UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE SERRA TALHADA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Aplicação de Técnicas de Mineração de Dados Educacionais para Previsão do Desempenho de Estudantes a partir de dados do ENADE

Por

Thacyo Eslley Burgo de Lima

Serra Talhada,
Março/2021

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

THACYO ESLLEY BURGO DE LIMA

Aplicação de Técnicas de Mineração de Dados Educacionais para Previsão do Desempenho de Estudantes a partir de dados do ENADE

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

Orientador: Prof. Paulo Mello da Silva

Serra Talhada,
Março/2021

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)

T363a Lima, Thacyo Eslley Burgo de
Aplicação de Técnicas de Mineração de Dados Educacionais para Previsão do Desempenho de Estudantes a partir de dados do ENADE / Thacyo Eslley Burgo de Lima. - 2021.
20 f. : il.

Orientador: Paulo Mello da Silva.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Serra Talhada, 2021.

1. Mineração de dados. 2. Sistema de informação. 3. ENADE. 4. Mineração de dados educacionais. I. Silva, Paulo Mello da, orient. II. Título

CDD 004

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

THACYO ESLLEY BURGO DE LIMA

Aplicação de Técnicas de Mineração de Dados Educacionais para Previsão do Desempenho de Estudantes a partir de dados do ENADE

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendido e aprovado por unanimidade em 04/03/2021 pela banca examinadora.

Banca Examinadora:

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

Prof. Hidelberg Oliveira Albuquerque
Universidade Federal Rural de Pernambuco

Prof. Maximiliano Carneiro da Cunha
Universidade Federal Rural de Pernambuco

AGRADECIMENTOS

Agradeço, primeiramente, a Deus e a Virgem Maria, que me deram forças e sabedoria para concluir este trabalho.

Agradeço a minha família e a minha noiva pelo apoio e compreensão.

Agradeço a Unidade Academica de Serra Talhada-UAST, meus professores e colegas do curso de Sistemas de Informação e de forma especial ao meu orientador.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Aplicação de Técnicas de Mineração de Dados Educacionais para Previsão do Desempenho de Estudantes a partir de dados do ENADE

Thacyo Eslley Burgo de Lima[1], **Paulo Mello da Silva**[1]

[1]Unidade Acadêmica de Serra Talhada - Universidade Federal Rural de Pernambuco (UFRPE)
Av. Avenida Gregório Ferraz Nogueira, S/N,
José Tomé de Souza Ramos, 56909-535 - Serra Talhada - PE

thacyo.burgo@ufrpe.br, paulomellosilva2@gmail.com

Abstract. *With the increasing growth in data, it is necessary to apply techniques to extract information and knowledge from the set of data that are available, with that comes data mining. This work aims to predict the performance of students in Information System courses from data from the National Student Performance Exam (Enade) held in the year 2017. For this, machine learning algorithms were used to discover knowledge and assist in decision making. Four algorithms were used for comparison that obtained accuracy greater than 60%, showing that it is feasible to make the forecast.*

Resumo. *Com o crescimento cada vez maior nos dados, faz-se necessário aplicação de técnicas para extrair informações e conhecimentos a partir do conjunto de dados que estão disponibilizados, com isso encontra-se a Mineração de Dados. Esse trabalho tem como objetivo prever o desempenho dos alunos dos cursos de Sistema de Informação a partir dos dados do Exame Nacional de Desempenho de Estudantes (Enade) realizado no ano de 2017. Para isto, foram utilizados algoritmos de aprendizagem de máquina para a descoberta do conhecimento e auxiliar na tomada de decisão. Foram utilizados 4 algoritmos para comparação, os quais obtiveram acurácia superior a 60%, mostrando que é viável realizar a predição.*

1. Introdução

Com o advento da Lei de Acesso à Informação (LAI), Lei Federal nº 12.527 [BRASIL 2011], as organizações públicas têm disponibilizados seus dados na Web para acesso livremente a todas as informações, de forma a garantir a transparência no serviço público. Para Isotani e Bittencourt (2015), consiste em dados que podem ser livremente acessados, utilizados, modificados e compartilhados para qualquer finalidade, estando sujeito a, no máximo, exigências que visem preservar sua proveniência e sua abertura [Isotani and Bittencourt 2015].

Diante do crescimento cada vez maior nos dados, não é possível de forma manual extrair informações e conhecimentos a partir do aglomerado de dados que estão disponibilizados, por isso faz-se necessário aplicação de técnicas computacionais, daí encontra-se a Mineração de Dados. Para [Côrtes et al. 2002] a Mineração de Dados é parte de um processo maior de pesquisa denominado Busca de Conhecimento em Banco de Dados (Knowledge Discovery in Database - KDD), o qual possui uma metodologia própria para

preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados.

