UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

WILTON RIBEIRO CRISPIM NETO

O USO DE MINERAÇÃO DE DADOS EDUCACIONAIS SOB O ENADE COMO APOIO AO PROCESSO DE TOMADA DE DECISÃO DE GESTORES DO ENSINO SUPERIOR

QUIXADÁ

2021

WILTON RIBEIRO CRISPIM NETO

O USO DE MINERAÇÃO DE DADOS EDUCACIONAIS SOB O ENADE COMO APOIO AO PROCESSO DE TOMADA DE DECISÃO DE GESTORES DO ENSINO SUPERIOR

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Prof[a]. Dra. Paulyne Matthews Jucá

QUIXADÁ

2021

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

C949u Crispim Neto, Wilton Ribeiro.
O uso de mineração de dados educacionais sob o ENADE como apoio ao processo de tomada de decisão de gestores do ensino superior / Wilton Ribeiro Crispim Neto. – 2021.
57 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2021.
Orientação: Profa. Dra. Paulyne Matthews Jucá.

1. Mineração de Dados. 2. Exame Nacional de Desempenho dos Estudantes. 3. Exploração de dados (Computação). I. Título.

CDD 005.1

WILTON RIBEIRO CRISPIM NETO

O USO DE MINERAÇÃO DE DADOS EDUCACIONAIS SOB O ENADE COMO APOIO AO PROCESSO DE TOMADA DE DECISÃO DE GESTORES DO ENSINO SUPERIOR

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em: _____/_____/_____

BANCA EXAMINADORA

Profª. Dra. Paulyne Matthews Jucá (Orientadora)
Universidade Federal do Ceará (UFC)

Profª. Ma. Antonia Diana Braga Nogueira
Universidade Federal do Ceará (UFC)

Profª. Ma. Lívia Almada Cruz Rafael
Universidade Federal do Ceará (UFC)

À todos aqueles que acreditaram em mim.

AGRADECIMENTOS

Agradeço inicialmente à Deus pelos cuidados enquanto estive tão longe de casa. Por tantas vezes foi minha única e maior companhia.

Aos meus pais, Rute e Maurício, e à minha esposa Vaniele, que sempre me apoiaram nos meus sonhos e nos caminhos que decidi seguir. Eu não seria nada sem vocês.

À minha vozinha, que com absoluta certeza de longe comemora mais um passo que seu menino está completando. Eu sou resultado de todo seu amor.

Agradeço também com enorme carinho à professora Paulyne, pela oportunidade de ser seu aluno, orientando e sobretudo, aprendiz, com certeza levarei para a vida toda um exemplo de profissional e pessoa extraordinária a me espelhar.

Aos meus amigos Clebson, Josué, Paulo, Pedro, Rafael e Victor, que sorte eu tive de ter tido a oportunidade de conhecer vocês, meu muito obrigado por estarem comigo nessa etapa tão importante da minha vida. Aliás os nomes estão em ordem alfabética para não gerar ciúmes.

Por fim, agradeço aos meus colegas de trabalho da Apicbase, em especial aos meus lideres Carl, Pieter e Vitor, pelos ensinamentos, momentos e por sempre lembrarem que neste momento tão crucial entre trabalho e estudos a graduação vem primeiro.

“Wwwrrrrrrgwwwrrrr aaararrrgwwwh rrwwwg”

(Chewbacca)

RESUMO

A grande quantidade de dados, encontradas na base do ENADE, é a coleta das informações de todos os alunos que fizeram o exame e seus resultados. Descobrir um conhecimento novo a partir dessa massa de dados pode garantir um melhor entendimento sobre os alunos, os cursos e universidades que frequentam. O foco dessa pesquisa é na descoberta de características institucionais que tem maior impacto no desempenho final do aluno no exame. Nisso, a Mineração de Dados Educacionais pode direcionar gestores e coordenadores ao planejamento de melhores ações que visem boas tomadas de decisões, ao saber quais características precisam de maior investimento para alcançar a excelência desejada. Foram aplicadas técnicas de mineração de dados educacionais sobre conjunto de dados reais dos cursos de Tecnologia da Informações presentes nas edições do exame de 2014 e 2017. Como resultado, são mostrados os esquemas adotados e análises dos resultados obtidos.

Palavras-chave: Mineração de Dados. Exame Nacional de Desempenho dos Estudantes. Exploração de dados (Computação).

ABSTRACT

The large amount of data, found at the base of the ENADE exam, is the collection of information from all students who took the exam and their results. Discover new knowledge from this mass of data can ensure a better understanding of students, courses and universities they attend. The focus of this research is on discovering institutional characteristics that have the greatest impact on the final performance of the exam. Related to this, Educational Data Mining can help and lead managers and coordinators to plan better actions that aim at good decision making. For this purpose, educational data mining techniques were applied to the real dataset of Information Technology courses present in the 2014 and 2017 exam editions. As a result, the adopted schemes and some analysis of the results obtained are shown.

Keywords: Data Mining. National Student Performance Exam. Data Exploration (Computing).

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Desde o advento da tecnologia de informação em diversas áreas do conhecimento humano, o crescimento de dados disponíveis atingiu proporções inéditas, em especial nos últimos 15 anos (WU *et al.*, 2013). É possível confirmar essa tendência através da predição feita pela International Data Corporation (IDC) no qual aponta um crescimento exponencial da quantidade de dados globais, que deve crescer de 33 Zettabytes em 2018 para 175 Zettabytes em 2025 (IDC, 2018). Atrelado a isso, as recentes oportunidades para descoberta de novos dados possibilitam uma profunda compreensão dos valores ocultos e as relações entre eles, de maneira a prever novas interpretações e resultados até então desconhecidos (CHEN *et al.*, 2014).

Esse contexto permitiu que a educação fosse revisitada, em relação aos seus meios (MARTUCCI, 2000). Dentro dessa perspectiva, há um novo empenho das organizações governamentais, em diversos âmbitos, em efetuar ações que possibilitem a melhoria e análise do ensino em todos os níveis da educação, desde a alfabetização até os níveis mais elevados de especialização. Assim, como parte dessas ações, o desenvolvimento de sistemas de avaliação visando à elaboração de diagnósticos para a análise, revisão e melhoria do ensino passam a ocupar um papel mais relevante na agenda política educacional (BAUER, 2012).

No Brasil, o órgão vinculado ao Ministério da Educação, responsável por realizar pesquisas, avaliações e estudos sob o sistema educacional no país é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP. O instituto tem como propósito financiar a criação e aplicação de políticas públicas e produzir dados seguros aos educadores, gestores e público em geral (INEP, 2011) por meio de seus exames, no qual objetivam avaliar e identificar os fatores que impactam a qualidade de ensino dos estudantes em geral (tais como elementos culturais, socioeconômicos, dentre outros), informações sobre os docentes, gestores, instituições, infraestrutura e o aprendizado do aluno em relação aos conteúdos propostos (FONSECA; NAMEN, 2016).

No que diz respeito à avaliação da qualidade de ensino da Educação Superior, o INEP aplica o Exame Nacional de Desempenho do Estudante (ENADE), principal objeto de estudo deste trabalho. O ENADE objetiva acompanhar e avaliar o desenvolvimento dos estudantes de cursos de graduação, com relação ao conteúdo disposto nas diretrizes curriculares de seus respectivos cursos (VISTA *et al.*, 2017). Desta maneira, a aplicação da avaliação, por meio de exames e questionários, permite a disponibilização de uma base de dados com uma grande quantidade de informações relevantes à observação de gestores, pesquisadores, educadores e a

comunidade.

A cada aplicação do ENADE, o INEP disponibiliza relatórios com análises descritivas, como por exemplo: média das notas por sexo, curso, região e/ou instituição. Em Reis e Reis (2002) esse tipo de análise é classificado como um método para sintetizar, organizar e apresentar os aspectos relevantes de um conjunto de dados observados ou comparar tais dados ou mais conjuntos. Já a estatística inferencial, não utilizada pelo INEP e comumente relacionada à mineração de dados, é outro método de análise estatística que se preocupa com o meio necessário para que, a partir de dados, possa-se obter conclusões gerais e/ou que seja possível fazer predições (SULLIVAN-BOLYAI; BOVA, 2014).

A aplicação de técnicas de mineração de dados no meio educacional deu origem a uma nova área de estudo denominada Mineração de Dados Educacionais (MDE) (BAKER *et al.*, 2011), área pela qual se encaixam os objetivos deste trabalho. Embora o âmbito da MDE seja único, o educacional, envolve diversos grupos de usuários ou participantes, dessa maneira, diferentes grupos podem ter noções e pontos de vista distintos a depender de sua missão, visão e objetivos no uso de Mineração de Dados (HANNA, 2004).

Os microdados do ENADE são disponibilizados à população, porém se faz necessário ter conhecimentos específicos para o entendimento dos dados, seja por meio do público em geral, quanto como educadores, diretores e/ou gestores do ensino superior.

O presente trabalho visa apresentar os principais aspectos institucionais relacionados ao desempenho dos discentes, a partir da mineração de dados coletados por meio do Questionário do Estudante presente no exame, com a finalidade de auxiliar e apoiar gestores, diretores e responsáveis pelos cursos de graduação no país, na compreensão de quais aspectos tenham maior relação à nota final do aluno.

1.1 Objetivos

1.1.1 Objetivo Geral

Identificar as relações entre as variáveis escolhidas a partir do Questionário do Estudante e o desempenho final do estudante no exame, a fim de auxiliar gestores de Instituições do Ensino Superior (IES) na compreensão sob quais aspectos institucionais que mais poderiam impactar no rendimento dos alunos no exame e consequentemente na avaliação da instituição de ensino.

1.1.2 Objetivos Específicos

Para atingir o objetivo geral deste trabalho, são estes os objetivos específicos:

- Relatar as questões de maior impacto em relação com a nota final do aluno, por meio da Mineração de Relações;
- Encontrar as relações entre as questões presentes do ENADE e as instituições de ensino;
- Comparar os resultados ao desempenho das instituições sob diferentes edições do exame, no espaço mínimo de 3 anos;

2 FUNDAMENTAÇÃO TEÓRICA

Nesta Seção, são demonstrados os fundamentos que conceituam o desenvolvimento deste trabalho. Na Seção 2.1, são introduzidos os conceitos sobre Mineração de Dados, bem como sua participação do processo de descoberta de conhecimento e técnicas. Na Seção 2.2, são apresentados os conceitos sobre Mineração de Dados Educacionais, desde sua definição às técnicas relacionadas. Na Seção 2.3, são detalhados os dados sobre a prova do ENADE, desde a história, processo avaliativo e relatórios.

2.1 Mineração de Dados

A capacidade de geração de dados nunca foi tão grande, desde a invenção e introdução da tecnologia da informação, e devido a isso, o maior desafio das aplicações que lidam com a exploração de grande quantidade de dados é a de extrair informações ou conhecimentos úteis para ações futuras (WU *et al.*, 2013). Segundo Bramer (2007), a Mineração de Dados (MD) consiste na descoberta de estruturas interessantes, inesperadas ou valiosas em um grande conjunto de dados. Seguindo nessa mesma linha, o autor Han *et al.* (2011) cita a MD como antes de tudo um processo, no qual o principal objetivo é a descoberta de padrões e conhecimento relevantes sob uma grande quantidade de dados. É durante a MD que ocorre a aplicação de algoritmos para extração de padrões, estruturas e tendências que sejam úteis e de interesse do usuário, no qual técnicas e procedimentos são usados em função da natureza dos dados e das informações que se deseja alcançar (OLSON; DELEN, 2008). A MD é considerada um dos passos fundamentais da descoberta de conhecimentos em bancos de dados - Knowledge Discovery in Databases (KDD), que consiste no processo inteiro da conversão de dados brutos em informações relevantes (HAN *et al.*, 2011).

2.1.1 Processo KDD

Segundo Fayyad *et al.* (1996), KDD é um processo não trivial de identificar padrões válidos, novos e potencialmente úteis sob os dados. Ao abordar o assunto como um processo, compreende-se que o KDD seja interativo e iterativo, conforme pode ser observado na Figura 1. Portanto, envolve diferentes etapas com muitas interações de tomada de decisão do usuário, como a preparação de dados, pesquisa de padrões, avaliação de conhecimento e refinamento, tudo repetido em várias iterações. Ainda segundo Fayyad *et al.* (1996), as etapas são definidas

da seguinte forma.

