem ter acesso irrestrito às informações governamentais. (VAZ; RIBEIRO; MATHEUS, 2010).

O estabelecimento de dados governamentais ocorreu em 2007, quando 30 defensores do governo aberto se reuniram na Califórnia e desenvolveram as três leis apontadas por (EAVES, 2009):

1. Se o dado não pode ser encontrado e indexado na Web, ele não existe;

2. Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado; e

3. Se algum órgão legal não permitir sua replicação, ele não é útil.

A Administração Pública está sempre em busca de se mostrar transparente dado o princípio constitucional da publicidade de seus atos e atuações diante da sociedade, com a promoção de políticas e investimentos de dados abertos para maior controle e participação social, auxiliando assim nos seguintes aspectos (MAGALHÃES; CARDOSO, 2016):

1. Realização de novos estudos no aspecto de inovação;

2. Realização de atividades e usos pelo próprio Governo e gerador desses dados, tendo como exemplo o INEP;

3. Realização de estudos pelos cidadãos, sejam eles estudantes universitários, pesquisadores, programadores, desenvolvedores, professores e outros.

2.2 Mineração de Dados Educacionais

A Mineração de Dados consiste em extrair ou "minerar" conhecimento a partir de uma grande quantidade de dados. É considerada como uma das principais etapa descoberta de

conhecimento em base de dados (KDD- Knowledge Discovery in Database), podendo ser dividido em seleção, pre-processamento, transformação, mineração de dados e interpretação e avaliação. (COSTA et al., 2013) afirmam que, em KDD, ainda há existência de duas grandes etapas: pré-processamento de dados, responsável pela preparação de dados, abrangendo mecanismos para captação, organização e tratamento dos dados e pós-processamento dos resultados obtidos na mineração de dados, responsável por verificar até que ponto os padrões encontrados contribuem na solução do problema inicialmente identificado.

A Mineração de Dados Educacionais (MDE) busca utilizar e/ou adaptar métodos e algoritmos de mineração de dados já existentes na literatura para encontrar a melhor forma de compreender os dados produzidos pela educação. A Mineração de Dados pode auxiliar a entender informações ocultas em base de dados, entretanto, na maioria das vezes, há a necessidade de adaptá-las devido às particularidades existentes em ambientes educacionais e seus dados (BRAVIN; LEE; RISSINO, 2019). Na MDE, surgiram varias subáreas e linhas de pesquisas derivadas diretamente da mineração de dado (BAKER; ISOTANI; CARVALHO, 2011).

2.3 Trabalhos Relacionados

O trabalho de (COLPANI, 2018) buscou realizar uma análise correlacional dos indicadores do Censo Escolar de 2017, no ensino médio, verificando assim, quais estavam relacionados com a evasão e propôs um modelo preditivo. Foi adotado, como método, processo padrão da indústria cruzada para mineração de dados (CRISP-DM -Cross-Industry Standard Process for Data Mining), aplicando técnicas de correlação e regressão linear. Para a validação, foi utilizado a métrica Raiz do Erro Médio Quadrático.

Como resultado, foi observado que a medida que a taxa de distorção idade-série aumenta, maior é a taxa de evasão, que as variáveis "Média de Alunos por Turma"e Média de "Horas-Aula Diária"apresentam uma fraca correlação negativa e que que não a uma correção entre as variáveis "Percentual de Docentes com Curso Superior"e "Taxa de Reprovação"com a taxa evasão. O modelo de regressão obteve um coeficiente de determinação de 0.33, significando que variação da taxa evasão pode ser explicado pela variação da taxa de distorção idade-série.

Uma das limitações apontadas foi a discrepância na base, o que dificultou na aplicação do modelo de regressão linear. Como trabalhos futuros, o autor pretende aplicar métodos robustos com o intuito de analisar o comportamento de dados discrepantes, explorar outras granularidades

de escolaridade e realizar uma análise histórica desses dados.

(MANHÃES; CRUZ, 2019) propuseram uma solução para prever o desempenho acadêmico dos estudantes de graduação e identificar aqueles que estão em risco de evadir do sistema de ensino, auxiliando assim os gestores acadêmicos. Como método, foi utilizado o EDM WAVE e realizado um estudo de caso, comparando o desempenho de 12 algoritmos classificadores a fim de obter o melhor conjunto preditivo.