O objetivo deste estudo é aplicar técnicas de Mineração de Dados Educacionais (EDM) para prever o desempenho dos alunos dos cursos de Sistema de Informação a partir dos resultados do Exame Nacional de Desempenho dos Estudantes (Enade), como também traçar o perfil dos alunos de Sistemas de Informação que realizaram o exame e identificar os fatores que influenciam o seu desempenho.

A partir dos resultados deste trabalho será possível identificar quais os fatores que influenciam no desempenho final dos alunos, e a partir disso seria possível propor estratégias que minimizassem a ocorrência do baixo desempenho dos alunos de SI. Como também identificar o perfil dos alunos que não obtiveram bons resultados para que este aluno tenha um melhor acompanhamento durante o curso. Segundo [Júnior et al. 2017]: essa possível detecção antecipada poderia fornecer informações que permitissem a tomada de decisões por gestores acadêmicos (coordenadores de curso, diretores de ensino, entre outros) para modificar essa predição detectada.

O artigo está organizado da seguinte maneira: Na Seção 2 apresenta-se o referencial teórico; na Seção 3 descreve-se como será o método utilizado, como os dados foram obtidos e o pré-processamento realizado nos mesmos; Os resultados são apresentados na Seção 4 e na Seção 5 são apresentadas as considerações finais.

2. Referencial Teórico

2.1. Exame Nacional de Desempenho dos Estudantes (Enade)

O Sistema Nacional de Avaliação da Educação Superior (Sinaes), criado pela Lei nº. 10.861, de 14 de abril de 2004, é formado por três componentes principais: a avaliação das instituições, dos cursos e do desempenho dos estudantes. Os resultados das avaliações possibilitam traçar um panorama da qualidade dos cursos e instituições de educação superior no país. Os processos avaliativos são coordenados e supervisionados pela Comissão Nacional de Avaliação da Educação Superior (Conaes) e a operacionalização é de responsabilidade do Inep.

O Exame Nacional de Desempenho dos Estudantes (Enade) é uma das avaliações que compõem o Sistema Nacional de Avaliação da Educação Superior (Sinaes). O objetivo do Enade é avaliar e acompanhar o processo de aprendizagem e o desempenho acadêmico dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação; suas habilidades para ajustamento às exigências decorrentes da evolução do conhecimento e competências para compreender temas exteriores ao âmbito específico da profissão escolhida, ligados à realidade brasileira e mundial e a outras áreas do conhecimento. [BRASIL 2019]

A cada ano o Enade é aplicado a um conjunto de áreas de ensino, o ciclo de todas as áreas é fechado a cada três anos. O ano I abrange as áreas da engenharia, arquitetura, urbanismo, ciências da saúde, produção alimentícia, ciências agrárias e afins. Ano II é formado pelas áreas de Ciências Biológicas, Ciências Exatas, licenciaturas e áreas afins e o ano III é composto pelas áreas de ciências sociais aplicadas, ciências humanas e áreas afins. Desta forma cada conjunto de área é avaliado a cada 3 anos.

A prova é composta pelo componente de Formação Geral (FG) que possui 10 (dez) questões, sendo 02 (duas) discursivas e 8 (oito) de múltipla escolha e pelos Componentes Específicos (CE) de cada área de avaliação que é composta de 30 (trinta) questões, sendo 3 (três) discursivas e 27 (vinte e sete) de múltipla escolha. Ambos envolvendo situações-problema e estudos de caso. O componente de Formação Geral possui peso de 25% na composição da nota final e o Componente Específico com peso de 75%.

Além da prova, tem o Questionário do Estudante que é destinado a levantar informações que permitam caracterizar o perfil dos estudantes e o contexto de seus processos formativos e o Questionário de Percepção de Prova que se destina a levantar informações que permitam aferir a percepção dos estudantes em relação à prova. Ambos auxiliando na compreensão dos resultados dos estudantes no Enade.

O Enade é componente curricular obrigatório dos cursos de graduação, a participação do estudante habilitado ao Enade é condição indispensável ao seu registro da regularidade no histórico escolar, assim como à expedição do diploma pela Instituição de Educação Superior (IES).

O Conceito do Enade é calculado para cada curso, tendo como unidade de observação a instituição de ensino superior – IES, o município da sede do curso e a área de avaliação e é apresentado em cinco categorias (1 a 5), sendo que 1 é o resultado mais baixo e 5 é o melhor resultado possível, na área.

Os resultados são publicados no Portal do Inep, no menu correspondente ao Enade, que estão disponíveis todos os relatórios produzidos a partir da primeira aplicação do Exame, como: Boletim de Desempenho do Estudante, Relatório do Curso, Relatório Síntese de Área, Relatório da Instituição e Resumo Técnico. Esses relatórios e resultados deverão contribuir para o aperfeiçoamento dos processos de ensino aprendizagem e das condições de ensino e do próprio sistema de avaliação dos cursos de graduação.

