DOI: 10.36558/rsc.v13i1.7995

APRENDIZADO DE MAQUINA PARA AGRUPAMENTO E ASSOCIACAO DE DADOS DO ENSINO SUPERIOR PUBLICO BRASILEIRO

MACHINE LEARNING FOR CLUSTERING AND ASSOCIATION OF DATA FROM BRAZILIAN PUBLIC HIGHER EDUCATION

Ebony Marques Rodrigues
Universidade Federal Rural de Pernambuco (UFRPE)
R. Dom Manuel de Medeiros, S/N, Dois Irmãos – 52.171-900
Recife – PE – Brasil
ebony.marquesr@ufrpe.br

Roberta Macêdo Marques Gouveia
Universidade Federal Rural de Pernambuco (UFRPE)
R. Dom Manuel de Medeiros, S/N, Dois Irmãos – 52.171-900
Recife – PE – Brasil
roberta.gouveia@ufrpe.br

Gabriel Alves de Albuquerque Junior
Universidade Federal Rural de Pernambuco (UFRPE)
R. Dom Manuel de Medeiros, S/N, Dois Irmãos – 52.171-900
Recife – PE – Brasil
gabriel.alves@ufrpe.br

Maria da Conceição Moraes Batista
Universidade Federal Rural de Pernambuco (UFRPE)
R. Dom Manuel de Medeiros, S/N, Dois Irmãos – 52.171-900
Recife – PE – Brasil
maria.cmbatista@ufrpe.br

RESUMO

Este trabalho visa à análise de características de Instituições de Ensino Superior (IES) públicas brasileiras e à descoberta de conhecimento sobre os contextos de formação de discentes de cursos de graduação das IES observadas. Para isso, técnicas previstas pelos métodos de *Knowledge Discovery in Databases* (KDD) e *Cross Industry Standard Process for Data Mining* (CRISP-DM) foram empregadas sobre bases de dados publicadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Experimentos de agrupamento e associação de dados, do Aprendizado de Máquina Não Supervisionado, foram executados em dois cenários de estudo. O primeiro cenário usa o algoritmo K-Means para agrupar IES públicas, com a observação de dados sobre despesas, quantidades de docentes e técnicos, localização e categoria administrativa das IES, entre outros. A análise dos quatro grupos obtidos no agrupamento possibilitou a identificação de similaridades e dissimilaridades entre as instituições. Os grupos, além de dados de concluintes de cursos de graduação nas IES (como faixa etária, tempo de graduação e forma de ingresso), são considerados no segundo cenário do estudo, que usa o algoritmo Apriori para geração de regras de associação que podem basear a caracterização dos perfis socioeconômicos dos estudantes.

Palavras-chave

ENADE; Censo da Educação Superior; KDD; CRISP-DM; Aprendizado de Máquina.

ABSTRACT

This work aims to analyze the characteristics of Brazilian public Higher Education Institutions (IES) and to discover knowledge about students in undergraduate courses at the observed IES. For this, techniques provided by the methods of Knowledge Discovery in Databases (KDD) and Cross Industry Standard Process for Data Mining (CRISP-DM) were used on databases published by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (National Institute of Educational Studies and Research Anísio Teixeira). Unsupervised Machine Learning data clustering and association experiments were performed in two study scenarios. The first scenario uses the K-Means algorithm to cluster public IES, observing data about expenses, number of professors and technicians, location and administrative category of IES, among others. The analysis of the four clusters obtained in the experiment enabled the identification of similarities and dissimilarities between the institutions. The clusters, in addition to data from students of undergraduate courses at IES (such as age group, length of stay at graduation and form of admission), are considered in the second study scenario, which uses the Apriori algorithm to gener-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ate association rules that can base the characterization of the students' socioeconomic profiles.

Keywords

ENADE; Higher Education Census; KDD; CRISP-DM; Machine Learning.

1. INTRODUÇÃO

O Censo da Educação Superior é uma ferramenta de pesquisa sobre as Instituições de Ensino Superior (IES) brasileiras que subsidia o Ministério da Educação no cumprimento de suas atribuições. Além disso, com a disponibilização de dados sobre infraestrutura, cursos, vagas ofertadas, docentes e discentes das IES, entre outros, o Censo possibilita a compreensão do sistema brasileiro de educação superior e contribui para o trabalho de gestores de instituições públicas e privadas, de pesquisadores e especialistas brasileiros e estrangeiros e de organismos internacionais [9].

O Exame Nacional de Desempenho dos Estudantes (ENADE) é um instrumento utilizado para a avaliação de competências e habilidades necessárias à formação geral de concluintes de cursos de graduação no Brasil. Realizado com a observação de um ciclo que abrange três anos como período de avaliação, o exame é aplicado, a cada ano, para estudantes de cursos de áreas do conhecimento determinadas [11]. Além da prova, o ENADE possui um questionário cujo objetivo é coletar informações que permitam designar os perfis socioeconômicos dos discentes, assim como os contextos de seus processos formativos [16].

O Índice Geral de Cursos (IGC) designa um indicador obtido a partir de uma média ponderada das notas dos cursos de graduação e pós-graduação de cada IES brasileira. Dessa forma, o IGC, divulgado anualmente, é capaz de sintetizar e estimar a qualidade dos cursos das instituições [12].

Este estudo tem como objetivos aplicar métodos do Aprendizado de Máquina Não Supervisionado para (1) identificar similaridades e dissimilaridades entre características de IES públicas federais e estaduais brasileiras, considerando dados sobre infraestrutura e investimento, por meio de um agrupamento, e (2) identificar perfis socioeconômicos de concluintes de cursos de graduação de graus bacharelado e licenciatura das IES observadas, contemplando os resultados do agrupamento realizado, a partir de regras de associação.

Visando ao atingimento dos objetivos, este trabalho compreende o emprego de técnicas de mineração de dados, com a observação dos processos de *Knowledge Discovery in Databases* (KDD) e *Cross Industry Standard Process for Data Mining* (CRISP-DM), sobre as bases de dados das edições de 2018 do Censo da Educação Superior, de 2019 do IGC e de 2016, 2017 e 2018 do ENADE [14, 13, 15]. As edições do Censo e do IGC citadas foram observadas por serem as mais recentes disponíveis durante a coleta dos dados, enquanto as três edições do ENADE em questão foram consideradas pois caracterizam o ciclo de avaliação completo mais recente concluído até o início da pesquisa.

A motivação deste estudo está fundamentada no interesse em adquirir o respaldo científico necessário para evidenciar a possibilidade de descobrir conhecimento, a partir dos dados coletados, por meio da detecção de padrões e da geração de regras significativas e não triviais. Espera-se que o conhecimento objetivado permita uma reflexão sobre determinadas características de IES públicas federais e estaduais de todo o Brasil, assim como a identificação e a compreensão de contextos de formação de estudantes dessas instituições, o que pode contribuir tanto para o aprimoramento de processos de ensino-aprendizagem quanto para a melhoria da gestão de recursos de IES públicas no país.

A Figura 1 apresenta uma visão geral sobre discentes concluintes de cursos de graduação de graus bacharelado e licenciatura em IES públicas federais e estaduais brasileiras que realizaram o ENADE entre 2016 e 2018. Operações de pré-processamento e transformação, que são abordadas nas seções seguintes, foram executadas sobre os dados. A breve análise tratada na imagem expõe, por região do Brasil, detalhes sobre a maior parte dos estudantes, considerando informações sobre sexo, cor/raça, faixa etária e ingresso na graduação, além da área do conhecimento do curso de graduação, contemplando a Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica do Brasil (Cine Brasil) de 2018 [10].

As próximas seções exibem informações pertinentes deste estudo, como trabalhos relacionados, o método e as ferramentas empregadas, os experimentos de aprendizado de máquina executados, os resultados obtidos nos experimentos, as considerações finais sobre a pesquisa e as referências.