Foi observado que os algoritmos SimpleLogistic, Decision trees, BayesNet e Naive-Bayes apresentaram maior homogeneidade entre as acurácias nos experimentos. O algoritmo Naive Bayes (NB) ainda obteve acurácia superior a 80% e apresentou um modelo de predição mais interpretável.

Limitações encontradas neste trabalho foram a de obter permissão de acesso aos dados e alteração do sistema SGA da UFRJ e compreender o modelo de dados do SIGA. Como trabalhos futuros, os autores propuseram ampliar os estudos experimentais, avaliar a arquitetura EDM WAVE para cursos na modalidade on-line ou EAD e incorporar alertas e relatórios para os próprios estudantes.

(CARVALHO; CRUZ; GOUVEIA, 2017) realizaram uma Mineração de Dados Educacionais (EDM) para a previsão do desempenho acadêmico de alunos e investigaram e diagnosticaram as deficiências e obstáculos existentes no ensinos fundamental, médio e superior do estado Pernambuco.

O método, utilizado foi o KDD, aplicando os algoritmos de Aprendizado Supervisionado: Árvore de Decisão e Classificação Bayesiana, utilizando a ferramenta WEKA na base de dados dos Censos da Educação Básica e Superior.

Como resultado, obtiveram em relação ao cenário "Perfis das Escolas quanto à Infraestrutura", que a classe Federal não se ajustou ao conjunto de teste (ano 2015) e foi constatada a ocorrência do fenômeno denominado de *overfitting*, caracterizado pelo ajuste em excesso para o conjunto de treinamento das hipóteses induzidas pelo algoritmo. Já no cenário "Perfis dos Alunos do Ensino Básico", a aplicação do algoritmo de árvore de decisão obteve uma taxa de acurácia de 63%. No cenário "Situações dos Alunos dos Cursos de TIC", foi visto a ocorrência do fenômeno chamado de *underfitting*, caracterizado por um aprendizado insatisfatório pelo modelo gerado, com o algoritmo de árvore de decisão J48. E no cenário "Perfis dos Alunos do Ensino Superior", o algoritmo Naive Bayes gerou o modelo com a maior taxa de acurácia, em torno de 77%.

Como trabalhos futuros, os autores pretendem aplicar análises mais profundas das regras que foram geradas no que se refere aos padrões que justifiquem as dificuldades presentes

na educação.

(CALIXTO; SEGUNDO; GUSMãO, 2017) realizaram uma Mineração de Dados Educacionais (EDM), com a finalidade de identificar as variáveis referentes à evasão escolar e comparar os dados relativos aos estados do Ceará e Sergipe. O método utilizado foi CRISP-DM. Foram aplicadas técnicas de Regressão Linear e Indução a Regra na base de dados do Censo Escolar do INEP, entre os anos de 2014 e 2016.

No estado de Sergipe, percebeu-se que, a cada ano, torna-se maior a probabilidade do aluno evadir; Alunos com 14 e 15 anos apresentam menor chance de evadir; O primeiro e segundo ano do ensino médio apresentam baixo índice de evasão, como também, escolas com ausência de salas de atendimento especiais ou laboratórios apresentaram alto índice de evasão. Em comparação, no estado do Ceará: estudantes com idade menor ou igual a 18 ou 19 anos apresentam menores índices de evasão; Escolas com dependência administrativa municipal possuem menor taxa de evasão e, por fim, também foram notados altos índices de evasão em escolas com ausência de salas de atendimento ou laboratórios.

2.3.1 Análise Comparativa

De modo a facilitar o entendimento de trabalhos relacionados em comparação com este trabalho apresentado, a Tabela 2.1 apresenta um resumo comparativo. Neste resumo, são expostos os trabalhos relacionados segundo suas proposta de bases de dados, algoritmos, tecnologias abordadas e métodos empregados.

Cada um dos trabalhos utilizou tecnologias bem distintas, sendo que, as que mostraram melhor resultado e praticidade foram utilizadas neste trabalho, como a ferramenta RapidMiner e a linguagem python. Os algoritmos utilizados nos trabalhos de (COLPANI, 2018), (CARVALHO; CRUZ; GOUVEIA, 2017) e (CALIXTO; SEGUNDO; GUSMãO, 2017) foram os principais algoritmos a serem testados para compor o comitê de classificador por já terem sido aplicados em base de dados abertas como o presente trabalho e tiveram um bom resultado. O dicionário de dados disponibilizado pelos autores (COLPANI, 2018) e (MAGALHÃES; CARDOSO, 2016) facilita no entendimento das variáveis utilizadas. A integralização de bases de dados foi utilizada apenas pelo autor (CARVALHO; CRUZ; GOUVEIA, 2017).