Então com todos esses dados disponibilizados pelo Inep faz-se necessário a aplicação de técnicas computacionais para melhor entender e obter informações a partir dessa quantidade imensa de dados que são gerados anualmente pelo Inep, a partir disso podemos citar a Mineração de Dados Educacionais.

2.2. Mineração de Dados Educacionais

A Mineração de dados constitui-se como uma etapa de um processo maior de descoberta de informações denominado de Knowledge Discovery in Database (KDD), ou Busca de Conhecimento em Banco de Dados [Côrtes et al. 2002], nota-se que a Mineração de Dados é apenas uma etapa no processo de descoberta de conhecimento em base da dados, onde é responsável pela aplicação de algoritmos que irão fazer extração nos dados e posteriormente serão convertidas em informações.

Outro conceito importante no tocante a dados e informações é o de Big Data, cujo significado contemporâneo mais comum alude a um massivo conjunto de dados armazenados, que podem advir da Internet, dos mais variados dispositivos tecnológicos como Smartphones, Notebooks, Tablets, Computador Pessoal (PC), Internet das Coisas – trata de dispositivos como TVs, geladeiras, carros, relógios, óculos, e demais objetos do cotidiano que possuem a capacidade de estarem conectados à Internet; entre outros [Patricio and Magnoni 2018].

Técnicas específicas para manipulação de big data, já bastante utilizadas na área de negócios, têm começado a serem utilizadas no campo educacional na tentativa de entender o comportamento e os interesses dos estudantes e os fatores que podem levá-los a níveis maiores de engajamento [Scaico et al. 2014].

A Mineração de Dados Educacionais (do inglês, "Educational Data Mining", ou EDM) é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem [Baker et al. 2011].

O principal objetivo da EDM é estudado há muito tempo, que é conseguir compreender como ocorre o processo da aprendizagem. A diferença é que agora os pesquisadores possuem uma grande escala de dados e conseguem analisar uma aprendizagem prática e real (sem a antiga necessidade de realizar experimentos para obter dados) [Souza 2017].

Muitos dados educacionais estão disponíveis para utilização, como as avaliações que são realizadas pelo Ministério da Educação no âmbito Federal, como em cada instituição educacional com os sistemas de informação, como também nos cursos à distância com a utilização dos Ambientes Virtuais de Aprendizagem (AVA). Como os autores citaram que a utilização crescente de AVA no processo de ensino-aprendizagem proporciona a geração de um grande volume de dados, que envolve interações dos alunos e docentes com os ambientes de aprendizagem [Costa et al. 2012].

A utilização das técnicas de Mineração de Dados no ambiente educacional é de grande valia para a comunidade acadêmica e contribuições no processo de ensino-aprendizagem. Como frisou os autores, que a utilização do Big Data no setor da educação é deveras importante para capturar informações sobre os alunos, assim como suas interações com os ambientes de aprendizagem, conteúdos educacionais, avaliações, e assim favorecer para que haja melhorias no processo de aprendizagem, auxiliando gestores, educadores e formuladores de políticas educacionais [Scaico et al. 2014].

O processo de descoberta de conhecimento educacional pode variar de acordo com diferentes pontos de vista, mas de forma didática, pode ser representado por um ciclo iterativo [Romero 2011]. Como pode ser observado na Figura 1 que o processo de mineração de dados educacional não é apenas para transformar dados em conhecimento, mas também em modificar o ambiente educacional para melhorar a aprendizagem do aluno de forma contínua.

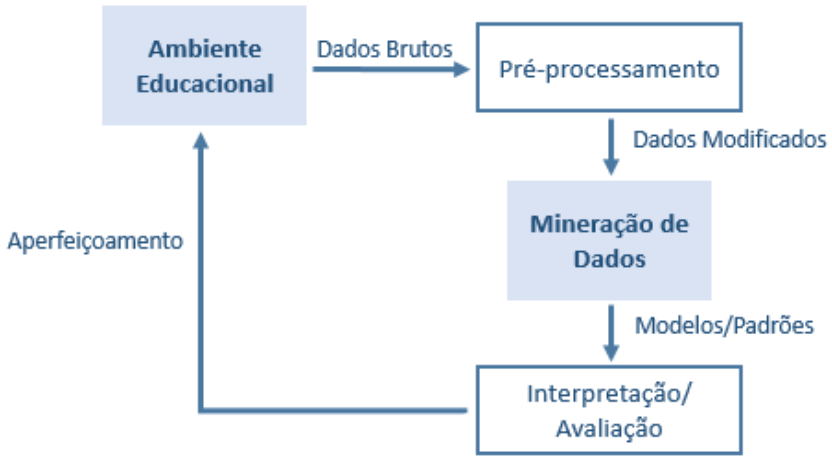

Figura 1. Processo de Mineração de Dados Educacionais
[Romero and Ventura 2013]

2.3. Trabalhos Relacionados

Alguns autores realizaram trabalhos semelhantes a este como: [CRETTON and GOMES 2016] utilizaram de técnicas de Mineração de dados com o uso do algoritmo J48 para extrair conhecimento a partir da base de dados do Enade, especificamente dos cursos de medicina no ano de 2013, foram levadas em consideração a idade e o sexo dos estudantes, juntamente com suas respectivas notas no componente específico do exame e suas respostas sobre o grau de dificuldade desta parte da prova, além das informações sobre as instituições que estes frequentam.