2. TRABALHOS RELACIONADOS

Esta seção apresenta trabalhos relacionados ao estudo realizado. Estes trabalhos tratam de mineração de dados sobre bases educacionais com o uso de métodos de Aprendizado de Máquina Supervisionado e Não Supervisionado, visando à descoberta de conhecimento a partir da consideração de dados sobre a infraestrutura de instituições de ensino e da caracterização de perfis socioeconômicos de estudantes.

O estudo de caso de Mansilha, Sellitto, Lacerda & Serrano (2022) [17], que empregou dados de docentes do Bacharelado em Engenharia de Produção de uma IES privada do Rio Grande do Sul, compreendeu a execução de um agrupamento visando explorar o processo de análise de *clusters* como auxílio para tomada de decisão, contemplando os interesses de pesquisa dos docentes. Os autores efetuaram entrevistas e análises documentais e averiguaram os cadastros dos docentes na plataforma Lattes. Souza, Rover, Gallon & Ensslin (2009) [26] realizaram um agrupamento de IES brasileiras que ofertam cursos da área de Ciências Contábeis, usando a análise de *clusters*, a partir de dados sobre artigos publicados em eventos. O trabalho buscou identificar similaridades entre as instituições, observando características relacionadas à produção científica e ao número de pesquisadores.

A pesquisa descritiva de Torres-Samuel et al. (2019) [29] considerou as 85 IES latino-americanas que figuram entre as 50 primeiras posições de quatro *rankings* mundiais. Os autores construíram grupos de instituições a partir de atributos como frequência de presença e posição das universidades nos *rankings*, bem como o seu país. Stoupas, Sidiropoulos, Katsaros & Manolopoulos (2021) [27] utilizaram em seu estudo dados de universidades de Taiwan visando à execução de agrupamentos das instituições para ranqueá-las de acordo com suas características.

O trabalho de Dionisio, Rego, Ramos, Baltazar & Lucas (2015) [4] teve o objetivo de agrupar IES públicas portuguesas, observando atributos como tipo e forma de organização, por meio de métodos de análise estatística multivariada. A pesquisa tratou de identificar similaridades e dissimilarida-

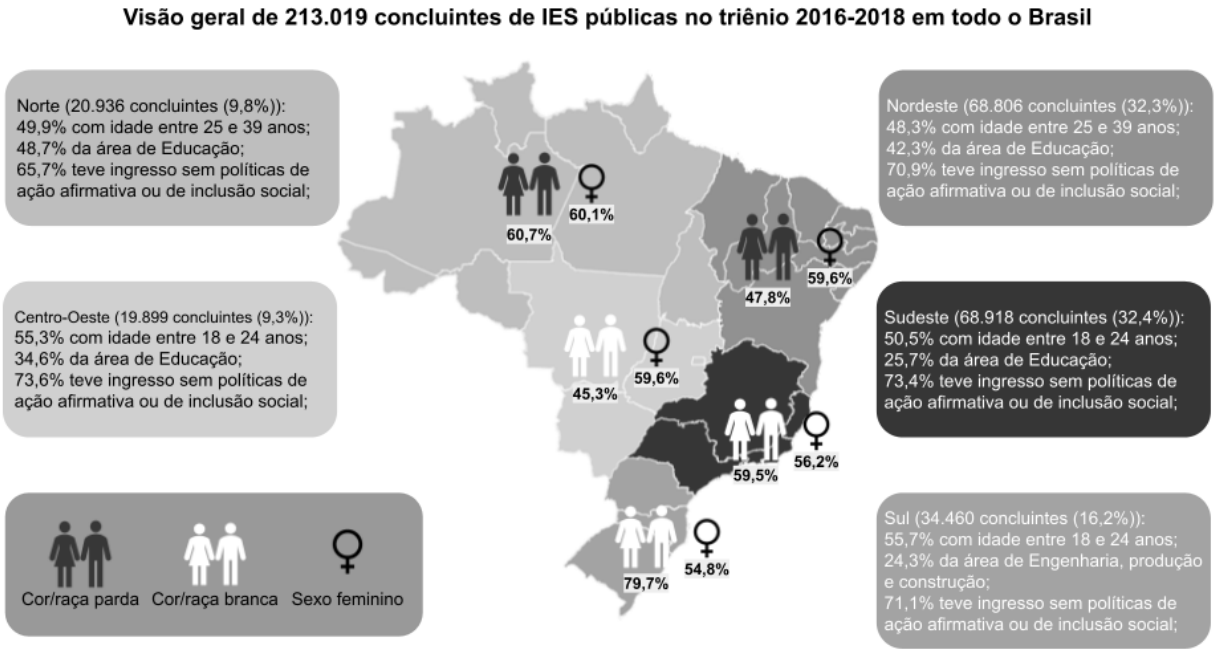


Figura 1: Visão geral dos concluintes de IES públicas federais e estaduais de todo o Brasil que participaram do ENADE entre 2016 e 2018. Fonte: os autores (2022).

des entre as instituições abordadas. Oliveira & Brito (2019) [19] empregaram em seu trabalho dados de 286 IES públicas brasileiras provenientes do Censo da Educação Superior de 2017 com o objetivo de realizar agrupamentos, observando o Coeficiente da Silhueta e o Índice de Calinski-Harabasz, para analisar as características das instituições. Os algoritmos K-Means e K-Medoids foram utilizados.

Os estudos supracitados empregaram dados de instituições de ensino superior diversas, considerando aspectos variados, com vistas à realização de ao menos um agrupamento, bem como o presente trabalho. Tanto Mansilha et al. (2022) [17] quanto Souza et al. (2009) [26] trataram de interesses de pesquisa de docentes. Torres-Samuel et al. (2019) [29] e Stoupas et al. (2021) [27] abordaram o agrupamento de IES visando ao ranqueamento de instituições. Por sua vez, Dionisio et al. (2015) [4] e Oliveira & Brito (2019) [19] realizaram agrupamentos de instituições de ensino superior de forma a permitir a identificação de similaridades e dissimilaridades entre as suas características.

Em especial, o trabalho de Oliveira & Brito (2019) [19] utilizou dados de IES oriundos do Censo da Educação Superior para a realização de um agrupamento com o uso do algoritmo K-Means, bem como o trabalho apresentado neste artigo. Entretanto, os métodos de processamento de variáveis considerados são distintos. Enquanto aqueles autores empregaram no agrupamento variáveis com categorias de formato semelhante ao original, nós realizamos, antes da execução do agrupamento propriamente dito, a *discretização* de determinadas variáveis, de forma a alterar as suas categorias, seguida de sua *binarização*, que permitiu o uso do K-Means. Nosso método é tratado em detalhes na seção Método e Ferramentas.

O trabalho de Rojanavasu (2019) [21] aplicou técnicas de associação e classificação, por meio dos algoritmos Apriori e Árvore de Decisão ID3, sobre dados de estudantes de uma universidade pública da Tailândia, buscando descobrir conhecimentos sobre o seu ingresso no mercado de trabalho. O estudo de Teodoro & Kappel (2020) [28] empregou dados do Censo da Educação Superior buscando identificar, por meio de técnicas de classificação, padrões característicos de estudantes do ensino superior público brasileiro que tendem a abandonar o curso de graduação.

A pesquisa de Cechinel, Araujo & Detoni (2015) [2] abordou a predição da reprovação de estudantes em cursos de graduação a distância utilizando dados da Universidade Federal de Pelotas e considerou métodos de classificação, com a comparação de vários algoritmos. Orji & Vassileva (2020) [20] utilizaram dados de estudantes de graduação de uma universidade pública do Canadá com o objetivo de analisar a relação entre engajamento e desempenho acadêmico. Em seu trabalho, métodos de agrupamento e classificação foram implementados por meio de algoritmos de Maximização de Expectativa e Floresta Aleatória.