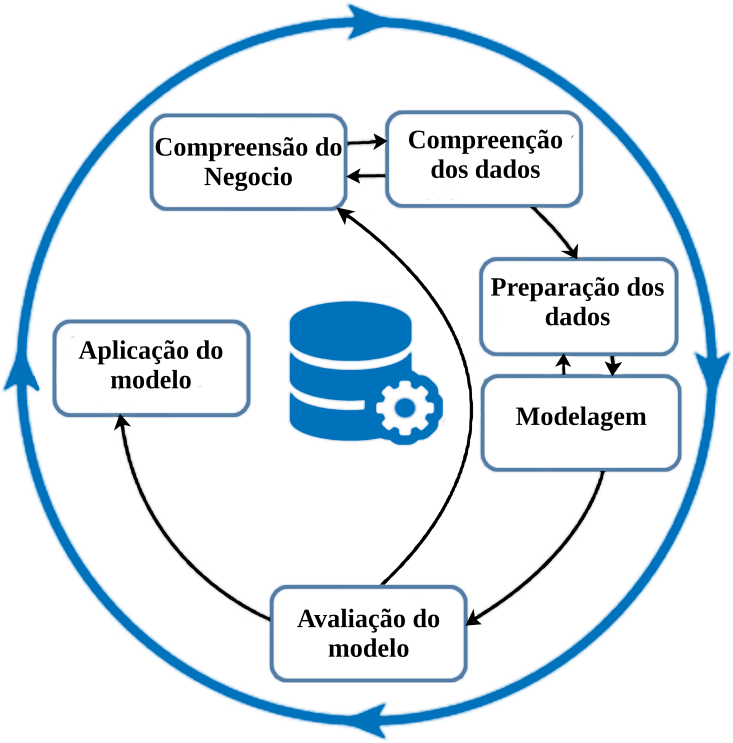
Tabela 2.1 – Análise Comparativa dos Trabalhos Relacionados

	[Colpani 2018]	[Manhães and da Cruz]	[Carvalho et al. 2017]	[Calixto et al. 2017
Validação	Coeficiente de determinação Raiz do Erro Médio Quadrático	Acurácia Validação Cruzada Matrix de Confunsão Kappa	Taxa F-measure TP Rate Acuracia Desempenho Matriz de confusão (Recall) Precisão	Acurácia
Base dados	Censo da Educação Básica	SGA da UFRJ (SIGA)	Censo da Educação Básica Censo da Educação Superio	Censo da Educação Básica
Algoritmos	regressão linea análise de correlação	BayesNet Decision trees C4.5 RIPPER SVM with Poly Kernel (SVM1) SVM with RBF Kernel (SVM2) Multilayered Perceptrons AdaBoost; NaiveBayes SimpleLogistic OneR RandomForest DecisionTable	Naive Bayes J48	Regressão Logística Indução de Regra
Integração de bases			X	
Dicionário de dados	X	X		
Metodo	CRISP-DM	EDM WAVE	KDD	CRISP-DM
Tecnologias	Python		WEKA	R IBM SPSS Statistics PostgreSQL RapidMiner

3 Método

O método utilizado nesta trabalho foi o CRISP-DM, por ser um dos mais utilizados e aceitos em várias aplicações de Descoberta de Conhecimento em Base de Dados (DCBD), além de apresentar um conjunto bem definido, não rígidas de seis fases, que permite a construção e implementação de um modelo de mineração para ser usado em um ambiente real (COLPANI, 2018), como mostra a Figura 3.1.

Figura 3.1 – Fases do modelo de referência CRISP-DM



Fonte: Adaptado de (CHAPMAN et al., 2000)

O ciclo externo na figura representa o ciclo natural da mineração de dados, pois o processo de mineração de dados continua após a solução ter sido desenvolvida, como por exemplo com a validação e atualização dos modelos desenvolvidos. As próximas subseções detalham cada uma das fases, as quais estão estruturadas em várias tarefas gerais e específicas, onde são definidas as ações que são desenvolvidas para situações específicas.