[Brito et al. 2014] propuseram a utilização de técnicas de Mineração de Dados para tentar predizer o desempenho dos alunos no primeiro período do curso de Ciência da Computação da UFPB, através das suas notas de ingresso no vestibular. Os resultados mostraram que é possível inferir o desempenho dos estudantes com uma acurácia superior a 70%, sendo esta informação útil para a realização de ações para evitar a evasão, aprimorando o sistema de ensino.

[Oliveira et al. 2018] analisaram o desempenho dos alunos nas disciplinas de Português e Matemática em duas escolas de nível médio. Analisaram a correlação entre a nota na avaliação final e as demais variáveis existentes e as variáveis: número de falhas de classe anteriores e tempo de estudo semanal, foram as que obtiveram maior correlação. O algoritmo utilizado foi o REPTree que gerou árvores com poucos nodos e maior capacidade preditiva.

[Filho et al. 2020] procuraram identificar associações das variáveis socioeconômicas com o desempenho dos estudantes de licenciatura no Enade. Realizaram análise de desempenho dos estudantes de Educação Física no Enade 2017. Os resultados apontaram um maior desempenho por parte dos estudantes que possuem renda, ainda recebem ajuda da família ou de outras pessoas para financiar os gastos durante a formação e apontou também que os maiores desempenho referem-se aos estudantes que os pais tiveram um maior nível de escolaridade.

O trabalho de [da Fonseca and Namen 2016] objetivou identificar o perfil dos pro-

fessores que pudessem influenciar o processo de ensino-aprendizagem de Matemática dos seus alunos, foi utilizado a base de dados correspondente à Prova Brasil 2011 do estado do Rio de Janeiro, o presente trabalho deu enfoque aos dados relacionados ao questionário dos professores que lecionam Matemática para alunos do 9º ano do Ensino Fundamental, juntamente com os dados relacionados à proficiência dos seus alunos. Foi aplicado o algoritmo de mineração de dados denominado Naïve Bayes.

[Francelino and Machado 2020] aplicaram técnicas de mineração de dados na base do Enade referente ao ano de 2013 para os cursos de Ciência da Computação, através do algoritmos de clusterização k-means que divide o conjunto de dados em grupos, de forma que os objetos contidos na base de dados fiquem agrupados de acordo com a semelhança entre eles, os autores mostraram as diferenças de duas clusterizações com os mesmos dados.

[Brito et al. 2019] tiverem como objetivo identificar padrões e classificar os discentes com o perfil mais propenso à evasão além de descobrir os possíveis motivos que contribuem para o crescimento da evasão, com dados dos 196 discentes do curso de graduação em Sistemas de informação da UFRN, os algoritmos utilizados através da ferramenta WEKA foram o K-Means e o J48. Os resultados dos experimentos mostram que: reprovar nas quatro disciplinas base do curso, não participar de nenhum tipo de projeto, junto com a extrapolação dos 8 semestres normais do curso e ter uma faixa etária superior a 26 anos, são os fatores que mais colaboraram para a evasão do curso.

Nota-se que já é possível encontrar vários trabalhos utilizando a Mineração de Dados no contexto educacional, onde este será mais um que irá contribuir e complementar a área da EDM. Este trabalho se diferencia dos demais por realizar uma abordagem à todos os cursos de Sistemas de informação e tentar predizer o desempenho dos alunos a partir das notas do Enade em 2017 comparando o desempenho dos algoritmos.

3. Materiais e Métodos

Será utilizada a base de dados do Enade referente ao ano de 2017, a qual contém avaliação de diversos cursos, entre eles os cursos de Sistemas de Informação. Esta avaliação mede o desempenho dos estudantes e avalia os conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos de graduação.

Nesse trabalho, será analisado somente o curso de Sistema de Informação, obter informações a cerca do desempenho dos alunos na avaliação do Enade, obter informações relevantes por meio do processo de Mineração de Dados, e foi optado para este estudo o método do KDD, que é um método bastante utilizado na área e simples de implementar.