1. Seleção: esta etapa compreende primeiramente a descoberta e conhecimento do domínio, e assim a seleção dos dados como um subconjunto de interesse.
2. Pré-processamento: é recorrente que após a seleção dos dados, exista dentro da amostra possíveis distorções, erros ou duplicações. Portanto, é nessa etapa que há o tratamento dessa inconsistência e padronização dos valores selecionados.
3. Transformação: a partir de métodos como normalização, redução, e discretização dos dados, é nessa etapa que os dados são processados e dispostos de maneira diferente do original, no qual objetiva-se a procura de atributos úteis para os objetivos do trabalho e aperfeiçoamento do modelo para as etapas seguintes, ainda que mantenha as mesmas propriedades.
4. Mineração de dados: essa etapa, já discutida anteriormente, é segundo Goldschmidt e Passos (2005) a etapa mais importante do processo de KDD. É a partir da MD que diferentes métodos podem ser usados com o objetivo de elaborar um modelo que represente um conjunto de dados, identificar dados aglomerados, inferir classificações e descobrir padrões.
5. Interpretação e Avaliação: por último, são extraídos os padrões descobertos. Mesmo depois da realização do processo de KDD.

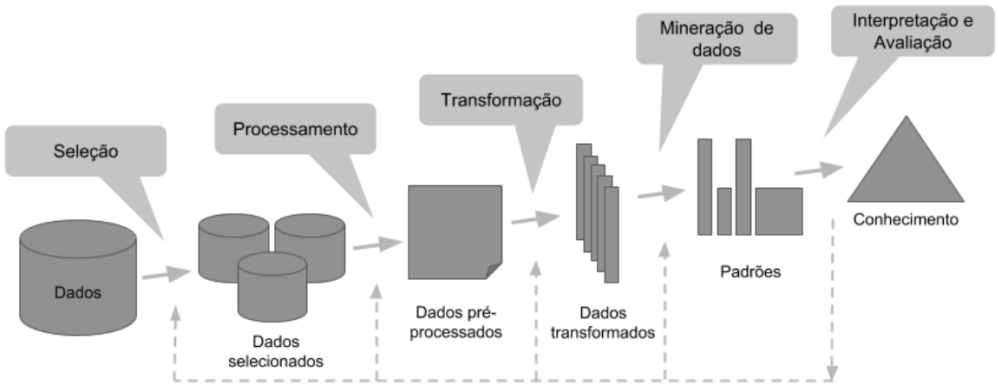
Ainda entre as etapas, a depender da visão do analista, os passos podem ser refeitos sob novos métodos de MD, ou diferentes processos de transformação e pré-processamento, a fim de verificar resultados diferentes ou mais precisos para o objetivo previamente proposto.

2.1.2 Técnicas de Mineração de Dados

Segundo Fayyad *et al.* (1996), maioria das técnicas de MD são baseadas em técnicas já aplicadas e aplicadas sob as áreas de aprendizado de máquina, reconhecimento de padrões e estatísticas. Apesar de diversos modelos da computação aplicados à MD, sobre aprendizado supervisionado as técnicas mais comumente utilizadas incluem regressão e classificação, e sobre aprendizado não supervisionado incluem agrupamento e regras de associação (OLSON; DELEN, 2008). A seguir são descritas, de maneira resumida, as técnicas de MD normalmente utilizadas:

1. Regressão: ainda segundo Fayyad *et al.* (1996), a técnica de regressão consiste em uma função que mapeia um item de um conjunto de dados para uma variá-

Figura 1 – Processo da Descoberta de Conhecimento em Bases de Dados



Fonte: Adaptado de Fayyad et al. (1996)

vel de previsão com valor real. As aplicações de regressão são muitas, como por exemplo, estimar a probabilidade de um paciente sobreviver, a partir dos resultados de um conjunto de testes de diagnóstico.

2. Classificação: A classificação é uma das técnicas de mineração de dados que é útil para prever a associação do grupo para instâncias de dados. A classificação é um tipo supervisionado de aprendizado de máquina em que há fornecimento de dados rotulados com antecedência. Ao fornecendo treinamento, os dados podem ser treinados e podemos prever o futuro dos mesmos, e nesse sentido, previsão é busca dizer à qual classe os dados podem pertencer (GERA; GOEL, 2015).
3. Agrupamento: O principal objetivo da técnica de agrupamento é separar o conjunto de dados não rotulados em finitos e conjunto discreto de estruturas de dados naturais e ocultas. No entanto, não há intuito de fornecer caracterização precisa das amostras (GERA; GOEL, 2015).
4. Regras de Associação: Consiste em identificar quais atributos do conjunto de dados estão associados entre si, estabelecendo uma correlação estatística entre os dados (GOEBEL; GRUENWALD, 1999). Um exemplo conhecido é o do carrinho de compras, no qual é possível identificar quais produtos são levados junto com os clientes (ARAÚJO *et al.*, 2019).

Devido a natureza do estudo, as técnicas utilizadas serão apresentadas com mais

detalhes a seguir na Seção de Mineração de Dados Educacionais.

2.1.3 Ferramentas para Mineração de Dados

No mercado, e meio acadêmico, existem diversas ferramentas disponíveis e bem avaliadas para a realização da MD. Dentre elas é possível citar DBMiner, Clementine, Intelligent Miner e WEKA, que oferecem coleções de algoritmos de mineração e/ou de pré-processamento, bem como técnicas de visualização dos dados (ROMERO; VENTURA, 2007). A seguir será dada uma breve descrição sobre as ferramentas analisadas:

1. O WEKA (*Waikato Environment for Knowledge Analysis*) é um software escrito em Java, de acesso livre e desenvolvido pela Universidade de Waikato, Nova Zelândia. É formando por uma coleção de algoritmos de diversas técnicas de mineração e ferramentas de pré-processamento, além de incorporar os principais métodos de MD, entre eles a classificação, regressão, regras de associação, agrupamento e seleção de atributos (HALL *et al.*, 2009).
2. DBMiner é um software de MD, desenvolvido para mineração interativa de conhecimentos de vários níveis em grandes bancos de dados relacionais. O sistema implementa um amplo espectro de funções de mineração de dados, incluindo generalização, caracterização, associação, classificação e previsão (HAN *et al.*, 1996).
3. Clementine foi desenvolvida pela SPSS, é um software que suporta a metodologia CRISP-DM, além de possuir outras facilidades no domínio da MD (ADDERLEY, 2004).
4. Intelligent Miner é uma ferramenta desenvolvida pela IBM com o objetivo de acelerar o processo de descoberta de informações, mantendo a qualidade das informações extraídas. O software destina-se ao uso de analistas de dados e tecnólogos de negócios em áreas como marketing, finanças, gerenciamento de produtos e gerenciamento de relacionamento com clientes (CABENA *et al.*, 1999).

Para esta pesquisa foi escolhida a ferramenta WEKA, devido a facilidade de manipulação dos dados, funcionamento em todas as plataformas operacionais por se tratar de um software Java, e por possuir os algoritmos de mineração de dados abordados neste estudo já implementados e testados nativamente.

2.2 Mineração de Dados Educacionais

Nos últimos anos a Mineração de Dados Educacionais (MDE) surge como uma ponte entre duas áreas distintas: Educação e Ciência da Computação (BAKHSHINATEGH *et al.*, 2018). A MDE é uma disciplina, preocupada em: desenvolver métodos para explorar os tipos únicos de dados provenientes de ambientes educacionais e em usá-los para entender melhor os alunos e as configurações em que eles aprendem (BAKER; YACEF, 2009). Além disso, a MDE pode ser executada em diferentes bases de dados educacionais, como plataformas de ambiente virtual de aprendizagem; e dados provenientes de avaliações institucionais como, por exemplo, o ENADE (objeto de pesquisa deste trabalho), ENEM, Censo Escolar, Censo da Educação, dentre outros meios de avaliação aplicados no país (ARAÚJO *et al.*, 2019).

Como citado anteriormente, a MDE traz consigo os métodos e abordagens da computação para a obtenção do conhecimento proposto, transformando-se em um campo que explora principalmente algoritmos estatísticos, de aprendizado de máquina e de mineração de dados nos diferentes tipos de dados educacionais (BAKHSHINATEGH *et al.*, 2018).

Dentre as abordagens, o autor Baker *et al.* (2010) propõe cinco categorias ou grupos de abordagens principais em MDE, dentre elas:

1. Predição: tem por objetivo modelar estruturas que permitam inferir aspectos específicos dos dados, conhecidos como variáveis preditivas, através da análise e combinação de diversos aspectos encontrados em outros dados, chamados de variáveis preditoras (BAKER *et al.*, 2010). Existem dois benefícios de se utilizar métodos de predição em MDE:

   > Primeiro, métodos de predição são utilizados para estudar quais aspectos de um modelo são importantes para predição, dando informação sobre o construto sendo examinado. Esta estratégia é frequentemente utilizada em pesquisas que tentam, de forma direta, predizer os benefícios educacionais para um conjunto de estudantes, sem primeiro predizer os fatores mediantes ou intermediários. Ou seja, o objetivo é verificar o quanto o aluno aprende sem considerar as diversas variáveis que influenciam a aprendizagem como, por exemplo, variáveis relacionadas ao comportamento do estudante. Segundo, os métodos de predição auxiliam a predizer o valor das variáveis utilizadas em um modelo. Este tipo de técnica pode auxiliar no desenvolvimento e uso de atividades instrucionais, pois consegue estimar os benefícios educacionais antes mesmo da atividade ser aplicada com os alunos. (BAKER *et al.*, 2011).

2. Agrupamento (*Clustering*) : é uma forma de modelagem de dados historicamente enraizada em áreas como matemática e estatística. O termo refere-se a divisão

de dados em grupos com características semelhantes, no qual cada grupo ou cluster consiste em dados que são semelhantes entre si e diferentes de dados em outros grupos (BERKHIN, 2006). Segundo Berkhin (2006), ao olhar sob a perspectiva do aprendizado de máquina, é possível observar que os clusters correspondem a padrões ocultos, portanto a busca por clusters é um aprendizado não supervisionado e o resultado representa um conceito de dados. Já sob a perspectiva de mineração de dados é que se encontram complicações adicionais, como grandes bancos de dados, dados com muitos atributos e atributos de tipos diferentes. Para Baker *et al.* (2010), a utilização do Agrupamento dentro da MDE é útil em casos em que as categorias mais comuns no conjunto de dados não são conhecidas antecipadamente. Desta forma os agrupamentos podem ser criados em vários tamanhos possíveis: por exemplo, as escolas podem ser agrupadas (com o objetivo de explorar semelhanças e diferenças entre elas), os estudantes podem ser agrupados (com o objetivo de explorar semelhanças e diferenças entre os alunos) ou as ações dos estudantes podem ser agrupados (para investigar padrões de comportamento dos mesmos).

3. Mineração de Relações: possui como objetivo descobrir possíveis relações entre as demais variáveis em um conjunto de dados. Esta abordagem pode envolver a descoberta de relações entre quaisquer variáveis presentes nos dados, ou pode envolver a tentativa de aprender quais variáveis são mais fortemente associadas a uma variável específica, previamente conhecida e relevante ao estudo (BAKER *et al.*, 2011). Em geral, os relacionamentos encontrados devem atender a dois critérios: i) Significância Estatística (statistical significance), no qual é frequentemente avaliada através de testes estatísticos, com o objetivo de aumentar a confiança de que um relacionamento encontrado não se tenha sido gerado ao acaso; ii) Interesse (interestingness), que tem por objetivo definir quais descobertas são as mais distintas e bem suportadas pelos dados, e podendo remover as semelhantes, caracterizadas como duplicadas ou repetitivas (BAKER *et al.*, 2010).
4. Descoberta em Modelos: é quando um modelo é utilizado como componente para outro modelo, no qual um ou mais modelos gerados via previsão, agrupamento ou, em alguns casos, desenvolvido usando o raciocínio humano em vez de

métodos automatizados, são utilizados como artefatos em uma outra análise, como previsão ou mineração de relacionamento (BAKER *et al.*, 2010). A descoberta em modelos é provavelmente a categoria mais incomum na taxonomia de EDM de Bakers, ao olhar sob uma perspectiva clássica da MD, e também um dos métodos menos conhecidos na área de pesquisa em MDE (ALGARNI, 2016).