3.1 Compreensão do Negócio

Esta é uma das fases mais importantes do processo, pois é nela que é realizada o entendimento do contexto educacional com o objetivo de identificar e descrever os principais problemas educacionais. Ou seja, identificar as dimensões dos fatores associados aos problemas e suas principais causas. Para esse trabalho, foram analisados materiais já elaborados na literatura, relacionados aos cenários educacionais, mineração de dados educacionais e evasão. Em seguida, os objetivos específicos a serem alcançados foram formulados, sendo eles:

(1) mapear os atributos utilizados na análise de correlação da evasão do ensino superior brasileiro e

(2) avaliar desempenho de algoritmos de classificação (ML) na predição da evasão.

3.2 Entendimento dos Dados

A fase de entendimento de dados envolve um olhar detalhado dos dados disponíveis, verificando as variáveis disponíveis e se as instâncias possuem valores discrepantes, valores ausentes, dados duplicados ou dados errados. Esse passo é essencial para evitar problemas inesperados durante a etapa seguinte. Nesta fase, é possível analisar os fatores e suas dimensões e os indicadores que serão medidos dependendo do contexto educacional, na qual os dados são coletados e organizado todos os dados que se encontram disponível para realizar a análise exploratória. Esta fase começa com a coleta de dados, e prossegue com atividades que permitem familiarizar-se, identificar problemas de qualidade, *insights* sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

A base de dados utilizada nesse estudo é o Censo da Educação Superior, disponibilizados anualmente pelo Inep em seu portal, desde 1995 (INEP, 2020). Foram utilizados os dados dos indicadores educacionais referentes aos anos de 2009 a 2018. Para entendimentos dos dados, foi estudado o dicionário de dados que é disponibilizado juntamente com a base para melhor entendimento. Com isso, foi possível perceber que algumas variáveis mudaram ao longo dos anos, como as variáveis descritivas, deixando apenas os códigos que as representam e outras deixaram de ser uma única variável para serem quebradas em mais categorias, como o tipo de deficiência.

3.3 Preparação dos Dados

A terceira fase do CRISP-DM é uma das mais importantes e frequentemente demanda de mais tempo, chegando normalmente a consumir em media de 50% a 70% do tempo empregado no projeto, sendo empregadas as tarefas de seleção de atributos, limpeza de dados e transformação para as ferramentas de modelagem.

A primeira tarefa utilizada foi seleção de dados, que compreende na identificação das informações, dentre as bases de dados existentes, consideradas durante o processo. Assim, nesse estudo foi empregada a ferramenta Pentaho para efetuar a filtragem e a integração dos dados, utilizando os atributos chaves, sendo selecionados os dados pelos critérios relevância para análise da evasão.

Após o processo de filtragem, foi executada a segunda tarefa, limpeza dos dados, este processo geralmente envolve filtrar e preencher valores vazios. Para filtrar os dados foi utilizado o o *preview* do Pentaho Data Integration, tendo uma visão geral das colunas, sendo possível verificar quais colunas não possuem grande parte de suas instancias *missing values*, valores nulos, sendo necessário filtrar e eliminar as que não há exemplo nos dados, ou simplesmente as que a quantidade é insuficiente para o aprendizado do modelo. Em seguida, foi executada o preenchimento de valores nulos, utilizando a ferramenta RapidMiner foram tratados os valores nulos que restaram, para que o número de *missing values* reduza o máximo possível, para que não atrapalhe o desenvolvimento do algoritmo, tendo assim o mínimo de perda de instâncias.

Na terceira tarefa, transformação dos dados, utilizando a ferramenta Pentaho, foi necessário corrigir erro no esquema de codificação de caracteres que vem nos arquivos CSVs disponibilizados no INEP, sendo necessário realizar a mudança de utf-8 para iso-8859-1, para que a ferramenta conseguisse reconhecer todos os caracteres e tratar os valores para retirada de acentos, caracteres especiais, colocar em caixa e criar intervalos de valores, como faixa de idades. Também será utilizado regras para padronizar variáveis, como por exemplo o sexo que pode esta representado binário, 0 e 1, ou escrito por extenso. Nessa fase também, foi ajustada a escala dos valores dos atributos para que os valores fiquem em pequenos intervalos, tais como entre 0 e 1. Tal ajuste se faz necessário para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa na mineração de dados.

Com a data que o senso foi realizado e o ano em que o aluno ingressou é possível criar a variável tempo de permanência.