O KDD abrange diversas fases que são potencialmente o caminho que os dados percorrem até tornarem-se conhecimento que seja útil. Ou seja, consiste em um processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente úteis e interpretáveis [Fayyad et al. 1996]. A descoberta de padrões é um processo que começa pela escolha dos dados que servem para responder as questões de pesquisa do estudo. Os dados são integrados e pré-processados para que sejam entregues estruturados, limpos, selecionados e padronizados à tarefa de mineração de dados. Na etapa de mineração aplica-se alguma técnica inteligente que possibilite a geração de informações. Os resultados devem ser pós-processados e apresentados de maneira que possam auxiliar na tomada de decisão.

3.1. Etapas do processo KDD

O KDD (Knowledge Discovery in Databases) representa uma metodologia de descoberta de informações a partir de uma sequência definida de passos, sendo eles: Limpeza de dados; Integração dos dados; Seleção dos dados; Transformação dos dados; Mineração dos dados; Avaliação dos Padrões; Apresentação e Assimilação do conhecimento [Han et al. 2014].

De acordo com a figura 2 é possível observar as etapas do processo KDD que serão realizadas nesse trabalho.

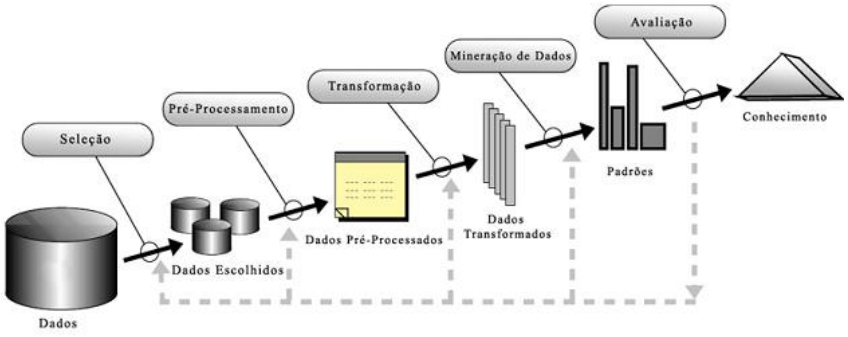

Figura 2. Etapas do processo KDD
[Fayyad et al. 1996]

3.1.1. Seleção

A base de dados do Enade do ano de 2017, possui 537.436 registros de alunos em que cada registro contém 150 variáveis distintas, que relacionando cada registro com os seus atributos totaliza mais 80 milhões de dados nesta base. Nessa etapa, são selecionados apenas os dados que serão relevantes para este estudo.

Foi selecionado os alunos dos cursos de Sistema de Informação, aplicando um filtro na variável CO_GRUPO = 4006, como também foi utilizado um filtro nas variáveis TP_PRES, TP_PR_GER, TP_PR_OB_FG, TP_PR_DI_FG, TP_PR_OB_CE e TP_PR_DI_CE que representam somente os estudantes que estavam presentes e com respostas válidas na parte objetiva e discursiva de formação geral e no componente específico, resultando 9.454 registros de alunos.

A retirada desses registros se faz necessária pois será analisado somente os alunos dos cursos de Sistema de Informação e que realizaram à prova e tiveram suas notas preenchidas, por isso faz-se necessário excluir os registros de alunos que não estavam presentes na avaliação.

3.1.2. Pré-processamento

Esta etapa consiste em verificar a qualidade dos dados armazenados, como também limpar, corrigir ou remover dados inconsistentes, verificar dados ausentes ou incompletos, identificar anomalias (outliers).

Como pode ser observado na Figura 3, mesmo após o filtro dos alunos que estavam presentes e tiveram suas repostas válidas, ainda existia na base registros com o atributo da nota em branco ou com nota igual a zero. Estes registros foram excluídos, uma vez que a nota obtida pelo aluno na avaliação é um atributo de grande relevância para este estudo.

CO_REGIAO_CURSO	NU_IDADE	TP_SEXO	NT_GER	QE_I01	QE_I02	QE_I04	QE_I05	QE_I08	QE_I10	QE_I17	QE_I22	QE_I23
4	33	M	53,4	E	D	B	E	C	E	A	C	B
4	23	M	42,5	A	A	D	D	E	E	A	A	C
4	24	M	62,1	A	A	D	D	F	E	B	A	B
4	27	M	54,1	A	A	F	D	F	E	B	B	B
4	25	M	33,6	A	A	C	C	D	E	B	B	E
4	23	M	53,5	A	A	F	C	E	E	B	A	B
4	24	M	52,2	A	A	D	D	F	E	B	A	A
4	30	M	52,7	E	A	F	C	F	E	B	B	E
4	22	M	53,7	A	A	E	E	F	E	B	B	B
4	35	M	54,4	A	A	D	C	F	E	B	B	B
4	27	M										
4	36	M	64,2	B	F	C	C	D	E	B	B	B
4	48	M										
4	47	M	40,6	A	A	B	B	D	E	E	B	B
4	24	M										
4	25	M	24,6	A	A	D	F	F	E	B	C	C
4	27	M	30,1	A	F	C	D	C	E	A	A	C
4	23	F	40,4	A	A	D	E	E	E	A	B	B
4	34	M	53,7	A	F	E	D	E	E	B	B	B

Figura 3. Base de dados com registros em branco

Com relação aos atributos que totalizam 150, foi realizado a exclusão de forma manual de alguns deles que não influenciariam nos resultados, como por exemplo o atributo NU_ANO=2017 que faz referência ao ano daquele registro, e como no estudo está sendo utilizado apenas um único ano, não se faz necessário.