V. Silva, Moreno, Gonçalves, Soares & Júnior (2020) [25] utilizaram dados do ENEM de 2019 para identificar desigualdades sociais de estudantes a partir de seu desempenho, empregando técnicas de agrupamento e associação. Os experimentos foram executados por meio dos algoritmos K-Means e Apriori. O estudo de A. Silva, Hoed & Saraiva (2019) [24] tratou da geração de regras de associação, por meio do algoritmo Apriori, a partir de dados do ENADE de 2017. Os autores tinham os objetivos de identificar e analisar fatores capazes de influenciar o desempenho de estudantes de cursos de Computação. O trabalho de Choji, Damasceno,

Bittencourt & Isotani (2021) [3] observou dados do ENADE de 2016, 2017 e 2018 visando analisar, após a geração de regras de associação com o algoritmo FP-Growth, dados de discentes do município de Araçatuba/SP. A pesquisa considerou perfis socioeconômicos de concluintes de cursos de graduação em instituições públicas e privadas.

Os trabalhos citados nos três últimos parágrafos trataram de caracterizar perfis socioeconômicos de estudantes do ensino básico ou superior, com o uso de técnicas de classificação, agrupamento ou associação. O estudo de Rojanavasu (2019) [21] observou o ingresso de discentes no mercado de trabalho, enquanto a pesquisa de Teodoro & Kappel (2020) [28] abordou a evasão estudantil. Os trabalhos de Cechinel et al. (2015) [2], Orji & Vassileva (2020) [20], V. Silva et al. (2020) [25] e A. Silva et al. (2019) [24] contemplaram a análise do desempenho acadêmico em contextos variados. Por sua vez, o trabalho de Choji et al. (2021) [3] abrangeu a interpretação de informações de discentes de um município do Estado de São Paulo. O nosso trabalho abrange a identificação de perfis socioeconômicos de discentes do ensino superior público no Brasil por meio da geração de regras de associação visando à identificação dos contextos de formação dos estudantes.

Em especial, os estudos de A. Silva et al (2019) [24] e Choji et al. (2021) [3] utilizaram dados de estudantes do ensino superior brasileiro, oriundos de bases do ENADE, para a geração e interpretação de regras de associação, observando os perfis socioeconômicos dos discentes, assim como o presente trabalho. Porém, os métodos de execução dos experimentos de associação considerados por aqueles autores são distintos dos que usamos. Enquanto eles selecionaram dados de estudantes por curso de graduação e localização, respectivamente, nós selecionamos dados por categoria administrativa da IES, grau acadêmico e duração do curso de graduação. O método utilizado no presente estudo está exposto em detalhes na seção Método e Ferramentas.

3. MÉTODO E FERRAMENTAS

Fayyad, Piatetsky-Shapiro & Smyth (1996) [5] dizem que o conhecimento é o produto final de uma descoberta baseada em dados. Segundo os autores, o processo de KDD tem cinco etapas que, contendo tarefas de seleção, pré-processamento, transformação e mineração de dados e interpretação e avaliação de resultados, tratam da descoberta de conhecimento útil a partir de dados, o que designa um método não trivial de identificação de padrões inéditos e válidos. Após a seleção de dados, cada etapa do KDD depende da finalização da etapa anterior, que pode ser repetida quantas vezes for preciso até que os dados estejam adequados para uso em etapas subsequentes.

O CRISP-DM possui como objeto de estudo questões de negócio que permitem a geração de conhecimento com a observação de seis etapas, que compreendem atividades de entendimento do negócio, entendimento e preparação dos dados, modelagem, avaliação e implantação de modelos [23]. É válido ressaltar que o CRISP-DM apresenta tarefas semelhantes às previstas pelo KDD em algumas de suas etapas.

O método utilizado para a execução deste estudo é constituído por uma adaptação dos processos de KDD e CRISP-DM, abrangendo as etapas que tratam de seleção, pré-processamento, transformação e mineração de dados. As etapas em questão são tratadas a seguir.

As linguagens de programação Python, de versão 3.8.3, e R, de versão 4.1.0, foram usadas, bem como as bibliotecas de Python Numpy, Pandas, Pandas Profiling, Feature Engine, Scikit-Learn, Imbalanced-Learn, XGBoost, Yellowbrick, Seaborn e Matplotlib e a biblioteca Arules do R, nas versões mais recentes. O RapidMiner Studio, de versão 9.8, ativado com licença educacional, também foi utilizado.

3.1 Seleção de Dados

A primeira etapa deste trabalho tratou de selecionar, das bases de dados coletadas, subconjuntos de dados de interesse para basear os experimentos de mineração de dados [5].

Com o uso do software RapidMiner Studio, por meio de sua ferramenta de automodelagem, no âmbito da tarefa de selecionar os atributos mais relevantes de uma base de dados de entrada observando os seus graus de correlação e estabilidade, subconjuntos de atributos de interesse foram selecionados a partir das bases de IES do Censo da Educação Superior e do ENADE [18]. De maneira específica, 17 atributos da base de dados de IES do Censo foram selecionados para, após a execução de operações de pré-processamento e transformação de dados, basear o experimento de agrupamento, enquanto 30 atributos das bases do ENADE foram selecionados para basear os experimentos de associação. É válido salientar que, após a seleção, foi realizada a tarefa de extração de dados propriamente dita, para criação dos subconjuntos, por meio da linguagem de programação Python e da biblioteca de Python Pandas.

Entre os atributos selecionados da base de IES do Censo, há dados como código de identificação no e-MEC, nome, categoria administrativa, organização acadêmica, localização, quantidade de técnicos, despesas da instituição etc. Entre os atributos selecionados das bases do ENADE, há informações sobre o ano de realização da prova, sobre o próprio estudante — como a faixa etária no ato de inscrição no exame, sexo, ano de conclusão do ensino médio, ano de início da graduação etc. —, sobre o curso de graduação — como a área do conhecimento, modalidade, turno, região etc. —, sobre a IES — a categoria administrativa — e sobre as respostas para o questionário socioeconômico — como o estado civil do estudante, sua cor/raça, renda familiar, motivo de escolha do curso de graduação, se o seu ingresso ocorreu por meio de políticas de ação afirmativa ou de inclusão social etc.

Neste estudo, foram utilizados dados de estudantes de cursos de graduação de graus bacharelado e licenciatura de IES de categoria administrativa federal ou estadual e de organização acadêmica universidade, faculdade ou centro universitário cujo tempo de graduação, definido neste trabalho como a diferença entre o ano de início da graduação e o ano de realização do ENADE, foi de 2 anos ou mais. Dados de estudantes de cursos de grau tecnólogo e dados de ingressantes no ensino superior não foram considerados. Dessa forma, foram selecionados para uso 226.612 registros de participantes do ENADE de 2016, 2017 e 2018, além de 108 registros da base de IES do Censo da Educação Superior de 2018, referentes às instituições vinculadas aos estudantes em questão.

Dados das bases de Docentes e de Cursos de Graduação do Censo da Educação Superior de 2018, assim como dados da base do IGC de 2019, foram selecionados para a criação de atributos posteriormente adicionados à base de dados de IES. A criação desses atributos é tratada na seção Transformação de Dados.

3.2 Pré-processamento de Dados

Esta etapa compreendeu a execução de processamentos gerais sobre atributos e registros após a identificação de inconsistências. Os dados inconsistentes foram corrigidos ou removidos de maneira a não comprometer a qualidade dos experimentos de aprendizado de máquina realizados [5].

Verificações de inconsistências foram executadas por meio da linguagem de programação Python e das bibliotecas de Python Pandas e Pandas Profiling sobre os dados selecionados, tal como determinadas correções. Operações de limpeza e padronização de dados também foram efetuadas, além da união de registros das bases das três edições do ENADE em uma única base.