5. Destilação de Dados para Julgamento Humano: propõe facilitar decisões humanas, de maneira a apresentar dados complexos de forma mais simples, para que seja possível a compreensão e observação das características mais importantes e relevantes ao dados (BAKER *et al.*, 2011). É através da destilação de dados que é possível inferir aspectos sobre os dados e assim tomar decisões que até então não poderiam ser tomadas e/ou automatizadas com apenas o uso de qualquer outro método da EDM. Este método é aplicado, geralmente, com dois principais propósitos: i) Classificação, no qual trata-se de uma preparação para a construção de um modelo de predição; ii) Identificação, no qual visa projetar os dados de maneira que sejam facilmente identificáveis por meio de padrões conhecidos, porém não formalizados (ALGARNI, 2016).

Como mencionados anteriormente, tais métodos usados na MDE, são divididos em cinco categorias propostas por Baker *et al.* (2010), sendo elas: predição, agrupamento, mineração de relações, descoberta em modelos e destilação de dados, ilustradas no Quadro 1.

Para esta pesquisa será utilizada, mais especificamente, a abordagem de Mineração de Relações, com objetivo de descobrir as possíveis relações entre as demais variáveis escolhidas a partir do Questionário do Estudante e o desempenho final do estudante no exame.

Além disso, as aplicações da MDE podem atingir qualquer um dos stakeholders envolvidos nos sistemas e ambientes educacionais. Para Bakhshinategh *et al.* (2018), as partes interessadas e os respectivos impactos da MDE nos grupos, foram definidos em quatro conjuntos:

1. Estudantes: fornecer feedback, personalização e recomendações pode melhorar o processo de aprendizado dos alunos.
2. Educadores: fazer descobertas e fornecer sistemas de apoio à decisão pode ajudar os educadores a melhorar o desempenho do ensino e tomada de decisões.
3. Pesquisadores: entender melhor as estruturas educacionais e avaliar a eficácia da aprendizagem, conforme as descobertas no campo da educação sejam

Quadro 1 – Principais abordagens na mineração de dados educacionais

Categoria	**Objetivos**	**Aplicações**
Predição	Desenvolve um modelo para prever determinadas variáveis com base em outras variáveis. As variáveis preditoras podem ser constantes ou extraídas do conjunto de dados.	Identificar alunos em risco; Compreender os resultados educacionais dos alunos.
Agrupamento	Agrupa uma quantidade específica de dados em diferentes clusters com base nas características dos dados. O número de clusters podem ser diferentes com base no modelo e os objetivos do processo de agrupamento.	Encontrar semelhanças e diferenças entre os alunos ou escolas; Identificar comportamento categorizado de novos alunos.
Mineração de Relações	Extrai o relacionamento entre duas ou mais variáveis no conjunto de dados.	Encontrar a relação entre o nível de educação dos pais e a evasão escolar; Descobrir as associações curriculares em sequências de cursos; Descobrir quais estratégias pedagógicas levam a uma aprendizagem mais eficaz
Descoberta em Modelos	O objetivo é desenvolver um modelo usando agrupamento, previsão ou engenharia de conhecimento, conforme um componente em um modelo mais abrangente de previsão ou mineração de relacionamento.	Descobrir as relações entre características dos alunos ou contexto variáveis e os comportamentos dos mesmos; Analisar questões de pesquisa em toda a variedade de contextos.
Destilação de Dados para Julgamento Humano	O principal objetivo deste modelo é encontrar uma nova maneira de permitir que os pesquisadores identifiquem e/ou classifiquem os recursos em dados mais facilmente.	Identificar a partir de conhecimento humano, padrões na aprendizagem dos alunos, comportamento ou colaboração; Descobrir dados de rotulagem para uso em desenvolvimento posterior do modelo de previsão.

Fonte: Elaborado pelo autor (2020).

evidenciadas.

4. Administradores (grupo foco deste trabalho): receber recursos e ferramentas para tomada de decisões, organização e coordenação das instituições.

2.2.1 Mineração de Relações

Historicamente, a técnica mais utilizada dentro da área de pesquisa de MDE, é a Mineração de Relações (ALGARNI, 2016). Ainda segundo Baker *et al.* (2011), a Mineração de Relações, possui as seguintes abordagens e definições:

1. Mineração de Regras de Associação (*Association Rule Mining*): tem por objetivo

encontrar regras que permitam a inferência de dados no estilo se-então, em que dado um conjunto de valores de variáveis assumidos, outra variável terá um valor específico, ou seja, caso uma condição seja verdadeira (por exemplo, a variável Y é verdade), e uma regra associe essa condição ao valor de uma outra variável (por exemplo, uma variável X), então Y poderá inferir em X. Por exemplo, ao analisar um conjunto de dados de domínio acadêmico, seria possível que a partir de variáveis como "objetivo aluno"(uma variável binária de valores: alcançado ou inalcançado), e uma outra variável "pedir ajuda ao professor"(também binária de valores: verdadeiro ou falso) fosse possível identificar uma regra que faça associação entre elas. Dado o contexto acima, se o aluno tem como objetivo aprender alguma matéria específica, mas está com dificuldade (i.e. a variável objetivo do aluno tem valor inalcançado), então é provável que ele peça ajuda do professor (i.e. a variável pedir ajuda ao professor seria verdade).

2. Mineração de Correlações (*Correlation Mining*): tem por objetivo encontrar correlações lineares (positivas ou negativas) entre variáveis do conjunto. Por exemplo, ao analisar um conjunto de dados de domínio acadêmico, seria possível identificar uma correlação positiva entre uma variável que indica a quantidade de tempo que um aluno passa com comportamentos que não estão relacionados aos deveres escolares durante o horário das aulas (e.g. conversas paralelas, brincadeiras e outras perturbações) e a nota que este aluno recebe na próxima prova.
3. Mineração de Padrões Sequenciais (*Sequential Pattern Mining*): tem por objetivo encontrar a associação temporal entre eventos e os impactos destes eventos nos valores das variáveis. Desta forma, é possível determinar qual trajetória de eventos e ações podem, de alguma maneira, levar a uma aprendizagem mais efetiva e/ou com melhores resultados.
4. Mineração de Causas (*Causal Mining*): para este método desenvolve-se algoritmos e técnicas para checar se um evento ou ação causa outro evento através da análise dos padrões de covariância ou usando informações sobre como um dos eventos foi acionado. Por exemplo, se um evento pedagógico é escolhido aleatoriamente usando experimentação automatizada, e frequentemente leva a um resultado positivo de aprendizagem, uma relação causal pode ser inferida.

Baker *et al.* (2011) cita um ótimo exemplo para a Mineração de Causas sob o domínio acadêmico:

> ... se considerarmos o exemplo anterior no qual um aluno externaliza comportamentos inadequados que não contribuem para resolver a tarefa dada pelo professor. Nesta situação o aluno, em muitos casos, recebe uma nota ruim na prova final. Desta, maneira, o comportamento do aluno pode ser a causa dele não aprender e, assim, resultando em uma performance ruim na prova. Contudo, pode ser que o aluno externalize tal comportamento inadequado devido a dificuldade em aprender, e portanto, a causa da performance ruim na prova não é o comportamento em si, mas sim a dificuldade de aprendizagem do aluno. Analisando o padrão de covariância, a mineração de causa pode inferir qual evento foi a causa do outro.

Para este estudo, será utilizada o método de Mineração de Correlações, pelo o qual, o algoritmo *CfsSubsetEval* foi escolhido para a descoberta de subconjuntos de atributos de acordo com maior correlação à classe do domínio.

A seleção do subconjunto de recursos é o processo de identificar e remover tanto irrelevantes e informações redundantes quanto possível. Isso reduz a dimensionalidade dos dados e pode permitem que os algoritmos de aprendizagem operem com mais rapidez e eficácia Hall (1999). Além disso, a seleção de atributos com base na correlação levará em consideração a capacidade preditiva individual de cada recurso e o grau de redundância entre eles para avaliar o valor do subconjunto de atributos. É preferível selecionar um subconjunto de recursos que sejam altamente relacionados à categoria e tenham baixa correlação cruzada (HALL *et al.*, 2009). O autor Hall *et al.* (2009) mostra claramente que associação é um termo em um sentido amplo, pois não se destina a referir-se especificamente à correlação linear clássica, mas para se referir ao grau de dependência ou previsibilidade de uma variável em relação a outra.

O *CfsSubsetEval* é um algoritmo que seleciona atributos com base na relevância por meio de filtragem simples, que classifica um subconjunto de atributos com base em uma função de avaliação heurística. A função de avaliação procurará subconjuntos que contenham recursos altamente relacionados a uma categoria ou classe, mas não relacionados a outra. Portanto, recursos insignificantes serão ignorados porque têm menor relevância Gnanambal *et al.* (2018). A classe utilizada como referência no presente estudo será a nota final do aluno no exame.

2.3 Exame Nacional de Desempenho do Estudante

Para que seja possível medir os índices de aprendizado do graduando e das Instituições de Ensino Superior, são necessárias a aplicação de sistemas de avaliação em larga escala, desenvolvidos pela esfera pública (PRIMI *et al.*, 2011). Essas informações são essenciais para a gestão dos recursos públicos, de maneira que seja possível perceber as falhas e virtudes do sistema para que ações interventivas e regulatórias sejam criadas com objetivo amplo de melhorar a qualidade do sistema (PRIMI *et al.*, 2010).

Em 2004, foi estabelecido o Sistema Nacional de Avaliação da Educação Superior (SINAES) com o objetivo de analisar o ensino superior no país. O SINAES qualifica a educação superior por meio de vários instrumentos, que focam na verificação da pesquisa, desempenho dos alunos, o corpo docente, gestão da instituição, responsabilidade social e as instalações (SINAES, 2020).

Esses processos são coordenados e supervisionados pela Comissão Nacional de Avaliação da Educação Superior (CONAES), que organiza e coordena externamente o método aplicado. Já a operacionalização é de responsabilidade do INEP, que desenvolve as práticas avaliativas que lhe forem atribuídas, no qual além de mensurar o aprendizado dos conteúdos propostos, procura examinar também os vários fatores que possam afetar a qualidade do ensino dos discentes (FONSECA; NAMEN, 2016).

A prova do ENADE (ENADE, 2020) é atualmente o meio utilizado para a análise dos estudantes de ensino superior no Brasil. O exame é obrigatório e apesar de ser aplicado todos os anos no país, é efetuado em um ciclo de 3 anos para todos os cursos de uma determinada área acadêmica. Esses cursos pertencentes às áreas são divididos em Ano I (formado por cursos que envolvem saúde, ciências agrárias, recursos naturais e afins), Ano II (formado por ciências exatas, licenciaturas, comunicação e produção industrial) e Ano III (formado por ciências sociais, ciências humanas e afins, cultura, designer e etc). A prova do ENADE consiste através dos seguintes pontos:

- Provas de conhecimento geral e específico para alunos de graduação ingressantes (já cursaram até 25% da grade curricular) e concluintes (mais de 75%).
- Questionários para dos discentes e coordenadores do curso.

Desta maneira, além dos testes de conhecimento, os alunos devem responder a um questionário para levantar a percepção dos alunos sobre o próprio teste e outro sobre o próprio

perfil socioeconômico educacional (LIMA *et al.*, 2019). O questionário aplicado aos discentes é chamado de Questionário do Estudante, principal artefato a ser explorado nesse trabalho.

2.3.1 Relatórios

Os relatórios do ENADE são disponibilizados anualmente pelo INEP, e compreendem Curso, IES e Síntese de Área, com estatísticas geradas a partir dos dados e agrupadas conforme a sua especificidade. Os relatórios Síntese de Área apresentam informações detalhadas a respeito da composição das provas, do desempenho e do perfil dos estudantes da área, da distribuição dos cursos no país, além de uma visão sobre o desempenho das instituições brasileiras no ENADE. Já os Relatórios de IES apresentam as informações a respeito do perfil e do desempenho dos estudantes de todos os cursos que participaram da prova vinculados à instituição. Em seguida, os Relatórios de Curso apresentam panorama do perfil e desempenho dos estudantes de cada um dos cursos que realizaram o exame (ARAÚJO *et al.*, 2019).

As análises disponibilizadas pelos relatórios se resumem a estatísticas descritivas dos dados, que visam descrever e resumir as informações coletadas. Porém, não são disponibilizadas pelo INEP análises mais sofisticadas, envolvendo estatística inferencial e/ou mineração de dados, que propõe abstrair informações relevantes a partir de grandes volumes de dados (LIMA *et al.*, 2019).

3 TRABALHOS RELACIONADOS

Para o embasamento deste trabalho foi realizada uma busca bibliográfica, com o objetivo de identificar outras pesquisas relacionadas. Esta Seção apresenta os estudos que se relacionam a este trabalho.