Outros atributos também foram excluídos pelo fato de possuírem apenas um único dado para todos os registros, que foram: CO_GRUPO, TP_INSCRICAO_ADM, TP_INSCRICAO, NU_ITEM_OFG, NU_ITEM_OFG_Z, NU_ITEM_OFG_X, NU_ITEM_OFG_N, NU_ITEM_OCE, NU_ITEM_OCE_Z, NU_ITEM_OCE_X, NU_ITEM_OCE_N, TP_PRES, TP_PR_GER, TP_PR_OB_FG, TP_PR_DI_FG, TP_PR_OB_CE e TP_PR_DI_CE.

Os atributos que representava os vetores dos gabarito das respostas originais e das escolhas dos candidatos também foram eliminados da planilha de dados, que são: DS_VT_GAB_OFG_ORIG, DS_VT_GAB_OFG_FIN, DS_VT_GAB_OCE_ORIG, DS_VT_GAB_OCE_FIN, DS_VT_ESC_OFG, DS_VT_ACE_OFG, DS_VT_ESC_OCE e DS_VT_ACE_OCE. E as questões QE_I69 a QE_I81 foram excluídos pois são itens exclusivos para os estudantes das licenciaturas, o que não se aplica para os estudantes de Sistema de Informação os quais são objeto de estudo deste trabalho.

Do total de 150 atributos, após as exclusões manuais dos atributos que não tinham relevância para o estudo, restou 43 atributos, os quais a partir destes será utilizado na execução dos algoritmos de mineração de dados a partir do software Weka.

3.1.3. Transformação

Os atributos que foram selecionados manualmente para poderem serem utilizados no processamento dos dados, passaram por uma transformação, de modo que cada atributo pudesse possuir apenas duas opções de resultados, para com isso facilitar o classificador e obter melhores resultados na acurácia do modelo.

As variáveis de Idade e Notas, por exemplo, que são atributos que possuem um grande conjunto de valores diferentes entre si, a utilização de cada variação não seria viável para o estudo, com isso foi definido intervalos para esses atributos, que facilitam a interpretação dos dados.

Para o atributo de NT_GER, que corresponde a nota geral obtida pela aluno, classificamos ela como Nota Acima para as notas iguais ou maiores que 50 pontos e Nota Abaixo para as notas que foram menores que 50 pontos, como pode ser observado na Tabela 1 abaixo.

A tabela com o Dicionário dos dados e como se deu o processo de transformação dos mesmos poderá ser consultado no link: `encurtador.com.br/ghoJY`.

Tabela 1. Intervalo para o atributo Nota Geral.

Intervalos para NT_GER	Descrição
Nota Abaixo	Notas menores que 50 pontos
Nota Acima	Notas maiores ou iguais a 50 pontos

3.1.4. Mineração de Dados

Com o pré-processamento resolvido, ou seja, com os dados já selecionados, limpos e padronizados, pode-se ser iniciada a etapa de mineração de dados. Esta etapa é responsável pela extração dos padrões e conhecimentos, independente da quantidade de dados e de sua complexidade.

A mineração de dados é a principal etapa do processo de KDD, nessa etapa é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD. Alguns autores se referem à mineração de dados e à descoberta de conhecimento em bases de dados como sinônimos. Envolve a aplicação de algoritmos sobre os dados em busca de conhecimento implícitos e úteis [Boente et al. 2008].

O software que será utilizado para a realização deste trabalho é o WEKA 3.8 (Waikato Environment for Knowledge Analysis) que, de acordo com Abernethy (2010), trata-se de um software de código aberto e gratuito que possibilita a transformação de dados em conhecimento útil. O grande número de algoritmos de aprendizado de máquina implementados pelo WEKA é um dos maiores benefícios de usar a plataforma, ele tem como objetivo agregar algoritmos provenientes de diferentes abordagens dedicando-se ao estudo de aprendizagem de máquina [Silva 2018].

No software Weka na aba para seleção de atributos, foi executado o algoritmo de correlação que avalia o valor de um atributo medindo a correlação (de Pearson) entre ele e o atributo classe, que no nosso estudo o atributo classe é a classificação da nota. Na

lista geral do ranking das variáveis, foi selecionado apenas as 10 melhores ranqueadas da lista para serem incluídas na classificação do modelo, que são essas que estão descritas na Tabela 2 logo abaixo.