Observando os dados selecionados da base de IES do Censo da Educação Superior, as tarefas de pré-processamento realizadas designam verificações de inconsistências, em especial, sobre os dados financeiros das instituições, além de alterações da estrutura dos atributos, abrangendo, por exemplo, a tarefa de renomeá-los para tornar os seus nomes intuitivos.

Quanto à busca por inconsistências nos dados do ENADE, as análises trataram de verificar se os registros possuem valores para o atributo que armazena o ano de realização do exame iguais ou menores do que o ano de conclusão do ensino médio e do que o ano de início da graduação do estudante. Outras análises verificaram se o ano de conclusão do ensino médio e se o ano de início da graduação do estudante caracterizam ano posterior a 2018 e se a idade do estudante é igual ou menor do que a diferença entre o ano de realização do exame e o ano de conclusão do ensino médio. Os registros que se enquadraram em alguma das situações citadas foram descartados.

Ainda tratando dos dados oriundos das bases do ENADE, considerando operações de padronização, foram executadas tarefas que visaram uniformizar a estrutura de atributos que apresentam divergências nos conjuntos das três edições, como ocorre com aqueles que armazenam o turno do curso de graduação do estudante. Alterações gerais também foram realizadas sobre todos os atributos, como o ato de renomeá-los. Por fim, todos os registros do ENADE que têm valores ausentes – registros que não possuem valores para determinados atributos – foram descartados.

Diante do exposto, uma ampla variedade de inconsistências foi verificada e tratada, tornando a base de dados processada e unificada com as três edições do ENADE íntegra, contendo 213.019 registros. Considerando o conjunto de dados selecionado inicialmente, a base de dados processada tem registros de 94% dos concluintes de cursos de graduação de graus bacharelado e licenciatura em IES públicas federais e estaduais, de organização acadêmica universidade, faculdade ou centro universitário, com tempo de graduação igual ou maior do que 2 anos e presenças válidas no ENADE entre os anos de 2016 e 2018 em todo o Brasil. 6% (13.593) dos registros do conjunto inicial foram considerados inconsistentes e descartados. Os estudantes cujos dados foram observados para sequência do estudo estão vinculados a 108 instituições públicas federais e estaduais que constituem a base de dados processada de IES.

3.3 Transformação de Dados

Após a etapa de pré-processamento, atributos foram criados ou transformados a partir de dados existentes para assumir determinadas estruturas, designando, por exemplo, alterações de categorias de valores possíveis. Tarefas de *discretização* foram executadas sobre os dados, tal como métodos de *binarização* e *undersampling*, com o emprego da linguagem de programação Python e das bibliotecas de Python Pandas, Pandas Profiling, Feature Engine, Scikit-Learn, Imbalanced-Learn, Seaborn e MatplotLib.

Inicialmente, no âmbito das IES, três atributos foram criados visando à execução do agrupamento de instituições. O primeiro atributo armazena a quantidade de docentes da IES, utilizando dados oriundos da base de Docentes do Censo da Educação Superior. O segundo atributo possui a despesa total da instituição, criado por meio da soma dos valores dos sete atributos de despesas existentes na base de IES do Censo. O terceiro atributo adicionado à base de IES foi criado a partir do IGC da instituição, obtido na base homônima, para armazenar o índice. Um quarto atributo também foi criado para integrar a base de IES, mas somente após a execução do agrupamento almejado, considerando os resultados do experimento. Esse atributo será tratado na seção Experimentos de Aprendizado de Máquina.

Cinco atributos foram criados e adicionados à base do ENADE para fundamentar os experimentos de associação. Alguns dos atributos foram criados por meio de dados oriundos de outras bases — como o que tem o grau acadêmico do curso de graduação do estudante, criado a partir de dados constantes na base de Cursos de Graduação do Censo da Educação Superior —, enquanto outros foram criados com a observação de operações de transformação sobre dados constantes na própria base do ENADE — como o atributo que possui a grande área do conhecimento do curso, criado por meio da área de formação geral e do manual para classificação de cursos Cine Brasil de 2018. Outro atributo criado armazena a diferença entre o ano de realização do ENADE e o ano de início da graduação pelo estudante, denominada, neste estudo, tempo de graduação.

Atributos do questionário socioeconômico do ENADE têm categorias que representam uma forma de expressar o quanto o participante concorda com a afirmação contida na questão em consideração. Por exemplo, o atributo INFRAESTRUTURA_GERAL_IES, relacionado à questão nº 61 do questionário, contém a afirmação "As condições de infraestrutura das salas de aula foram adequadas" e originalmente possui oito categorias: "*discordo totalmente*", "*discordo*", "*discordo parcialmente*", "*concordo parcialmente*", "*concordo*", "*concordo totalmente*", "*não se aplica*" e "*não sei responder*". Após a operação de transformação que trata da redução da cardinalidade de categorias, o atributo passou a ter cinco: "*discordo*", "*discordo/concordo parcialmente*", "*concordo*", "*não se aplica*" e "*não sei responder*".

Atributos numéricos, que possuem valores contínuos como categorias, podem ser *discretizados* de maneira que as suas frequências ou quantidades de observações sejam equilibradas. A *discretização* trata de apresentar os dados de forma alternativa, por meio de intervalos ou faixas de valores, para que os atributos originalmente contínuos tornem-se discretos, ou seja, para que as categorias dos atributos passem a ter quantidades específicas de valores possíveis [7].

A *discretização* dos atributos de IES que têm a despesa total e as quantidades de docentes e de técnicos da instituição foi efetuada de forma automática por meio do método *EqualFrequencyDiscretiser* da biblioteca Feature Engine [6]. Tratando dos dados do ENADE, as tarefas de *discretização* dos atributos que armazenam a faixa etária e o tempo de graduação do estudante foram realizadas manualmente. No

primeiro caso, foram consideradas as faixas etárias usadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em seus estudos. No segundo caso, a *discretização* ocorreu após análises acerca do tempo de graduação previsto para cursos de graus bacharelado e licenciatura.

Há algoritmos de Aprendizado de Máquina que requerem o emprego de dados numéricos para a sua execução, como o K-Means, utilizado neste estudo. Considerando que a existência de dados não numéricos em uma base é comum, como aqueles resultantes de operações de *discretização*, é preciso converter os dados não numéricos em dados numéricos, o que pode ser feito de várias formas. Quando os dados têm o mesmo peso, deve-se utilizar um método de codificação que não altere essa característica. Variáveis *dummy* podem ser usadas, pois permitem representar de forma binária dados de atributos que, em seu formato original, possuem duas ou mais categorias [22]. O método *OneHotEncoder* da biblioteca de Python Scikit-Learn foi empregado para a implementação de variáveis *dummy*, objetivando a codificação dos dados oriundos da base de IES do Censo, que basearam o cenário de agrupamento deste estudo [22].

Note que a *discretização* de dados foi executada para representar os valores de determinadas variáveis de IES em intervalos e, em seguida, tais variáveis foram *binarizadas* para que pudessem ser utilizadas na realização do agrupamento de instituições por meio do K-Means.

Após as tarefas de transformação, que envolveram a criação de atributos a partir de atributos selecionados das bases de dados originais, assim como a alteração e a exclusão de alguns desses atributos originais depois de seu emprego para a criação de novos, a base de IES, usada para basear o agrupamento de instituições, passou a ter 13 atributos, enquanto a base do ENADE, cujos dados foram utilizados para a geração de regras de associação, passou a ter 21 atributos.