Silva *et al.* (2017) apresentam uma análise sob as variáveis que influenciam diretamente os alunos matriculados nos cursos pertencentes a área de ciências exatas que realizaram a prova do ENADE no ano de 2014, com o objetivo de que a partir dos resultados, fosse possível identificar e acertar os pontos a serem melhorados para alcançar um melhor desempenho. Os autores dividiram o conjunto de dados em duas bases específicas, com 31 variáveis do conjunto de amostras, denominadas de Modelo 1 e Modelo 2. Porém, para a base Modelo 2, foi aplicado o Stepwise, processo específico de seleção de dados relevantes.

Em ambas as bases, foram utilizados os conjuntos de testes para efetivar a MD através da técnica de Regressão Linear Múltipla. Como resultado, os autores, obtiveram as variáveis mais relevantes em relação a nota geral dos alunos, e após comparar os resultados dos dois modelos, identificou-se que o Modelo 1 era mais adequado, e, portanto, nesse caso, a utilização do método Stepwise para seleção das variáveis era irrelevante. No entanto, o trabalho restringe o conjunto de dados a somente uma edição do exame, possui um baixo número de variáveis analisadas, e não há dentre elas os atributos do Questionário do Estudante, mas apenas variáveis referentes à localização da instituição, notas da prova e descritivas do próprio aluno.

Cretton e Gomes (2016) apresentam uma análise dos dados do ENADE, especificamente na base do ENADE 2013, restringindo os dados aos estudantes de medicina, com o objetivo de encontrar dentre as variáveis selecionadas, as que possuem maior relevância em relação ao desempenho final do discente. Por isso, os autores tomaram como variáveis importantes para a análise, não somente o desempenho do aluno, mas também seu perfil e opinião sobre o nível de dificuldade da avaliação, uma vez que, quando sua opinião não condiz com seu rendimento, sugere uma possível falta de conhecimento ou confiança, sendo ambos os fatores essenciais para um médico. A técnica de MD utilizada pelos autores foi a árvore de decisão, no qual foi aplicada através do algoritmo J48, sob 6 variáveis com um total de 14.142 amostras. Como resultado, as variáveis mais relevantes sob o modelo foram reveladas, e notou-se um nível de confiança de 84% após avaliação da técnica utilizada, o que demonstra o potencial dos padrões gerados. No entanto, apenas uma edição foi utilizada, e apesar dos autores possuírem bons resultados a partir do algoritmo selecionado, devido ao foco do trabalho somente na área médica,

houve uma redução total de 97% de variáveis possíveis de serem analisadas e identificadas que poderiam ter igual ou maior relevância para o aprimoramento dos cursos, instituições e alunos.

Leão *et al.* (2018) apresentam a MD sob os dados coletados pelo Questionário do Estudante, formulário presente nas provas ENADE. O estudo escolheu como escopo às instituições vinculadas à Universidade Luterana do Brasil, sob as edições do ENADE de 2014 a 2016, cujo objetivo geral é identificar a as queixas mais importantes dos alunos, e, a partir delas, propor alternativas de melhoria para a instituição. Para a análise dos dados, foram escolhidas 22 variáveis, agrupadas pelos autores por identificadores (2 atributos), socioeconômicos (10 atributos), infraestrutura e instalações físicas (5 atributos), organização didática e pedagógica (3 atributos), oportunidades de treinamento acadêmico e profissional (2 atributos). O algoritmo Apriori foi escolhido para a descoberta de regras de associação. Como resultado, foram obtidas 753 regras de associação sob índice de confiança de 0.6, no qual foi possível identificar a descoberta de conhecimentos relevantes como: os fatores que avaliam questões relacionadas às práticas didático-pedagógicas, apesar de serem relatadas por alguns alunos, não representam o maior fator de insatisfação, mas sim as oportunidades de formação acadêmica e profissional. Comparado ao presente trabalho, este trabalho relacionado trata do mesmo grupo de técnicas de MD, inclui na análise os dados do Questionário do Estudante, classifica os atributos escolhidos e utilize mais de uma edição do ENADE, porém não houve relação entre os atributos selecionados, com a satisfação do estudante ao desempenho final do mesmo.

No Quadro 2 é possível analisar as comparações entre os diferenciais propostos por este trabalho, em relação aos trabalhos relacionados citados anteriormente.

Quadro 2 – Comparativo entre trabalhos relacionados e o presente estudo.

Características	Silva *et al.* (2017)	Cretton e Gomes (2016)	Leão *et al.* (2018)	Presente pesqusia
Utiliza mais de uma edição do ENADE			X	X
Compara modelos sob diferentes processos de aprendizagem	X			
Utiliza o Questionário do Estudante como parte das variáveis			X	X
Realiza filtro das amostras por cursos	X	X		X
Realiza filtro das amostras por IES			X	
Mapeia as variáveis ao desempenho do aluno	X	X		X

Fonte: Elaborado pelo autor (2020).

4 METODOLOGIA

Nesta Seção, são apresentados os procedimentos metodológicos para a execução deste trabalho. A Figura 2 apresenta os seguintes passos para a execução do trabalho: i) Coletar e analisar a estrutura os dados; ii) Escolher as edições e grupo de graduações que participarão da amostra; iii) Filtrar e classificar as questões a serem usadas no estudo; iv) Limpeza e normalização dos dados; v) Aplicar o algoritmo; e vi) Analisar e comparar os resultados.

4.1 Coletar e analisar a estrutura os dados

Para a coleta de dados, serão utilizados os microdados do ENADE, classificados como um dos demais dados abertos da educação e disponibilizados pelo site oficial do INEP. No total, ao observar as edições mais recentes a partir do ano de 2016, foram descobertos, em média, 150 variáveis distintas. As variáveis estão estruturadas em 9 partes, da seguinte maneira: i) Informações da instituição de ensino superior e do curso; ii) Informações do estudante; iii) Número de itens da parte objetiva; iv) Vetores; v) Tipos de presença; vi) Tipos de situação das questões da parte discursiva; vii) Notas na formação geral e componente específico; viii) Questionário de percepção da prova; ix) Questionário do estudante.

4.2 Escolher as edições e grupo de graduações que participarão da amostra

São disponibilizados pelo INEP, os microdados das edições de 2004 a 2018 do ENADE. Como mostrado anteriormente, o ENADE é aplicado todos os anos para determinadas áreas de conhecimento, num período cíclico de 3 anos para que uma determinada área refaça o exame. Para este trabalho, foram selecionadas as edições de 2014 e 2017, que compreendem as áreas formadas por ciências exatas, licenciaturas, comunicação e produção industrial.

4.3 Filtrar e identificar as questões a serem usadas no estudo

Dado o objetivo deste trabalho de identificar e classificar as questões de aspectos institucionais, relacionadas diretamente à administração das IES. Nesta seção, é proposto um filtro manual, para que seja possível selecionar somente os atributos relacionados diretamente às IES, assim como classificá-los de acordo com o que mais se caracterizam. Portanto, será possível que a partir dos resultados, as IES tenham ciência de quais aspectos poderiam ser melhorados ou

reavaliados para que melhor afetasse o desempenho do aluno no exame.

4.4 Limpar e normalizar os dados da amostra

Após finalizada a escolha das questões a serem usadas como amostra, é nessa etapa que será realizada a limpeza e normalização dos dados. A limpeza dos dados refere a remoção de dados irregulares ou corrompidos que comprometam a qualidade do conjunto, no qual neste contexto, trata-se principalmente da remoção de dados incompletos ou inconsistentes. A normalização dos dados redimensiona valores numéricos para um intervalo especificado, no qual neste contexto, é pretendido, principalmente, transformar os dados de forma linear em um intervalo 0 e 1, cujo o valor mínimo é dimensionado para 0 e o valor máximo para 1.

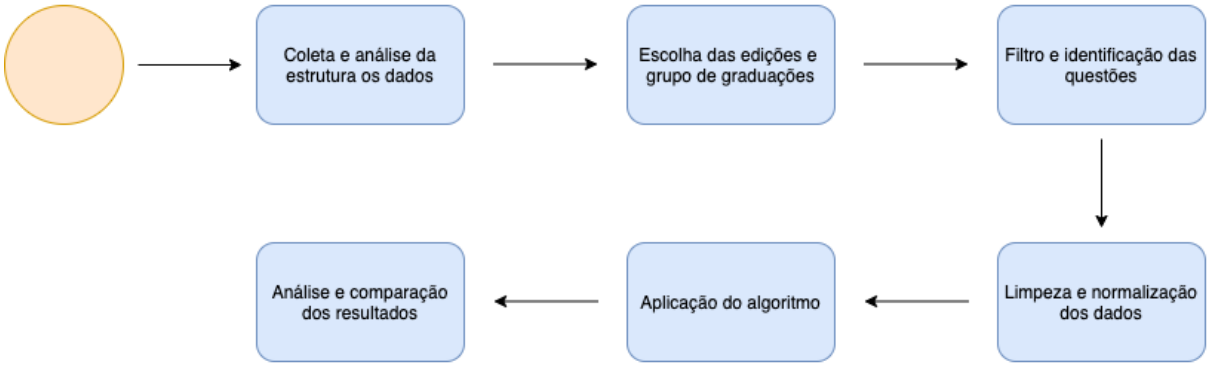
4.5 Aplicar o algoritmo

Com a amostra de dados limpa e normalizada, é nessa etapa que o conjunto de dados final será utilizado como entrada para a ferramenta WEKA que fará a aplicação do algoritmo abordado na pesquisa. Então, os atributos correlacionados gerados pelo algoritmo, serão finalmente sumarizados em um relatório que será utilizado na etapa seguinte.

4.6 Analisar e comparar os resultados

Após a geração do relatório, é nessa etapa que será analisado os resultados escolhidas e o desempenho final do aluno. Os resultados serão comparados entre as edições do exame, e entre cursos, no caso, incluem tanto os de nível tecnológico, quanto os de bacharelado.

Figura 2 – Procedimentos para a execução do trabalho



Fonte: Elaborada pelo autor (2020).

5 RESULTADOS

Nesta Seção, serão apresentados os resultados deste trabalho.

5.1 Coleta e análise dos dados

Após observar a estrutura dos dados do ENADE, percebeu-se que o Questionário do Estudante possui 80 variáveis das 150 existentes no escopo completo. Porém, 12 destas variáveis são exclusivamente destinadas aos cursos de licenciatura.

5.2 Edições de grupos de graduação selecionados

Para melhor acurácia do estudo, e melhor avaliação dos resultados futuros, foram escolhidos os cursos identificados como graduações relacionadas à área de Tecnologia da Informação (TI). Desta maneira, foram identificados, dentre todas os três grupos de graduação definidos pelo INEP (Ano I, Ano II e Ano III), 7 cursos de ensino superior caracterizados como pertencentes à área da TI, sendo eles: i) Ciência da Computação (Licenciatura); ii) Ciência da Computação (Bacharelado); iii) Sistemas de Informação; iv) Engenharia da Computação; v) Tecnologia em Análise e Desenvolvimento de Sistemas; vi) Tecnologia em Gestão da Tecnologia da Informação e vii) Tecnologia em Redes de Computadores. Para adequar os resultados as mesmas variáveis do questionário, foram descartados da pesquisa todos os cursos de licenciatura, e aqueles que não participaram de pelo menos duas edições do exame, restando apenas os de nível bacharelado e tecnólogo, somando um total de 6 cursos selecionados, como ilustrado no Quadro 3.

Após a identificação dos 6 cursos restantes, foram escolhidas as edições do exame a serem analisadas. Foi definido, para este estudo, que para cada curso, fossem escolhidas duas edições diferentes, de forma a permitir a comparação dos resultados entre as edições. Nesse contexto, foram escolhidas as edições mais recentes que pudessem englobar, ao menos suma maioria dos cursos selecionados, resultando nas edições do ENADE de 2017 e 2014. Como resultado, o curso de Tecnologia em Gestão da Tecnologia da Informação foi descartado, devido a ausência em um das edições do ENADE, a referente ao exame de 2014. Portanto, foram definidos 5 cursos a serem analisados, dentre eles: i) Ciência da Computação (Bacharelado); ii) Sistemas de Informação; iii) Engenharia da Computação; iv) Tecnologia em Análise e Desenvolvimento de Sistemas; v) Tecnologia em Redes de Computadores.