Tabela 2. Os 10 primeiros atributos do Ranking de correlação.

Atributos	**Descrição**
CO_CATEGAD	Categoria Administrativa (Pública ou Privada)
POSSUI_BOLSA_ACADEMICA	Se ao longo do curso o aluno recebeu algum tipo de bolsa acadêmica
RENDA_FAMILIA	Renda da família do aluno
ENUNCIADO_CE_OBJ	Se os enunciados das questões do componente especifico estavam claros e objetivos
TEMPO_GASTO_PROVA	Tempo gasto para a realização da prova
ATIVIDADES_EXTERIOR	Se durante o curso participou de programas ou atividades no exterior
CURSOU_EMEDIO	Tipo de escola que cursou o Ensino Médio
CO_TURNO_GRADUACAO	Turno do curso de graduação
HORA_ESTUDO_SEMANA	Quantidade de horas que dedicou aos estudos na semana, exceto as aulas
ESCOLARIZACAO_PAI	Até que etapa o pai concluiu os estudos

De total dos atributos que continham na base, restaram 11 que serão aplicados nos algoritmos de classificação, sendo um atributo classe que é a classificação da nota e os 10 atributos selecionados pelo algoritmo de correlação.

4. Resultados

Para realização do estudo foram utilizados 4 algoritmos de aprendizagem de máquina o J48, Naive Bayes, Random Forest e Regressão Logística com seus respectivos parâmetros default do software Weka. Como forma de avaliação do desempenho dos algoritmos, utilizou-se a acurácia e precisão para classificação das duas classes. Os dados das execuções podem ser vistos na Tabela 3.

Tabela 3. Precisão dos Algoritmos de aprendizagem de máquina

	J48	**Naive Bayes**	**Random Forest**	**Regressão Logística**
Acurácia	63,42%	64,09%	63,22%	64,49%
Precisão da Classe "Nota Acima"	58,8%	57,5%	57,5%	59,9%
Precisão da Classe "Nota Abaixo"	64,6%	66,6%	84,7%	65,8%

A divisão entre o conjunto de treinamento e testes foi feita utilizando a Validação Cruzada (do inglês, Cross Validation). Nesta abordagem, o conjunto de dados é dividido em 10 partes iguais (Folds), sendo 9 destas utilizadas para treinamento e a restante utilizada para teste, a fim de se criar um modelo. O processo é repetido 10 vezes no total, usando um segmento diferente de cada vez para o teste, totalizando 10 modelos de onde extrai-se a média representativa de todos os resultados dos modelos.

Verificando os resultados apresentados na Tabela 3, percebe-se que os algoritmos obtiveram valores semelhantes para a acurácia e precisão das classes, com acurácia superior a 60% em todos os algoritmos selecionados, porém o que obteve uma melhor acurácia foi a regressão logística com 64,49%.

Na precisão das classes, que é proporção de instâncias que são verdadeiramente de uma classe dividida pelo total de instâncias classificadas como aquela classe, observou-se uma precisão superior na classe "Nota Abaixo" que são os alunos que obtiveram notas inferiores a 50 pontos, isso mostra que o modelo classificou melhor esses alunos do que os que estavam na classe com "Nota Acima".

O fato dos alunos da classe "Nota Abaixo" terem sido melhores classificados é relevante, tendo em vista que são estes os alunos que necessitam de um melhor apoio didático, que consequentemente são os alunos que possuem um maior risco de evasão, como também que fazem com que a média geral do curso seja baixa.

5. Considerações Finais

Este estudo teve como objetivo predizer o desempenho dos alunos de Sistema Informação a partir de algoritmos de aprendizagem de máquina utilizando o software Weka e os dados da prova do Enade. Observou-se que os algoritmos obtiveram uma acurácia superior a 60%, e o melhor desempenho foi do algoritmo de Regressão Logística com 64,49%, que estudos posteriores poderão aperfeiçoar o modelo e encontrar outros fatores que melhorem esse resultado.

O modelo classificou melhor os alunos com "Nota Abaixo" do que aqueles alunos que foram classificados com "Nota Acima", pode-se inferir que outros fatores além dos que foram selecionados são determinantes na classificação dos alunos com notas altas. O algoritmo Random Forest, por exemplo, que obteve 84,7% de precisão para a classe de "Nota Abaixo", que é uma bom resultado e mostra que é viável realizar a predição do desempenho baseado nas características e perfil do aluno.

Outro objetivo era identificar os fatores que eram determinantes na classificação do desempenho, e com base no algoritmo de correlação foram selecionados os 10 melhores, que foram os citados na Tabela 2.