3.4 Modelagem

Considerada como a principal etapa do processo, nessa fase será decidida qual a melhor técnica de Data Mining a ser utilizada com base nos objetos identificados na primeira fase. Representa o desenvolvimento dos modelos para o problema, baseado nos dados que já estão adequados para serem utilizados. Nesse trabalho, será utilizado a técnica de aprendizagem de máquina, no qual será desenvolvido um comitê de classificador, englobando os algoritmos Random Forest, Naive Bayes, Decicion Tree, Regressão linear e Regressão logística.

Será realizada a análise correlacional entre as variáveis, avaliando assim o grau de relacionamento entre as variáveis das bases de dados, ou seja, descobrir o quanto uma variável (x) interfere no resultado de outra (y). Além disso o uso de indução de regra que fornece regras de um conjunto de dados, encontrando pedrões que são posteriormente expressos em regras se-então, e que é apresentado de acordo com a sua frequência. Para que dessa forma possa conhecer um pouco melhor o comportamento das variáveis.

3.5 Avaliação do Modelo

Os principais critérios de avaliação utilizados serão: (i) as acurácias avaliadas individualmente; (ii) particionamento das bases de dados utilizando a validação cruzada com k conjuntos (k-fold cross validation), dois conjuntos de dados: treinamento e teste e (iii) as medidas estatísticas calculadas a partir da matriz de confusão: taxa de acerto da classe positiva (verdadeiro positivo), taxa de acerto da classe negativa (verdadeiro negativo).

3.6 Aplicação do modelo

Todo o conhecimento obtido através do trabalho de mineração tornou-se subsídio para o desenvolvimento de estratégias que resolvam o problema proposto. Nesse trabalho, será aplicado esse modelo na base de dados do Censo da Educação Superior e identificar os perfis dos alunos que podem evadir.

4 Resultados

Este trabalho teve como objetivo apresentar um projeto para construção de uma aplicação de mineração de dados para analisar a evasão no ensino superior. Para isto, foram elencados os principais trabalhos na área de mineração de dados educacionais, bem as ferramentas, técnicas e algoritmos utilizados. O modelo proposto, baseado em um comitê de classificador, permitirá identificar alunos que poderão evadir de acordo com um determinado nível de acurácia.

REFERÊNCIAS

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03, 2011.

BRASIL. *LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011.Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências*. 2011. Http://www.planalto.gov.br/ccivil$_0 3/_a to 2011 - 2014/2011/Lei/L12527.htm.$ $ltimoacessoem 15set 2020.$

BRAVIN, G. F.; LEE, L.; RISSINO, S. das D. Mineração de dados educacionais na base de dados do enem 2015. *Brazilian Journal of Production Engineering-BJPE*, p. 186–201, 2019.

CALIXTO, K.; SEGUNDO, C.; GUSMãO, R. P. de. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1447.

CARVALHO, J.; CRUZ, L.; GOUVEIA, R. Descoberta de conhecimento com aprendizado de máquina supervisionado em dados abertos dos censos da educaçao básica e superior. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, n. 1, p. 674.

CHAPMAN, P. et al. Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, v. 9, p. 13, 2000.

COLPANI, R. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. *Informática na educação: teoria & prática*, v. 21, n. 3, 2018.

COSTA, E. et al. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, v. 1, n. 1, p. 1–29, 2013.

EAVES, D. The three laws of open government data. *Eaves. ca*, v. 30, n. 8, 2009.

INEP. *Microdados*. 2020. Http://portal.inep.gov.br/web/guest/microdados
Último acesso em 06 out 2020.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, v. 25, 2012.

MAGALHÃES, H. d. F.; CARDOSO, L. d. A. Análise de dados abertos sobre o ensino superior brasileiro. 2016.

MANHÃES, L. M. B.; CRUZ, S. M. S. da. PREDIÇÃO DO DESEMPENHO ACADÊMICO DE ALUNOS DA GRADUAÇÃO UTILIZANDO MINERAÇÃO DE DADOS. p. 1–15, 2019.

SANTOS, P.; FERREIRA, R.; MIRANDA, P. Dados abertos educacionais: Uma revisao da literatura brasileira. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 11.

VAZ, J. C.; RIBEIRO, M. M.; MATHEUS, R. Dados governamentais abertos e seus impactos sobre os conceitos e práticas de transparência no brasil. *Cadernos ppg-au/ufba*, v. 9, n. 1, 2010.