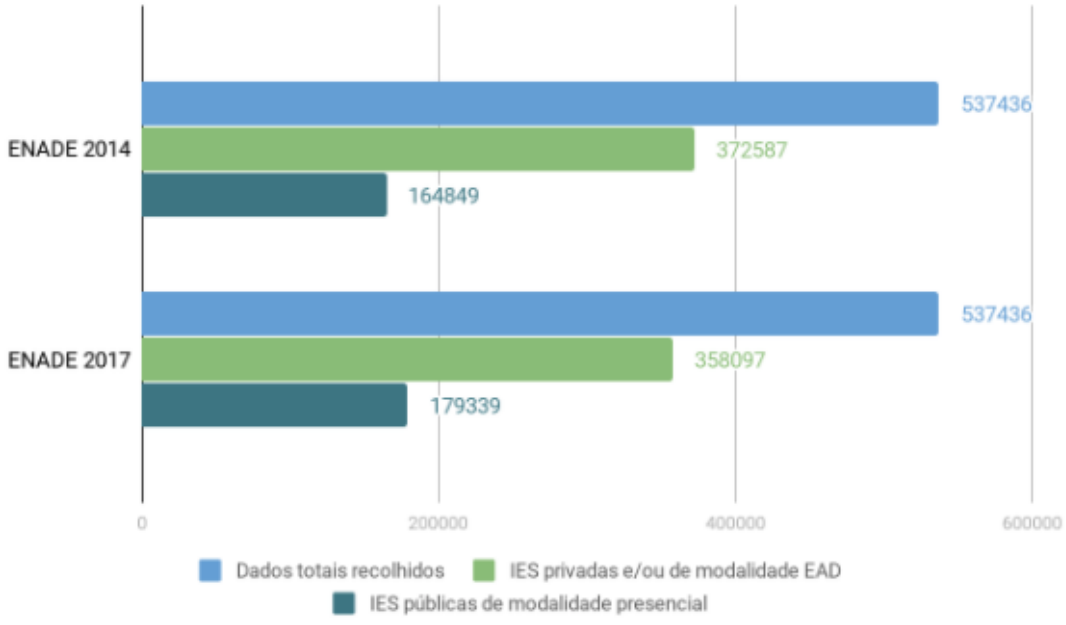
Quadro 3 – Cursos selecionados para análise da pesquisa

Grupo	**Objetivos**	**Aplicações**	**Descartado**
Ciência da Computação	Ano II	Licenciatura	SIM
Ciência da Computação	Ano II	Bacharelado	NÃO
Sistemas de Informação	Ano II	Bacharelado	NÃO
Tecnologia em Redes de Computadores	Ano II	Tecnólogo	NÃO
Tecnologia em Gestão da Tecnologia da Informação	Ano II	Tecnólogo	SIM
Tecnologia em Análise e Desenvolvimento de Sistemas	Ano II	Tecnólogo	NÃO
Engenharia da Computação	Ano II	Bacharelado	NÃO

Fonte: Elaborado pelo autor (2020).

Após o recolhimento dos dados, foi realizada a primeira etapa de pré-processamento, a fim de definir o escopo da aplicação. Primeiramente, foram escolhidos somente resultados de IES públicas cuja modalidade fosse exclusivamente presencial. Dessa forma, foram descartados dados estudantis de IES privadas (com ou sem fins lucrativos) e de modalidade EAD (Ensino a Distância). Ao analisar a Figura 3, é possível observar que o filtro do escopo reduziu os dados a serem analisados em 66.63% para a edição de 2017, e em 69.32% para a edição de 2014, respectivamente.

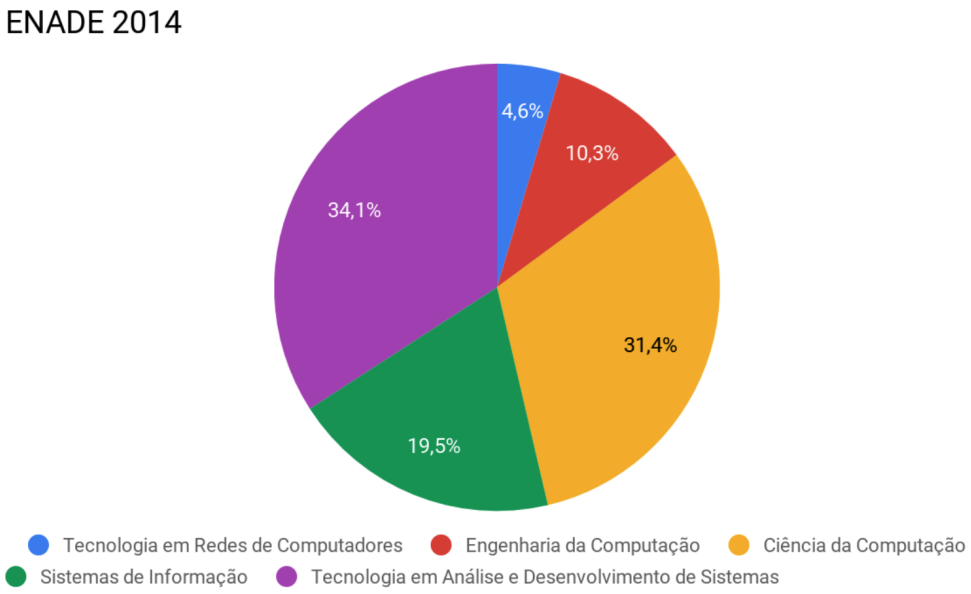
Figura 3 – Comparativo entre os dados das edições de 2017 e 2014 após o primeiro filtro.



Fonte: Elaborada pelo autor (2020).

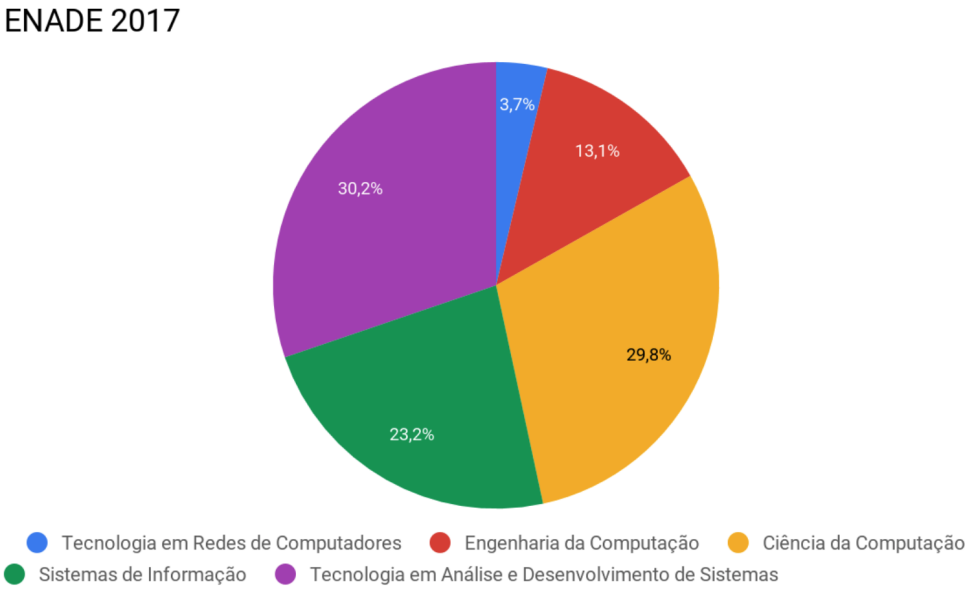
Também é possível visualizar, por meio da Figura 4 e Figura 5, as devidas proporções dos dados de cada curso de graduação selecionado das edições de 2014 e 2017, respectivamente.

Figura 4 – Gráfico comparativo entre os cursos selecionados sob a edição do ENADE 2014.



Fonte: Elaborada pelo autor (2020).

Figura 5 – Gráfico comparativo entre os cursos selecionados sob a edição do ENADE 2017.



Fonte: Elaborada pelo autor (2020).

5.3 Filtro e identificação dos atributos

No total, foram selecionados 44 atributos a serem analisados e utilizados nessa pesquisa. Dentre eles, 2 dos atributos foram selecionados como identificadores, sendo eles ano e curso, pelo o qual serão utilizados para fazer a separação dos dados e posterior comparação dos resultados de cada instituição ou grupo de graduação. Dessa maneira, os demais 42 atributos foram escolhidos a partir das questões presentes no Questionário do Estudante e identificados e agrupados em 5 categorias, sendo elas:

- Organização pedagógica: os atributos classificados como organização pedagógica são especialmente relacionados ao projeto pedagógico do curso, portanto são diretamente associados aos objetivos gerais do curso, peculiaridades, matriz curricular, carga horária das atividades didáticas e da integralização do curso.
- Planejamento didático: os atributos classificados como planejamento didático são relacionados às abordagens aplicadas dentro de sala de aula, ou ações especificamente executadas por professores. Por isso, não são caracterizados como parte do planejamento pedagógico do curso, mas sim como resultado da sua execução.
- Infraestrutura e facilidades físicas: os atributos classificados como infraestrutura e facilidades físicas são relacionados aos elementos estruturais que enquadram e suportam as atividades acadêmicas das IES, tendo como objetivo, garantir conforto e bem estar não apenas aos docentes, mas também para os discentes e a toda comunidade acadêmica relacionada.
- Treinamento acadêmico/profissional: os atributos classificados como treinamento acadêmico/profissional são relacionados às oportunidades de treinamento, seja com o foco no âmbito universitário ou de mercado, advindas diretamente e/ou indiretamente das IES ou parcerias entre IES e outros órgãos relacionados.
- Responsabilidade institucional: os atributos classificados como responsabilidade institucional são relacionados às questões de caráter extraordinário, ou seja, aquelas que não dependem exclusivamente da execução ou organização do planejamento pedagógico, mas sim de órgãos superiores aos de coordenação, pelo os quais são responsáveis por prover verba, incentivos culturais e atividades extra-curriculares.

Após a definição das categorizações a serem utilizadas neste estudo, os atributos do Questionário do Estudante foram devidamente agrupados nas 5 categorias definidas. É possível

visualizar a lista completa de todas as questões, bem como suas respectivas classificações e proporções, nos Quadros 4, 5, 6, 7 e 8. A partir dessa etapa, as classificações serão importantes para a análise das regras de associação geradas após o uso da MD sob o domínio proposto.

Quadro 4 – Atributos classificados como Infraestrutura e Facilidades Físicas.

Classificação (%)	Identificação	Descrição
Infraestrutura e facilidades físicas (16.6%)	QE_I59	A instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico.
	QE_I61	As condições de infraestrutura das salas de aula foram adequadas.
	QE_I62	Os equipamentos e materiais disponíveis para as aulas práticas foram adequados para a quantidade de estudantes.
	QE_I63	Os ambientes e equipamentos destinados às aulas práticas foram adequados ao curso.
	QE_I64	A biblioteca dispôs das referências bibliográficas que os estudantes necessitaram.
	QE_I65	A instituição contou com biblioteca virtual ou conferiu acesso a obras disponíveis em acervos virtuais.
	QE_I68	A instituição dispôs de refeitório, cantina e banheiros em condições adequadas que atenderam as necessidades dos seus usuários.

Fonte: Elaborado pelo autor (2020).

Quadro 5 – Atributos classificados como Organização Pedagógica.

Classificação (%)	Identificação	Descrição
Organização pedagógica (26.2%)	QE_I27	As disciplinas cursadas contribuíram para sua formação integral, como cidadão e profissional.
	QE_I28	Os conteúdos abordados nas disciplinas do curso favoreceram sua atuação em estágios ou em atividades de iniciação profissional.
	QE_I31	O curso contribuiu para o desenvolvimento da sua consciência ética para o exercício profissional.
	QE_I32	No curso você teve oportunidade de aprender a trabalhar em equipe.
	QE_I33	O curso possibilitou aumentar sua capacidade de reflexão e argumentação.
	QE_I34	O curso promoveu o desenvolvimento da sua capacidade de pensar criticamente, analisar e refletir sobre soluções para problemas da sociedade.
	QE_I35	O curso contribuiu para você ampliar sua capacidade de comunicação nas formas oral e escrita.
	QE_I36	O curso contribuiu para o desenvolvimento da sua capacidade de aprender e atualizar-se permanentemente.
	QE_I42	O curso exigiu de você organização e dedicação frequente aos estudos.
	QE_I47	O curso favoreceu a articulação do conhecimento teórico com atividades práticas.
	QE_I49	O curso propiciou acesso a conhecimentos atualizados e/ou contemporâneos em sua área de formação.