Segundo o algoritmo de correlação a categoria administrativa da instituição se pública ou privada foi relevante na classificação, como também os alunos que possuem bolsa acadêmica como bolsa de iniciação científica, extensão, monitoria e tutoria, a renda familiar, a percepção do aluno com relação aos enunciados da prova do componente específico se estavam claros e objetivos, o tempo gasto para a realização da prova, se o aluno participou de atividades no exterior como o programa Ciências Sem Fronteiras ou intercâmbios, o tipo de escola que cursou durante o ensino médio se pública ou privada e o turno do seu curso se durante o dia ou noturno; a quantidade de horas que foi dedicado aos estudos, excetuando as horas que estavam em sala de aula no curso e escolarização do pai do aluno. Esses atributos são relevantes na classificação da nota dos alunos.

Acredita-se que os resultados obtidos neste estudo possam ajudar os educadores, uma vez que é possível obter estimativas sobre o desempenho dos alunos, e então servir de base para o planejamento de estratégias e políticas que visem diminuir o baixo despenho e consequentemente o índice de evasão do curso.

Referências

Baker, R. S. J., Isotani, S., and de Carvalho, A. M. J. B. (2011). Mineração de Dados Educacionais : Oportunidades para o Brasil. 19:3–13.

Boente, A. N. P., Goldschmidt, R. R., and Estrela, V. V. (2008). UMA METODOLOGIA DE SUPORTE AO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS. pages 1–14.

BRASIL (2011). LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011.

BRASIL (2019). Ministério da educação - exame nacional de desempenho dos estudantes (enade).

Brito, D. M. D., Araújo, I., and Júnior, D. A. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. (Cbie):882–890.

Brito, I. P. D., Rabelo, H., Naschold, Â. M. C., Ferreira, A. M., Burlamaqui, A. M. F., Rabelo, D. S. d. S., and Valentim, R. A. d. M. (2019). Uso de Mineração de Dados Educacionais para a classificação e identificação de perfis de Evasão de graduandos em Sistemas de Informação da UFRN. (Cbie):159–168.

Côrtes, S. D. C., Porcaro, R. M., and Lifschitz, S. (2002). Mineração de Dados – Funcionalidades, Técnicas e Abordagens. *PUC-Rio Informática*, page 35.

Costa, E., Baker, R. S. J. d., Amorim, L., Magalhães, J., and Marinho, T. (2012). Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação - JAIE*.

CRETTON, N. N. and GOMES, G. R. (2016). Aplicação De Técnicas De Mineração De Dados Na Base De Dados Do Enade Com Enfoque Nos Cursos De Medicina. *Acta Biomédica Brasiliensia*, 7(1):74.

da Fonseca, S. O. and Namen, A. A. (2016). Mineração Em Bases De Dados Do Inep: Uma Análise Exploratória Para Nortear Melhorias No Sistema Educacional Brasileiro. *Educação em Revista*, 32(1):133–157.

Fayyad, U. M., Piatetsky Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *"Advances in Knowledge Discovery and Data Mining*.

Filho, A. E. C. M., Roseira, Í. B. R., and Junior, J. A. d. F. P. (2020). Perfil socioeconômico e desempenho de estudantes de licenciatura em educação fí-sica no ENADE/BRASIL Socioeconomic profile and the performance of physical education undergraduate students in ENADE/BRAZIL. *Dialnet.Unirioja.Es*, 35:90–101.

Francelino, W. L. and Machado, L. D. S. (2020). MINERAÇÃO DE DADOS NOS MICRODADOS ENADE COMPUTAÇÃO.

Han, J., Kamber, M., and Pei, J. (2014). *Data mining: Concepts and Techniques*.

Isotani, S. and Bittencourt, I. (2015). *Dados Abertos Conectados: em Busca da Web do Conhecimento*.

Júnior, J. G. D. O., Noronha, R. V., and Kaestner, C. A. A. (2017). Método de Seleção de Atributos Aplicados na Previsão da Evasão de Cursos de Graduação. 13:54–67.

Oliveira, M. J. S., Caetano, G., and Pedro, E. M. (2018). Usando a Mineração de Dados para predição de desempenho de alunos nas disciplinas de português e matemática. *Revista de Educação do Vale do Arinos*, 5:8–16.

Patricio, T. S. and Magnoni, M. d. G. M. (2018). Mineração de Dados e Big Data na Educação. *Revista GEMInIS*, 9(1):57–75.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Romero, C. e. a. (2011). Handbook of educational data mining. *CRC Press*.

Scaico, P. D., De Queiroz, R. J. G. B., and Scaico, A. (2014). O conceito big data na Educação. *Anais do XX Workshop de Informática na Escola (WIE 2014)*, 1(Cbie):328.

Silva, R. (2018). Introdução ao WEKA: Software para mineração de dados. page 10.

Souza, M. M. B. d. A. C. F. F. d. (2017). MineraÇÃo de dados educacionais: PrevisÃo de notas parciais utilizando classificaÇÃo. *Programa de Pós-Graduação em Informática - PPGI*.