4. EXPERIMENTOS DE APRENDIZADO DE MÁQUINA

Considerando os objetivos deste estudo, por meio de métodos do Aprendizado de Máquina Não Supervisionado, um agrupamento foi realizado com o uso do algoritmo K-Means no âmbito do primeiro cenário de mineração de dados, enquanto experimentos de associação foram efetuados no segundo cenário com o algoritmo Apriori. O Quadro 1 apresenta uma visão geral dos cenários, tratando de seus objetivos, método e ferramentas. Os dois cenários são expostos a seguir.

4.1 Agrupamento de Dados de IES

O K-Means é um algoritmo de agrupamento efetivo e bastante difundido. Partindo de um valor de entrada k, o algoritmo define, aleatoriamente, k pontos como centros dos grupos a serem gerados. Todas as instâncias de dados são atribuídas ao centro do grupo mais próximo, observando a métrica ordinária da distância euclidiana, e o centroide (média das instâncias) de cada grupo é calculado. Os centroides são tidos como os novos valores de centro dos grupos e todo o processo é repetido com os novos centros até que os mesmos pontos sejam atribuídos para cada grupo várias vezes consecutivas, o que sugere que os centros estabilizaram-se e permanecerão os mesmos para sempre. A inicialização *k-means++* aprimora tanto a velocidade quanto a acurácia do K-Means com uma escolha cuidadosa dos centros de grupo iniciais, por meio do que se chama de sementes (*seeds*) [30].

O Método do Cotovelo é um instrumento gráfico que auxilia na definição do número de grupos a serem gerados em um agrupamento a partir da observação da base de dados a ser usada. Se o gráfico retornado pelo método assemelha-se a um braço, o ponto de inflexão na curva, chamado de cotovelo, é uma boa indicação de que o agrupamento deve ter melhores resultados ao considerar tal valor como quantidade de grupos [31].

No primeiro cenário de mineração de dados, houve a execução de um agrupamento de IES públicas federais e estaduais brasileiras, a partir do algoritmo K-Means, com o uso da inicialização *kmeans++*, por meio da linguagem de programação Python e das bibliotecas de Python Pandas e Scikit-Learn, objetivando a geração de grupos de IES para a identificação de similaridades e dissimilaridades entre suas características. O Método do Cotovelo, implementado com o *KElbowVisualizer* da biblioteca de Python Yellowbrick, foi usado para definir a quantidade de grupos a serem gerados, com base na métrica da distorção.

O agrupamento utilizou 108 registros de IES públicas brasileiras e 13 atributos. Entre os atributos, há informações sobre despesas totais, quantidades de docentes e técnicos, localização, categoria administrativa e IGC da instituição. Com a análise de características das IES de cada grupo resultante do experimento, os grupos foram definidos. Dessa maneira, perceberam-se similaridades entre as IES de um mesmo grupo, assim como dissimilaridades entre as IES de grupos distintos e, por extensão, entre os próprios grupos. O dado que informa o grupo ao qual cada IES faz parte foi armazenado em um atributo criado e adicionado à base de dados de IES. Esse atributo compôs o conjunto de dados que baseou os experimentos de associação, tratados a seguir. Os resultados do agrupamento estão expostos em detalhes na seção Resultados.

4.2 Associação de Dados de Estudantes e IES

Regras de associação caracterizam um método de mineração de dados que relaciona as ocorrências de itens de uma base de dados de maneira que seja possível associar os dados para informar, a partir das frequências, o quanto a ocorrência de um conjunto de itens específico implica a ocorrência de outros conjuntos de itens da base. O algoritmo Apriori objetiva encontrar todos os conjuntos de itens possíveis em uma base de dados, recebendo valores mínimos para suporte e confiança como parâmetros [1].

Sejam X e Y dois conjuntos de itens (categorias) para um atributo de uma base. O suporte designa a porcentagem de transações (registros) da base que contêm X e Y, enquanto a confiança representa, considerando as transações que têm X, a porcentagem de transações que também têm Y. É ideal ter os valores de suporte e de confiança o mais próximos de 100% quanto for possível [1]. O *lift*, outro índice estatístico muito utilizado, informa o quão mais frequente Y torna-se quando X ocorre na regra. O *lift* pode ser maior, igual ou menor do que 1. Quando o *lift* é maior do que 1, há uma correlação positiva entre X e Y, ou seja, X e Y são positivamente correlacionados, o que significa que a ocorrência de um conjunto implica a ocorrência de outro. Quando o *lift* é igual a 1, os itens de X e Y são independentes, ou seja, não há correlação entre eles. Quando o *lift* é menor do que 1, diz-se que a ocorrência de X é negativamente correlacionada com a ocorrência de Y. É ideal ter regras com o *lift* muito

Quadro 1: Visão geral dos cenários de mineração de dados. Fonte: os autores (2022).

Cenário	Base de dados	Objetivo e método	Ferramentas
Agrupamento	13 atributos e 108 registros de IES	Agrupamento de IES federais e estaduais de todo o Brasil para a percepção de similaridades e dissimilaridades entre as instituições em quatro grupos	Python, Pandas, Scikit-Learn, Yellowbrick e outras
Associação	21 atributos e 213.019 registros do ENADE + 1 atributo de IES (que identifica o grupo ao qual a instituição está vinculada)	Geração de regras de associação para a percepção de perfis socioeconômicos de discentes de IES públicas federais e estaduais de todo o Brasil por meio de 92 experimentos	R, Arules e outras

maior do que 1 [8].

O segundo cenário do estudo tratou da execução de vários experimentos de associação empregando dados de estudantes de IES públicas federais e estaduais brasileiras e observando os resultados obtidos no agrupamento do primeiro cenário, com a utilização do algoritmo Apriori, por meio da linguagem de programação R e da biblioteca de R Arules.

Os experimentos de associação realizados foram baseados em 24 subconjuntos da base de 213.019 registros de discentes participantes do ENADE entre 2016 e 2018, de cursos de graduação de graus bacharelado e licenciatura de IES públicas federais e estaduais brasileiras, cujo tempo de graduação foi de 2 anos ou mais. 22 atributos foram usados.

Os 24 subconjuntos de dados, criados para separar as informações dos estudantes a partir do grupo ao qual as suas IES estão vinculadas, da forma de ingresso no ensino superior e do tempo de graduação, permitiram a busca pela geração de regras de associação com a observação da estrutura exibida na Figura 2, o que implica a modelagem de 96 experimentos de associação, uma vez que a estrutura em questão, que contempla 24 experimentos, foi considerada para cada grupo de IES.

Os experimentos objetivaram a obtenção de regras a partir de duas abordagens. A primeira abordagem considerou dados de estudantes que ingressaram no ensino superior público por meio de políticas de ação afirmativa ou de inclusão social, enquanto a segunda observou dados de estudantes que ingressaram sem essas políticas. Em cada abordagem, foram buscadas regras sobre estudantes com tempo de graduação de 2 a 4 anos, de 5 a 7 anos ou de 8 anos ou mais e com faixa etária de 18 a 24 anos, 25 a 39 anos ou de 40 anos ou mais. Porém, tendo em vista a baixa quantidade ou a ausência de dados de estudantes de 18 a 24 anos com tempo de graduação de 8 anos ou mais em alguns subconjuntos, os quatro experimentos vinculados à busca por regras desses estudantes foram desconsiderados. Diante disso, foram executados 23 experimentos de associação por grupo de IES, sendo, para os quatro grupos, realizados 92 experimentos.