Fonte: Elaborado pelo autor (2020).

Quadro 6 – Atributos classificados como Planejamento Didático.

Classificação (%)	Identificação	Descrição
Planejamento didático (26.2%)	QE_I29	As metodologias de ensino utilizadas no curso desafiaram você a aprofundar conhecimentos e desenvolver competências reflexivas e críticas.
	QE_I30	O curso propiciou experiências de aprendizagem inovadoras.
	QE_I37	As relações professor-aluno ao longo do curso estimularam você a estudar e aprender.
	QE_I38	Os planos de ensino apresentados pelos professores contribuíram para o desenvolvimento das atividades acadêmicas e para seus estudos.
	QE_I39	As referências bibliográficas indicadas pelos professores nos planos de ensino contribuíram para seus estudos e aprendizagens.
	QE_I48	As atividades práticas foram suficientes para relacionar os conteúdos do curso com a prática, contribuindo para sua formação profissional.
	QE_I51	As atividades realizadas durante seu trabalho de conclusão de curso contribuíram para qualificar sua formação profissional
	QE_I55	As avaliações da aprendizagem realizadas durante o curso foram compatíveis com os conteúdos ou temas trabalhados pelos professores.
	QE_I56	Os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.
	QE_I57	Os professores demonstraram domínio dos conteúdos abordados nas disciplinas.
	QE_I58	Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem).

Fonte: Elaborado pelo autor (2020).

Quadro 7 – Atributos classificados como Responsabilidade Institucional.

Classificação (%)	Identificação	Descrição
Responsabilidade institucional (12%)	QE_I41	A coordenação do curso esteve disponível para orientação acadêmica dos estudantes.
	QE_I54	Os estudantes participaram de avaliações periódicas do curso (disciplinas, atuação dos professores, infraestrutura).
	QE_I60	O curso disponibilizou monitores ou tutores para auxiliar os estudantes.
	QE_I66	As atividades acadêmicas desenvolvidas dentro e fora da sala de aula possibilitaram reflexão, convivência e respeito à diversidade.
	QE_I67	A instituição promoveu atividades de cultura, de lazer e de interação social.

Fonte: Elaborado pelo autor (2020).

Quadro 8 – Atributos classificados como Treinamento acadêmico/profissional.

Classificação (%)	Identificação	Descrição
Treinamento acadêmico/profissional (19%)	QE_I40	Foram oferecidas oportunidades para os estudantes superarem dificuldades relacionadas ao processo de formação.
	QE_I43	Foram oferecidas oportunidades para os estudantes participarem de programas, projetos ou atividades de extensão universitária.
	QE_I44	Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica.
	QE_I45	O curso ofereceu condições para os estudantes participarem de eventos internos e/ou externos à instituição.
	QE_I46	A instituição ofereceu oportunidades para os estudantes atuarem como representantes em órgãos colegiados.
	QE_I50	O estágio supervisionado proporcionou experiências diversificadas para a sua formação.
	QE_I52	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios no país.
	QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.

Fonte: Elaborado pelo autor (2020).

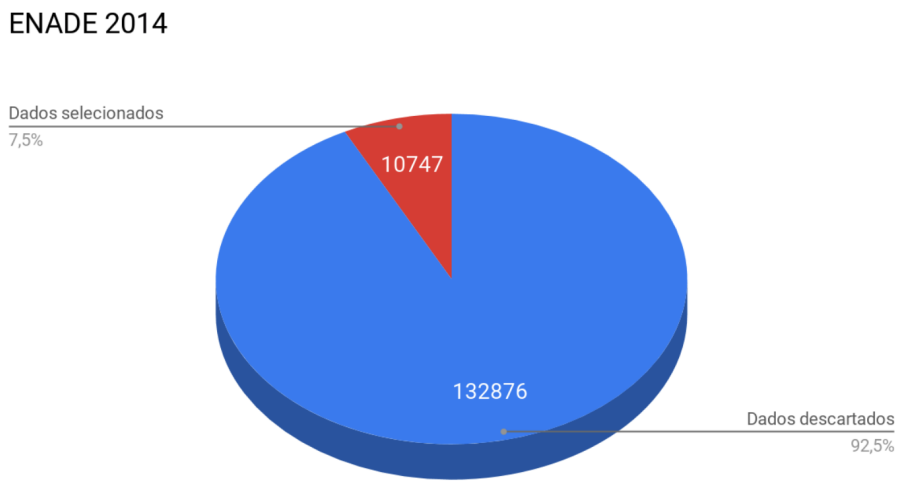
5.4 Limpeza e normalização da amostra

Após definir o escopo dos dados, foram removidos ainda todos os resultados dos estudantes que não cursaram um dos 5 cursos previamente escolhidos. Além disso, outro critério crucial para a limpeza dos dados deu-se a exclusão daqueles que não responderam ao menos uma das 44 questões identificadas anteriormente e presente no Questionário do Estudante, visto que o questionário é opcional para quem fez a prova, independentemente da edição.

É possível observar, a partir da Figura 6 e Figura 7, uma expressiva redução dos dados após serem selecionados somente aqueles que relacionados aos cursos pertinentes ao estudo, em comparação aos dados totais das edições de 2014 e 2017, respectivamente.

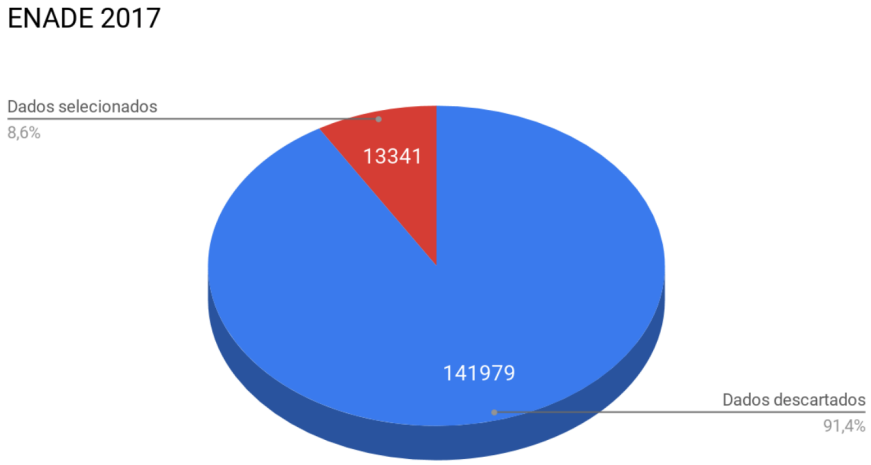
Em seguida, um dos passos relevantes para a normalização dos dados deu-se na

Figura 6 – Gráfico comparativo entre os dados filtrados por cursos selecionados e dados restantes sob a edição do ENADE 2014.



Fonte: Elaborada pelo autor (2020).

Figura 7 – Gráfico comparativo entre os dados filtrados por cursos selecionados e dados restantes sob a edição do ENADE 2017.



Fonte: Elaborada pelo autor (2020).

conversão da nota geral dos alunos, identificada como NET-GR, para números reais. Além disso, cerca de 16% das notas encontravam-se em formato não numérico, com virgulas sendo usadas como separador decimal, ao invés de ponto final, e portanto fazendo desta conversão extremamente necessária considerando que os dados seriam usados como valores matemáticos para processamento em passos futuros.

Outra alteração importante foi a normalização do dicionário de respostas. Para cada questão selecionada, ou seja, para cada resposta referente a um dos 44 atributos selecionados,

existem cerca de 8 possíveis repostas conforme o Quadro 9.

Quadro 9 – Dicionário de respostas do Questionário do Estudante

Resposta	**Valor numérico associado**
Discordo totalmente	1
Discordo	2
Discordo parcialmente	3
Concordo parcialmente	4
Concordo	5
Concordo totalmente	6
Não se aplica	7
Não sei responder	8

Fonte: Elaborado pelo autor (2020).

Portanto, para normalização do dicionário de respostas, primeiramente foram excluídos os dados estudantis daqueles que responderam "Não se aplica"em ao menos um dos 44 atributos selecionados, pois tal valor não se faz relevante ao intuito da pesquisa. Outra ação de normalização necessária foi alterar o valor associado da resposta "Não sei responder"para 0 (zero), a fim de que resposta em si pudesse ter menor relevância matemática que as demais. Dessa forma, ao final da etapa de normalização o dicionário de respostas normalizado se deu conforme o Quadro 10.

Quadro 10 – Dicionário de respostas do Questionário do Estudante após normalização

Resposta	**Valor numérico associado**
Não sei responder	0
Discordo totalmente	1
Discordo	2
Discordo parcialmente	3
Concordo parcialmente	4
Concordo	5
Concordo totalmente	6

Fonte: Elaborado pelo autor (2020).

5.5 Aplicação do algoritmo

Após o filtro dos dados e classificação dos atributos, foi utilizada a ferramenta Weka para fazer as primeiras inferências sobre os dados coletados. Nessa pesquisa, apesar de todas as

graduações selecionadas fazerem parte do mesma área de Tecnologia da Informação, os dados serão usados e analisados separadamente por curso de graduação e ano de exame do ENADE.

Foi utilizado nessa pesquisa o algoritmo de seleção de atributos CsfSubsetVal (*Correlation Based Feature Selection*) com o método de pesquisa *GreedyStepWise*. O foco da seleção de *features* é selecionar um subconjunto de variáveis que podem descrever com eficiência o próprio conjunto, além de escolher um subconjunto eliminando dados considerados não preditivos ou não co-relacionados com variável escolhida, no nosso caso a nota final do aluno.

Como mencionado anteriormente foi utilizada nessa etapa da pesquisa o Weka. A ferramenta é de fácil instalação e uso. A mesma já conta com vários métodos de *Machine Learning* previamente implementados, incluindo o algorítimo utilizado nessa pesquisa. Após instalado, basta fazer o envio do conjunto de dados a serem processados, selecionar a página Seleção de Atributos entre as opções da aplicação e escolher o algorítimo de avaliação e método de pesquisa desejado, nesta etapa foram selecionados CsfSubsetVal e GreedyStepWise respectivamente. Por último, após a escolha dos métodos utilizados para a seleção dos atributos, foi selecionado o atributo de referência cujo a pesquisa busca encontrar correlação, dentre todos os atributos, a variável escolhida para todos os cursos e edições foi a nota final do aluno.

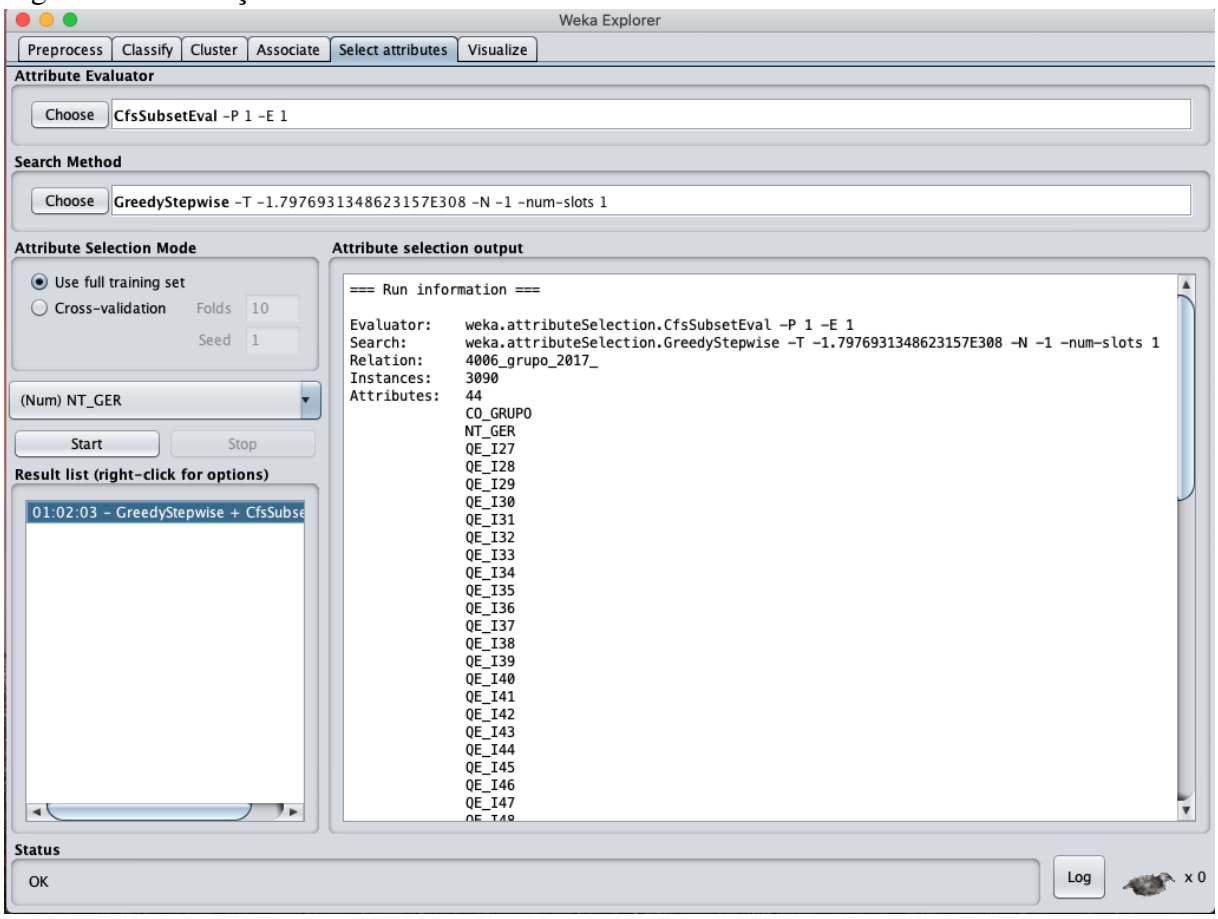
É possível observar na Figura 8 o resultado final do processo de uso do Weka para um dos conjuntos de dados utilizados nesse estudo.