Os experimentos buscaram regras com a observação de suporte mínimo de 30%, confiança mínima de 70% e *lift* maior do que 1. Os experimentos com essa configuração foram suficientes para a geração de regras com atributos desejados sobre a maioria dos estudantes, de faixa etária de 18 a 24 anos e de 25 a 39 anos, com tempo de graduação entre 2 e 4 anos e 5 e 7 anos. Os atributos desejados tratam de renda familiar, situação financeira e de trabalho, companhia de residência, estado civil e tipo de escola de ensino médio do estudante. Entretanto, tendo em vista perspectivas de associação mais específicas — como as que buscam regras de estudantes de todas as faixas etárias cujo tempo de graduação foi de 8 anos ou mais ou a que busca regras de estudantes de faixa etária de 40 anos ou mais cujo tempo de graduação foi de 2 a 4 anos —, foi necessário executar experimentos com a consideração de suportes menores para a geração de regras com os atributos buscados. Diante disso, além de 92 experimentos com suporte mínimo de 30%, também foram executados experimentos com suporte mínimo de 10%, 5%, 3% e 2% em casos específicos, com a confiança mínima de 70% e a seleção de regras com *lift* maior do que 1.

Caso a busca por regras ocorresse a partir da base de dados completa de 213.019 registros de participantes do ENADE, observando a execução de perspectivas de associação específicas, com quantidades de registros de estudantes do contexto desejado (tempo de graduação e faixa etária) muito baixas, o suporte a ser definido deveria ser diretamente proporcional, e isso não garantiria a percepção de regras com os atributos desejados. Dessa forma, o método empregado para a geração de regras de associação neste estudo, com a elevada quantidade de experimentos, é justificado.

Com os experimentos, buscou-se compreender os perfis socioeconômicos de concluintes de cursos de graduação de graus bacharelado ou licenciatura em IES públicas brasileiras cujo tempo de graduação foi de 2 a 4 anos ou 5 a 7 anos, que tinham de 18 a 24 anos, 25 a 39 anos ou 40 anos ou mais, assim como de concluintes com tempo de graduação de 8 anos ou mais, que tinham entre 25 e 39 anos ou 40 anos ou mais, observando se o ingresso no ensino superior ocorreu por meio de políticas de ação afirmativa ou de inclusão social ou não. É válido salientar que os dados utilizados neste estudo tratam de discentes de cursos de graduação em IES públicas que tiveram presenças válidas no ENADE. Diante disso, a evasão estudantil não foi considerada neste trabalho. Os resultados da associação de dados estão expostos na seção Resultados.

5. RESULTADOS

Esta seção apresenta os resultados percebidos após a execução dos experimentos dos dois cenários de mineração de dados do estudo.

5.1 Agrupamento de Dados de IES

Como exibe a Figura 3, o Método do Cotovelo retornou o valor de $k = 4$ como quantidade ideal de grupos a serem gerados no agrupamento que usou dados de 108 IES públicas federais e estaduais brasileiras. Após o experimento, os quatro grupos foram definidos com a consideração de carac-

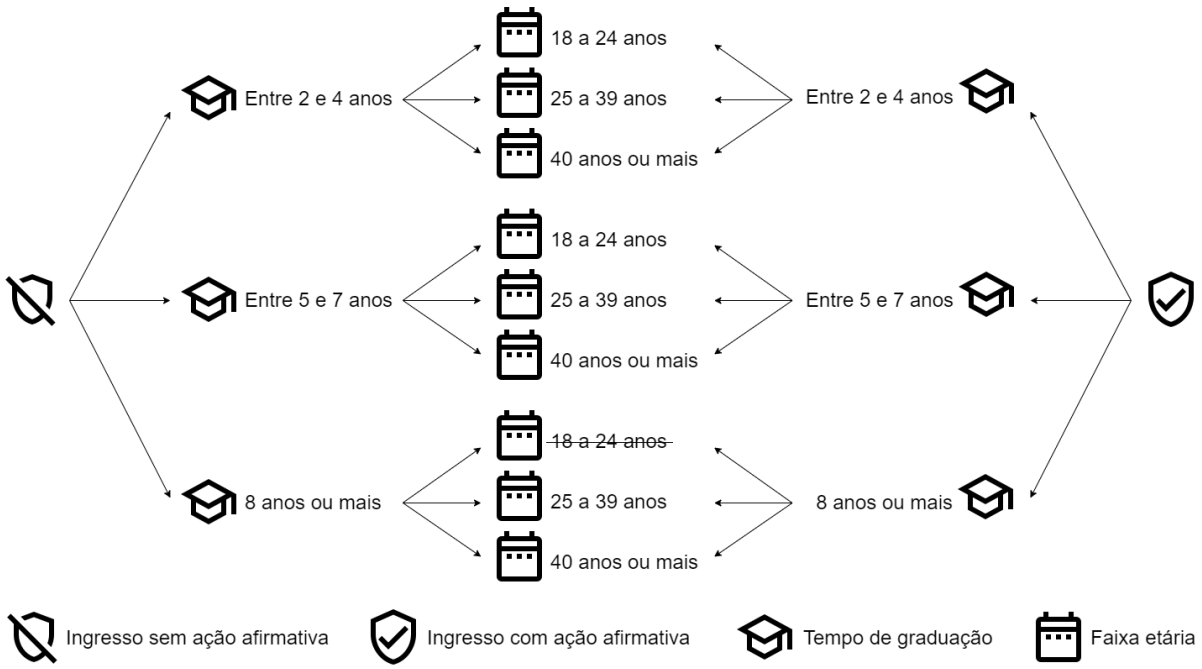


Figura 2: Visão geral da estrutura dos experimentos de associação. Fonte: os autores (2022).

terísticas da maior parte de suas instituições, a partir de informações sobre despesas totais, quantidades de docentes e de técnicos, localização e categoria administrativa. A Figura 4 apresenta uma visão geral sobre a maior parte das IES de cada grupo.

Nota-se que 20,3% das IES consideradas no agrupamento foram alocadas no Grupo 1. O Grupo 2, por sua vez, conta com 24,1% das IES, enquanto o Grupo 3 possui 28,7%. O Grupo 4 tem 26,9% das IES analisadas. Diante disso, é válido perceber que os grupos obtidos são balanceados por quantidade de instituições.

No Grupo 1, 95,45% das IES têm investimento baixo, com despesas entre R$ 2.035.717,73 e R$ 184.224.613,77, enquanto 57,69% das instituições do Grupo 2 possuem investimento médio, com despesas entre R$ 184.224.613,78 e R$ 311.016.578,32. No Grupo 3, 54,84% das instituições têm investimento alto, com despesas entre R$ 311.016.578,33 e R$ 794.566.153,67. No Grupo 4, 82,76% das IES possuem investimento muito alto, com despesas entre R$ 794.566.153,68 e R$ 4.016.243.944,85.

Quanto às quantidades de docentes e técnicos, a maior parte das IES do Grupo 1 tem entre 40 e 540 docentes e entre 29 e 371 técnicos. Por sua vez, a maior parte das IES do Grupo 2 tem entre 993 e 1.841 docentes e entre 372 e 771 técnicos, enquanto a maior parte das IES do Grupo 3 possui entre 541 e 992 docentes e entre 772 e 1.723 técnicos. Considerando as IES do Grupo 4, a maior parte delas tem entre 1.842 e 4.197 docentes e entre 1.724 e 14.581 técnicos.

Percebe-se que 72,73% das IES do Grupo 1 têm a categoria administrativa de IES estadual, tal como 57,69% das IES do Grupo 2. Por outro lado, 74,19% das IES do Grupo 3 têm a categoria de IES federal, bem como 86,21% das IES do Grupo 4. Quanto à localização, a maior parte das IES do Grupo 1 encontra-se na região Sudeste, bem como a maior parte das IES do Grupo 3. Na região Nordeste, está localizada a maior parte das IES do Grupo 2, assim como a maior parte das IES do Grupo 4. Quanto ao Índice Geral de Cursos, segundo dados de 2019, 40,9% das IES do Grupo 1 têm o conceito 3, assim como 53,85% das IES do Grupo 2 no Grupo 3, 77,42% das IES possuem o conceito 4, tal como 72,41% das IES do Grupo 4.

Sobre a alocação de IES por grupo, observando o Grupo 1, notam-se IES federais como a Universidade Federal de Ciências da Saúde de Porto Alegre e a Universidade Federal do Cariri, além de IES estaduais como a Universidade Estadual do Paraná e a Universidade Estadual de Roraima. No Grupo 2, estão IES federais como a Universidade Federal Rural do Semi-Árido e a Universidade Federal do Vale do São Francisco, bem como IES estaduais como a Universidade do Estado de Minas Gerais e a Universidade de Pernambuco. Por sua vez, o Grupo 3 possui IES federais como a Universidade Federal de Campina Grande e a Universidade Federal de Pelotas, tal como IES estaduais como a Universidade Estadual de Londrina e a Universidade Estadual da Paraíba. Por fim, no Grupo 4, existem IES federais como a Universidade Federal de Pernambuco, Universidade Federal Rural de Pernambuco, Universidade Federal do Paraná, Universidade de Brasília, Universidade Federal de Minas Gerais e Universidade Federal do Rio Grande do Sul, além de IES estaduais como a Universidade do Estado do Rio de Janeiro e a Universidade Estadual Paulista Júlio de Mesquita Filho.

A partir dos resultados obtidos, percebe-se que o Grupo 1, em sua maioria, é constituído de IES de categoria administrativa estadual, da região Sudeste, de investimento baixo, tendo baixas quantidades de docentes e técnicos em comparação com os demais grupos. No Grupo 2, a maior parte das IES também possui a categoria administrativa estadual, mas com localização na região Nordeste e investimento médio, tendo maiores quantidades de docentes e técnicos em relação às IES do Grupo 1. No Grupo 3, a maior parte das

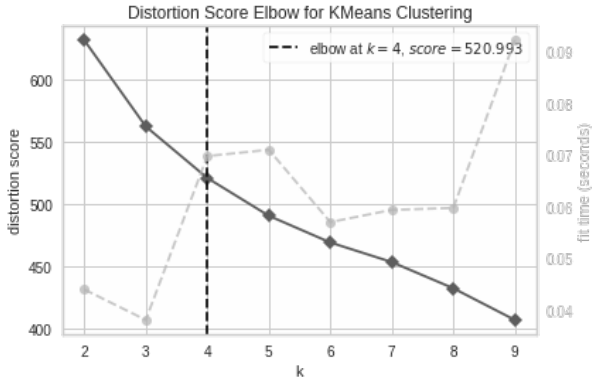


Figura 3: Método do Cotovelo para o K-Means com a inicialização $k\text{-}means++$ e a métrica da distorção. Fonte: os autores (2022).

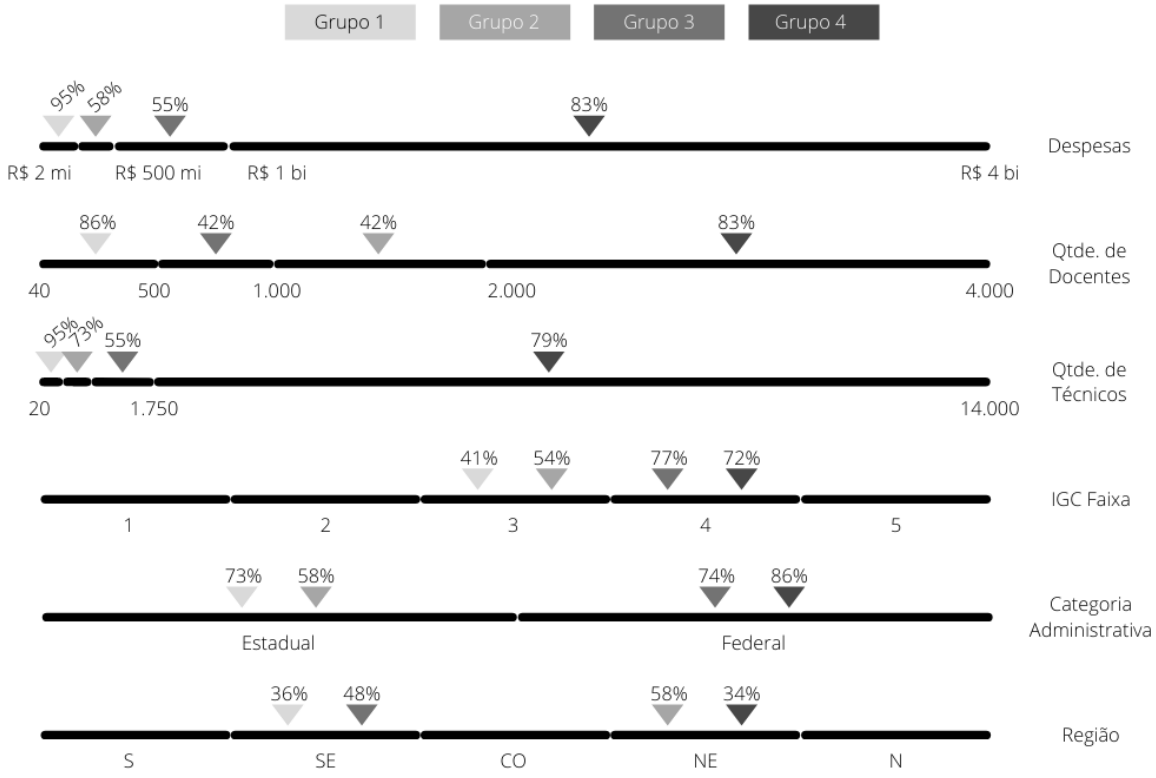


Figura 4: Visão geral dos grupos observando a maior parte de suas IES. Fonte: os autores (2022).

IES conta com investimento alto, categoria administrativa federal e altas quantidades de docentes e técnicos, com localização nas regiões Sudeste e Sul. A maior parte das IES do Grupo 4 tem investimento muito alto, as maiores quantidades de docentes e técnicos entre as IES analisadas e categoria administrativa federal, com localização nas regiões Nordeste e Sudeste.

5.2 Associação de Dados de Estudantes e IES

Com a execução dos experimentos de associação, as regras obtidas que apresentaram os maiores valores para suporte, confiança e *lift* foram selecionadas, interpretadas e empregadas na construção dos resumos visuais expostos a seguir.

A Figura 5 exibe a visão geral das regras de associação geradas usando dados de estudantes vinculados a IES que integram o Grupo 1, grupo em que a maior parte das IES tem investimento baixo. A Figura 6 trata das regras obtidas quando foram observados dados de estudantes de IES do Grupo 2, onde a maior parte das IES tem investimento médio. A Figura 7 expõe regras de estudantes relacionados a IES do Grupo 3, cuja maior parte das IES tem investimento alto. A Figura 8 resume as regras sobre estudantes associados a IES do Grupo 4, grupo em que a maior parte das IES tem investimento muito alto.

A partir das regras obtidas, observando os estudantes de 18 a 24 anos, percebe-se que aqueles que ingressaram no ensino superior POR MEIO de políticas de ação afirmativa ou de inclusão social têm perfis semelhantes independentemente do grupo ao qual as suas IES estão vinculadas: são solteiros, moram com pais e/ou parentes, cursaram todo ou a maior parte do ensino médio em escola pública e têm renda familiar de até 3 salários mínimos. Há exceção que trata de estudantes de IES do Grupo 4, que, a depender do tempo de graduação, podem ter renda familiar acima de 3 salários. Por sua vez, os estudantes de 18 a 24 anos que ingressaram no ensino superior SEM políticas de ação afirmativa são solteiros, moram com pais e/ou parentes, cursaram todo ou a maior parte do ensino médio em escola privada e têm renda familiar acima de 3 salários mínimos. Há exceções relacionadas a estudantes de IES dos grupos 1 e 2, que, a depender do tempo de graduação, podem ter estudado em escola pública e ter renda familiar de até 3 salários.