Sob o uso do Weka foram recolhidos os subconjuntos de variáveis selecionados pelo algoritmo, sendo estes agrupados por curso comentados a seguir. A definição do significado de cada variável pode ser revisto nos nos Quadros 4, 5, 6, 7 e 8. Posteriormente na sessão 5.6.1 os resultados serão discutidos com mais detalhes.

5.5.1 Tecnologia em Análise e Desenvolvimento de Sistemas

Para o curso de Tecnologia em Análise e Desenvolvimento de Sistemas foram analisados 4032 e 3669 instâncias de dados para as edições de 2017 e 2014 respectivamente. Após o processamento do algoritmo para a edição de 2017 foram selecionados somente 3 atributos, sendo eles QE_I28, QE_I56 e QE_I58. Já para a edição de 2014, foram selecionados os atributos QE_I28, QE_I53, QE_I55, QE_I56 e QE_I59.

Figura 8 – Utilização do WEKA.



Fonte: Elaborada pelo autor (2020).

5.5.2 Tecnologia em Rede de Computadores

Para o curso de Tecnologia em Rede de Computadores foram analisados 495 e 496 instâncias de dados para as edições de 2017 e 2014 respectivamente. Após o processamento do algoritmo para a edição de 2017 foram selecionados os atributos QE_I46, QE_I54, QE_I56 e QE_I62. Já para a edição de 2014, foram selecionados os atributos QE_I53 e QE_I58.

5.5.3 Sistemas de Informação

Para o curso de Sistema de Informação foram analisados 3090 e 3377 instâncias de dados para as edições de 2017 e 2014 respectivamente. Após o processamento do algoritmo para a edição de 2017 foram selecionados os atributos QE_I44, QE_I53, QE_I56, QE_I57, QE_I58 e QE_I59. Já para a edição de 2014, foram selecionados os atributos QE_I44, QE_I53 e QE_I58.

5.5.4 Ciência da Computação

Para o curso de Ciência da Computação foram analisados 3972 e 2908 instâncias de dados para as edições de 2017 e 2014 respectivamente. Após o processamento do algoritmo para a edição de 2017 foram selecionados os atributos QE_I27, QE_I43, QE_I44, QE_I50, QE_I53, QE_I58, QE_I60, QE_I63 e QE_I64. Já para a edição de 2014, foram selecionados os atributos QE_I44, QE_I53, QE_I58 e QE_I63.

5.5.5 Engenharia da Computação

Para o curso de Engenharia da Computação foram analisados 1752 e 1107 instâncias de dados para as edições de 2017 e 2014 respectivamente. Após o processamento do algoritmo para a edição de 2017 foram selecionados somente 3 atributos, sendo eles QE_I27, QE_I44 e QE_I53. Já para a edição de 2014, foram selecionados os atributos QE_I44, QE_I53 e QE_I61.

5.6 Análise e comparação de resultados

Nessa etapa, iremos comentar sobre os grupos selecionados, observando a relação entre as edições e cursos escolhidos, além da análise sob a classificação proposta nesse estudo.

5.6.1 Avaliação geral

Sob a análise de todos os cursos e edições, é possível notar atributos que foram comumente selecionados entre os conjuntos. O atributo QE_I53 cuja descrição é: "Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país", esteve presente ao menos em uma edição em todos os cursos abordados nesse estudo, principalmente ao se tratar da edição de 2014, onde para todos os cursos, o mesmo sempre fora escolhido pelo algoritmo. Outro atributo a ser observado é o atributo QE_I58, cuja descrição é: "Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem)", observa-se presente entre maioria dos cursos, com exceção somente do curso de Engenharia da Computação, no qual não fora selecionado em nenhuma das edições de 2014 e 2017.

Sobre as classificações dos atributos, foi possível observar que as questões mais relevantes para a nota do aluno foram associados aos grupos de Treinamento acadêmico/pro-

fissional e Planejamento Didático, no qual juntos englobam mais de 65% de todos os atributos selecionados pelo algoritmo.

5.6.2 Tecnologia em Análise e Desenvolvimento de Sistemas

No total, para o curso de Tecnologia em Análise e Desenvolvimento de Sistemas, foram selecionados 6 atributos distintos. No Quadro 11, é possível observar identificação, significados e categorização de cada um. Vale a pena observar, que os atributos QE_I28 e QE_I56 foram repetidos em ambas as edições, permanecendo relevantes para a nota final do aluno. Outro ponto é a relevância das questões classificadas como Planejamento Didático sob o curso, no qual das 6 questões selecionados, 50% eram somente de Planejamento Didático, enquanto todos os os outros 3 possuíam classificações distintas, sendo elas Organização Pedagógica, Treinamento acadêmico/profissional e, Infraestrutura e Facilidades Físicas.

Quadro 11 – Atributos selecionados para o curso de Tecnologia em Análise e Desenvolvimento de Sistemas

Identificação	**Descrição**	**Grupo**	**Edições**
QE_I28	Os conteúdos abordados nas disciplinas do curso favoreceram sua atuação em estágios ou em atividades de iniciação profissional.	Organização Pedagógica.	2014 e 2017
QE_I56	Os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.	Planejamento Didático.	2014 e 2017
QE_I58	Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem).	Planejamento Didático.	2017
QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.	Treinamento acadêmico/profissional.	2014
QE_I55	As avaliações da aprendizagem realizadas durante o curso foram compatíveis com os conteúdos ou temas trabalhados pelos professores.	Planejamento Didático.	2014
QE_I59	A instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico.	Infraestrutura e Facilidades Físicas.	2014

Fonte: Elaborado pelo autor (2020).

5.6.3 Tecnologia em Rede de Computadores

No total, para o curso de Tecnologia em Rede de Computadores, também foram selecionados 6 atributos distintos. No Quadro 12, é possível observar identificação, significados e categorização de cada um. Desta vez, todos foram distintos entre as edições, ou seja, não houve repetição entre os dois exames. Dentre eles, os atributos QE_I53 e QE_I58, por sinal os mais comumente relacionados a todos os cursos, foram selecionados somente na edição de 2014. Outro ponto de observação é que de entre todos os cursos, Tecnologia em Rede de Computadores foi o único a selecionar QE_I62 como uma dos principais atributos, muito provavelmente devido ao seu caráter técnico específico e que dentre os cursos de TI, talvez seja o que mais demande de infraestrutura e materiais extras, além obviamente de recursos mais comuns à área como computadores e *software* diversos.

Quadro 12 – Atributos selecionados para o curso de Tecnologia em Rede de Computadores

Identificação	**Descrição**	**Grupo**	**Edições**
QE_I46	A instituição ofereceu oportunidades para os estudantes atuarem como representantes em órgãos colegiados.	Treinamento acadêmico/profissional.	2017
QE_I54	Os estudantes participaram de avaliações periódicas do curso (disciplinas, atuação dos professores, infraestrutura).	Responsabilidade Institucional.	2017
QE_I56	Os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.	Planejamento Didático.	2017
QE_I62	Os equipamentos e materiais disponíveis para as aulas práticas foram adequados para a quantidade de estudantes.	Infraestrutura e Facilidades Físicas.	2017
QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.	Treinamento acadêmico/profissional.	2014
QE_I58	Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem).	Planejamento Didático.	2014

Fonte: Elaborado pelo autor (2020).

5.6.4 Sistemas de Informação

No total, para o curso de Sistemas de Informação, também foram selecionados 6 atributos distintos. No Quadro 13, é possível observar identificação, significados e categorização de cada um. É possível observar que todos os 3 atributos selecionados na edição de 2014, sendo eles QE_I44, QE_I53 e QE_I58, permaneceram presentes na edição de 2017, evidenciando consistência e permanecendo relevantes para a nota final do aluno em ambas as edições. Desta vez, 5 das 6 questões selecionados são classificados como Planejamento didático ou Treinamento acadêmico/profissional, mostrando à relevância dos dois grupos para o curso de graduação.

Quadro 13 – Atributos selecionados para o curso de Sistemas de Informação

Identificação	**Descrição**	**Grupo**	**Edições**
QE_I44	Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica.	Treinamento acadêmico/profissional.	2014 e 2017
QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.	Treinamento acadêmico/profissional.	2014 e 2017
QE_I58	Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem).	Planejamento Didático.	2014 e 2017
QE_I56	Os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.	Planejamento Didático.	2014
QE_I57	Os professores demonstraram domínio dos conteúdos abordados nas disciplinas.	Planejamento Didático.	2014
QE_I59	A instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico.	Infraestrutura e Facilidades Físicas.	2014

Fonte: Elaborado pelo autor (2020).

5.6.5 Ciência da Computação

No total, para o curso de Ciência da Computação, foram selecionados 9 atributos distintos. No Quadro 14, é possível observar identificação, significados e categorização de cada um. O curso de Ciência da Computação foi o que obteve o maior número de questões

selecionadas dentre os cursos analisados nesse estudo. Uma das possibilidades de explicação para isso talvez seja devido à uniformidade da graduação, uma vez que dentre as graduações do escopo, o curso é o mais antigo, conciso entre as instituições de ensino, e com um dos maiores números de instâncias analisadas. Outro ponto levantado para o devido resultado, é que dentre as demais graduações, Ciência da Computação é comumente mais teórico, acadêmico e tende a abranger as principais áreas da TI, inclusive aquelas nos quais os outros cursos tem como espacialidade.

Quadro 14 – Atributos selecionados para o curso de Ciência da Computação

Identificação	**Descrição**	**Grupo**	**Edições**
QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.	Treinamento acadêmico/profissional.	2014 e 2017
QE_I58	Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem).	Planejamento Didático.	2014 e 2017
QE_I63	Os ambientes e equipamentos destinados às aulas práticas foram adequados ao curso.	Infraestrutura e Facilidades Físicas.	2014 e 2017
QE_I27	As disciplinas cursadas contribuíram para sua formação integral, como cidadão e profissional.	Organização Pedagógica.	2017
QE_I43	Foram oferecidas oportunidades para os estudantes participarem de programas, projetos ou atividades de extensão universitária.	Treinamento acadêmico/profissional.	2017
QE_I50	O estágio supervisionado proporcionou experiências diversificadas para a sua formação.	Treinamento acadêmico/profissional.	2017
QE_I60	O curso disponibilizou monitores ou tutores para auxiliar os estudantes.	Responsabilidade Institucional	2017
QE_I64	A biblioteca dispôs das referências bibliográficas que os estudantes necessitaram.	Infraestrutura e Facilidades Físicas.	2017
QE_I44	Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica.	Treinamento acadêmico/profissional.	2014

Fonte: Elaborado pelo autor (2020).

Já sob os atributos, é possível observar que 3 dos 4 atributos selecionados na edição de 2014, sendo eles QE_I53, QE_I58 e QE_I63, permaneceram presentes na edição de 2017,

evidenciando consistência e permanecendo relevantes para a nota final do aluno em ambas as edições. Vale ressaltar o destaque para o atributo QE_I63, que além de ter sido selecionado somente nesta graduação, também foi considerado importante para a nota do aluno em ambas as edições analisadas. O curso também foi o único a compreender questões em todas as 5 classificações levantadas nessa pesquisa, e levantar tópicos até então não selecionados em outros cursos, mas comumente discutidos em âmbito acadêmico como pontos chaves para o sucesso de uma graduação, tais como disponibilização de monitores ou tutores para auxiliar nas disciplinas, estágio supervisionado, projetos de extensão e acervo bibliográfico.

5.6.6 Engenharia da Computação

No total, para o curso de Engenharia da Computação, foram selecionados apenas 4 atributos distintos. No Quadro 15, é possível observar identificação, significados e categorização de cada um. Foram 2 atributos repetidos entre as ambas as edições, são eles QE_I44 e QE_I53. O curso foi o que obteve menor número de atributos selecionados pelo algoritmo, o que nos leva a pressupor que para os dados utilizados do curso, não foi possível encontrar um subconjunto de grupos de significativa relevância para a nota final do aluno. Acredita-se que por se tratar de um curso relativamente novo, com poucas instâncias, e que a depender da instituição de ensino possa ter o foco em diferentes áreas, seja ela em computação ou engenharia elétrica por exemplo, os dados apesar de agrupados podem possuir contextos diferentes e portanto não dizer muito sobre o conjunto total.

Quadro 15 – Atributos selecionados para o curso de Engenharia da Computação

Identificação	**Descrição**	**Grupo**	**Edições**
QE_I44	Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica.	Treinamento acadêmico/profissional.	2014 e 2017
QE_I53	Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.	Treinamento acadêmico/profissional.	2014 e 2017
QE_I27	As disciplinas cursadas contribuíram para sua formação integral, como cidadão e profissional.	Organização Pedagógica.	2017
QE_I61	As condições de infraestrutura das salas de aula foram adequadas.	Infraestrutura e Facilidades Físicas.	2014

Fonte: Elaborado pelo autor (2020).

6 CONCLUSÕES E TRABALHOS FUTUROS

O desenvolvimento do presente estudo possibilitou uma análise de quais questões presentes no Questionário do Estudante, e pertinentes somente à caráter institucional, poderiam ter maior correlação com a nota final do aluno de graduação na prova do ENADE. Além disso, o estudo também permitiu a classificação das questões em diferentes grupos, com o intuito de também observar a correlação entre a nota e o contexto das questões, e não somente ao significado isolado de cada atributo.

Ainda sobre as classificações dos atributos, ao olhar sob a ótica geral de todos os cursos analisados e edições selecionadas, foi possível observar que as questões mais relevantes para a nota do aluno foram associados aos grupos de Treinamento acadêmico/profissional e Planejamento Didático. Pressupõe-se que a constante correlação entre os dois grupos e os cursos escolhidos, em especial aos atributo referentes à utilização de TICs como estratégia de ensino e oportunidades de estágio nacional e internacional, seja dada pela direta relação da área com a tecnologia e a internacionalização do conhecimento, além da comum demanda por aulas práticas, acesso à computadores e laboratórios de informática como meio eficaz para alcançar os objetivos pedagógicos dá área. Tais resultados permitem auxiliar gestores da área a entender quais pontos merecem atenção durante a trajetória do curso, como exemplo, para o curso de Sistema de Informação, o investimento em características relacionadas ao Planejamento didático ou Treinamento acadêmico/profissional trariam os resultados de maior impacto ao curso.

Apesar da análise possuir interesse em se aproximar a uma correlação perfeita entre atributos e a nota do aluno, não é possível dizer com certeza a relação com o mundo real, já que o subconjunto de dados utilizado nessa pesquisa limitou-se somente a questões de caráter institucional, desconsiderando atributos de aspectos sociais, familiares e econômicos do aluno, que outrora poderiam ter inclusive maior relevância sobre os dados institucionais. Outro aspecto que tende a influenciar os resultados são agrupamentos de caráter geográfico, nesse estudo o conjunto utilizado não limitou-se a analisar cursos por região, cidade, de caráter metropolitano ou interiorano, e portanto, os dados foram analisados sob o mesmo peso.

Como trabalho futuro pretende-se reduzir o escopo a fim de conseguir um melhor conjunto de atributos relevantes à nota do aluno, como a escolha de um único curso específico, análise dos dados de instituições que tiveram presentes nos dois exames, e possuem óticas similares como abordagem pedagógica e contexto geográfico. Desta maneira seria possível ir mais fundo e com mais precisão no aprendizado de máquina, a fim de tentar prever a partir de

tais dados, quanto de investimento em determinadas áreas seria essencial para que tal graduação alcançasse a excelência desejada.

REFERÊNCIAS

ADDERLEY, R. The use of data mining techniques in operational crime fighting. In: SPRINGER. **International Conference on Intelligence and Security Informatics**. [*S. l.*], 2004. p. 418–425.

ALGARNI, A. Data mining in education. **International Journal of Advanced Computer Science and Applications**, [*S. l.*], v. 7, n. 6, p. 456–461, 2016.

ARAÚJO, R. A. *et al.* **Análise dos microdados do Enade**: Proposta de uma ferramenta de exploração utilizando mineração de dados. [*S. l.*]: Universidade Federal de Goiás, 2019.

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

BAKER, R. *et al.* Data mining for education. **International encyclopedia of education**, Elsevier Oxford, UK, [*S. l.*], v. 7, n. 3, p. 112–118, 2010.

BAKER, R. S.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **JEDM| Journal of Educational Data Mining**, [*S. l.*], v. 1, n. 1, p. 3–17, 2009.

BAKHSHINATEGH, B.; ZAIANE, O. R.; ELATIA, S.; IPPERCIEL, D. Educational data mining applications and tasks: A survey of the last 10 years. **Education and Information Technologies**, Springer, [*S. l.*], v. 23, n. 1, p. 537–553, 2018.

BAUER, A. É possível relacionar avaliação discente e formação de professores? a experiência de são paulo. **Educação em revista**, SciELO Brasil, [*S. l.*], v. 28, n. 2, p. 61–82, 2012.

BERKHIN, P. A survey of clustering data mining techniques. In: **Grouping multidimensional data**. [*S. l.*]: Springer, 2006. p. 25–71.

BRAMER, M. **Principles of data mining**. [*S. l.*]: Springer, 2007. v. 180.

CABENA, P.; CHOI, H. H.; KIM, I. S.; OTSUKA, S.; REINSCHMIDT, J.; SAARENVIRTA, G. Intelligent miner for data applications guide. **IBM RedBook SG24-5252-00**, [*S. l.*], v. 173, 1999.

CHEN, M.; MAO, S.; LIU, Y. Big data: A survey. **Mobile networks and applications**, Springer, [*S. l.*], v. 19, n. 2, p. 171–209, 2014.

CRETTON, N. N.; GOMES, G. R. Aplicação de técnicas de mineração de dados na base de dados do enade com enfoque nos cursos de medicina. **Acta Biomedica Brasiliensia**, [*S. l.*], v. 7, n. 1, p. 74–89, 2016.

ENADE. 2020. Disponível em: http://portal.inep.gov.br/enade. Acesso em: 18 maio. 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, [*S. l.*], v. 17, n. 3, p. 37–37, 1996.

FONSECA, S. O. d.; NAMEN, A. A. Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. **Educação em Revista**, SciELO Brasil, [*S. l.*], v. 32, n. 1, p. 133–157, 2016.

GERA, M.; GOEL, S. Data mining-techniques, methods and algorithms: A review on tools and their validity. **International Journal of Computer Applications**, Citeseer, [*S. l.*], v. 113, n. 18, 2015.

GNANAMBAL, S.; THANGARAJ, M.; MEENATCHI, V.; GAYATHRI, V. Classification algorithms with attribute selection: an evaluation study using weka. **International Journal of Advanced Networking and Applications**, Eswar Publications, [*S. l.*], v. 9, n. 6, p. 3640–3644, 2018.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, [*S. l.*], v. 1, n. 1, p. 20–33, 1999.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, [*S. l.*], v. 11, n. 1, p. 10–18, 2009.

HALL, M. A. **Correlation-based feature selection for machine learning**. [*S. l.*]: University of Waikato Hamilton, 1999.

HAN, J.; FU, Y.; WANG, W.; CHIANG, J.; GONG, W.; KOPERSKI, K.; LI, D.; LU, Y.; RAJAN, A.; STEFANOVIC, N. *et al.* Dbminer: A system for mining knowledge in large relational databases. In: **KDD**. [*S. l.*: *s. n.*], 1996. v. 96, p. 250–255.

HAN, J.; PEI, J.; KAMBER, M. **Data mining**: concepts and techniques. [*S. l.*]: Elsevier, 2011.

HANNA, M. Data mining in the e-learning domain. **Campus-wide information systems**, Emerald group publishing limited, [*S. l.*], 2004.

IDC, I. D. C. **The Digitization of the World From Edge to Core**. Framingham, USA: [*S. n.*], 2018. Disponível em: ttps://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 14 abr. 2020.

INEP, M. da Educação. Instituto Nacional de Estudos e P. E. A. T. **PDE/PROVA BRASIL Plano de Desenvolvimento da Educação**. Brasília, DF: [*S. n.*], 2011. Disponível em: http://portal.mec.gov.br/dmdocuments/prova/brasil_matriz2.pdf. Acesso em: 15 abr. 2020.

LEÃO, H. A. T.; CANEDO, E. D.; LADEIRA, M.; FAGUNDES, F. Mining enade data from the ulbra network institution. In: **Information Technology-New Generations**. [*S. l.*]: Springer, 2018. p. 287–294.

LIMA, P. d. S. N.; AMBRÓSIO, A. P. L.; FERREIRA, D. J.; BRANCHER, J. D. Análise de dados do enade e enem: uma revisão sistemática da literatura. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, SciELO Brasil, [*S. l.*], v. 24, n. 1, p. 89–107, 2019.

MARTUCCI, E. M. Informação para educação: os novos cenários para o ensino fundamental. **Informação & Sociedade**, Universidade Federal da Paraíba-Programa de Pós-Graduação em Ciência da . . . , [*S. l.*], v. 10, n. 2, 2000.

OLSON, D. L.; DELEN, D. **Advanced data mining techniques**. [*S. l.*]: Springer Science & Business Media, 2008.

PRIMI, R.; CARVALHO, L. F. d.; MIGUEL, F. K.; SILVA, M. C. R. d. Análise do funcionamento diferencial dos itens do exame nacional do estudante (enade) de psicologia de 2006. **Psico-USF**, SciELO Brasil, [*S. l.*], v. 15, n. 3, p. 379–393, 2010.

PRIMI, R.; HUTZ, C. S.; SILVA, M. C. R. da. A prova do enade de psicologia 2006: concepção, construção e análise psicométrica da prova. **Avaliação Psicológica**, Instituto Brasileiro de Avaliação Psicológica, [*S. l.*], v. 10, n. 3, p. 271–294, 2011.

REIS, E. A.; REIS, I. A. Análise descritiva de dados. **Síntese numérica Estatística**, [*S. l.*], 2002.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert systems with applications**, Elsevier, [*S. l.*], v. 33, n. 1, p. 135–146, 2007.

SILVA, L. F.; ROCHA, M. E. P. S. da; FAGUNDES, R. A. de A. Enade: Math and science students' performance analysis. **IEEE Latin America Transactions**, IEEE, [*S. l.*], v. 15, n. 9, p. 1742–1746, 2017.

SINAES. 2020. Disponível em: http://portal.inep.gov.br/web/guest/sinaes. Acesso em: 18 maio. 2020.

SULLIVAN-BOLYAI, S.; BOVA, C. Data analysis: Descriptive and inferential statistics. **Nursing Research-E-Book: Methods and Critical Appraisal for Evidence-Based Practice**, Elsevier Health Sciences, [*S. l.*], p. 310, 2014.

VISTA, N. P. B.; FIGUEIRÓ, M. F.; CHICON, P. M. M. Técnicas de mineração de dados aplicadas aos microdados do enade para avaliar o desempenho dos acadêmicos do curso de ciencia da computação no rio grande do sul utilizando o software r. **I Seminário de Pesquisa Científica e Tecnológica**, [*S. l.*], v. 1, n. 1, 2017.

WU, X.; ZHU, X.; WU, G.-Q.; DING, W. Data mining with big data. **IEEE transactions on knowledge and data engineering**, IEEE, [*S. l.*], v. 26, n. 1, p. 97–107, 2013.