Observando os estudantes de 25 a 39 anos, aqueles de IES dos grupos 1 e 2, independentemente da forma de ingresso no ensino superior, cursaram todo ou a maior parte do ensino médio em escola pública e têm renda familiar de até 3 salários, com o estado civil e a companhia de residência variando de acordo com o tempo de graduação. Os estudantes de 25 a 39 anos de IES dos grupos 3 e 4 que ingressaram no ensino superior POR MEIO de políticas de ação afirmativa ou de inclusão social são solteiros, moram com pais e/ou parentes, cursaram todo ou a maior parte do ensino médio em escola pública e têm renda familiar de até 3 salários. Há exceção designada por estudantes de IES do Grupo 3, que, dependendo do tempo de graduação, podem ter renda familiar acima de 3 salários. Tratando dos estudantes de 25 a 39 anos que ingressaram no ensino superior SEM políticas de ação afirmativa ou de inclusão social, aqueles de IES do Grupo 3 têm perfis bastante distintos, com estado civil, companhia de residência, tipo de escola de ensino médio e renda familiar mudando a depender do tempo de graduação. O tipo de escola de ensino médio e a renda familiar dos estudantes de 25 a 39 anos de IES do Grupo 4 também variam de acordo com o tempo de graduação, mas, em geral, são solteiros e moram com pais e/ou parentes.

As regras mostram que os estudantes de 40 anos ou mais, independentemente do grupo ao qual as suas IES estão vinculadas, são casados, moram com cônjuge e/ou filhos e cursaram todo ou a maior parte do ensino médio em escola pública, com a renda familiar variando de acordo com a forma de ingresso no ensino superior e o tempo de graduação. Além disso, as regras mostram que os estudantes de 40 anos ou mais de IES do Grupo 4 que ingressaram no ensino superior SEM políticas de ação afirmativa ou de inclusão social têm renda familiar acima de 3 salários, independentemente do tempo de graduação.

Os resultados expostos indicam que, independentemente da faixa etária, estudantes que ingressam no ensino superior POR MEIO de políticas de ação afirmativa ou de inclusão social cursaram todo ou a maior parte do ensino médio em escola pública e/ou têm renda familiar de até 3 salários mínimos. Por outro lado, percebem-se regras que mostram que estudantes de 18 a 39 anos que ingressaram no ensino superior SEM políticas de ação afirmativa ou de inclusão social cursaram todo ou a maior parte do ensino médio em escola privada e têm renda familiar acima de 3 salários. Por fim, observando os estudantes de 40 anos ou mais, é válido perceber que, segundo as regras obtidas, mesmo aqueles que ingressaram SEM políticas de ação afirmativa ou de inclusão social cursaram todo ou a maior parte do ensino médio em escola pública.

6. CONSIDERAÇÕES FINAIS

Este trabalho trata do uso de métodos de agrupamento e associação sobre dados provenientes do Censo da Educação Superior e do ENADE visando à descoberta de conhecimento sobre o ensino superior público no Brasil. O primeiro cenário de mineração de dados do estudo compreendeu a execução de um agrupamento de IES públicas federais e estaduais brasileiras para a identificação de similaridades e dissimilaridades entre suas características. No segundo cenário, foram geradas regras de associação buscando identificar perfis socioeconômicos de estudantes de cursos de graduação de graus bacharelado e licenciatura das IES observadas no primeiro cenário, considerando os resultados do agrupamento.

Com o agrupamento de IES públicas federais e estaduais, houve a geração de quatro grupos, definidos com a análise da maior parte de suas instituições, observando informações sobre despesas, quantidades de docentes e técnicos, localização e categoria administrativa da IES, entre outras. Percebe-se que IES estaduais, que representam a maioria das instituições dos grupos 1 e 2, recebem menos investimento do que IES federais, que caracterizam a maioria das IES dos grupos 3 e 4. Nota-se, também, que as instituições de maior investimento estão localizadas, principalmente, nas regiões Nordeste, Sudeste e Sul do Brasil.

As regras de associação geradas permitiram uma percepção geral sobre perfis socioeconômicos de discentes de IES públicas brasileiras, observando os grupos aos quais as IES estão vinculadas, resultantes do agrupamento realizado no primeiro cenário. As regras indicam que estudantes de 18 a 39 anos que ingressam no ensino superior por meio de políticas de ação afirmativa ou de inclusão social estudaram em escola pública e/ou têm renda familiar menor no momento de conclusão da graduação quando comparados a estudantes de mesma faixa etária que ingressaram no ensino superior

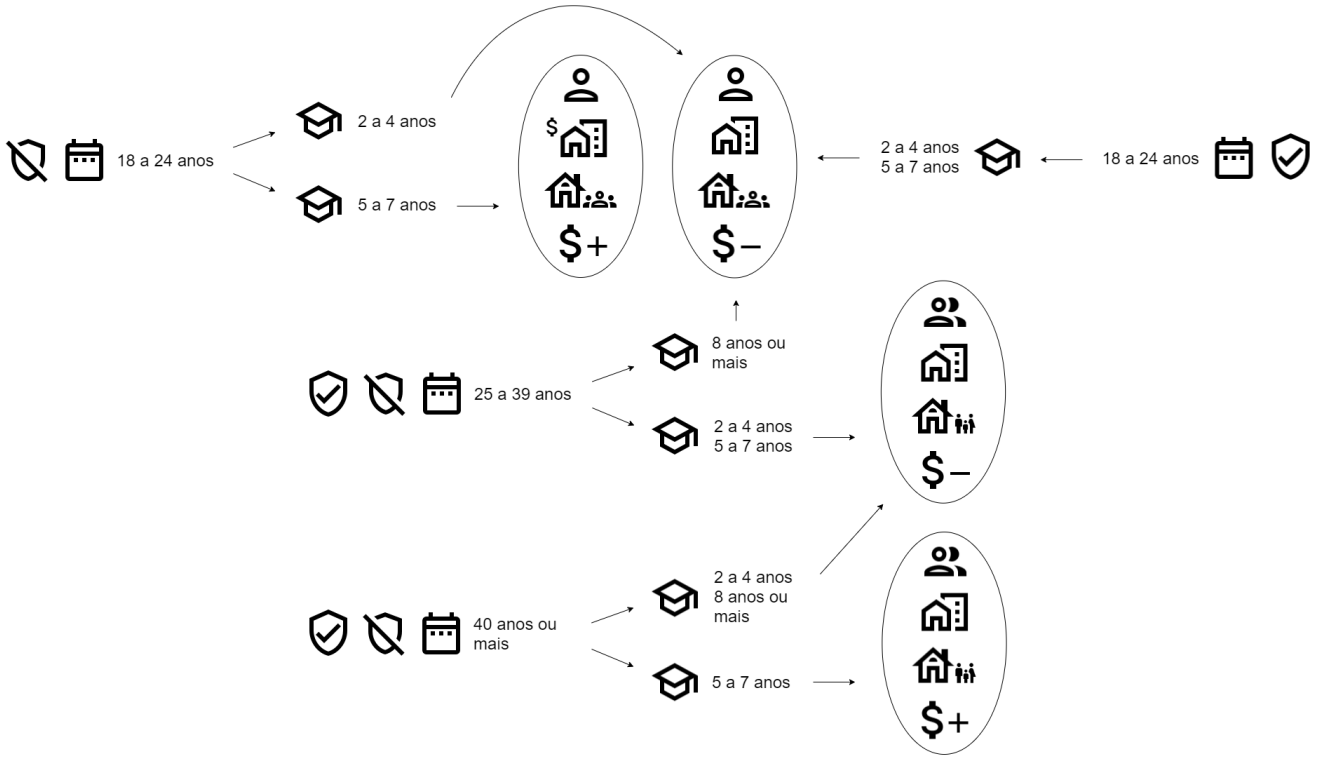
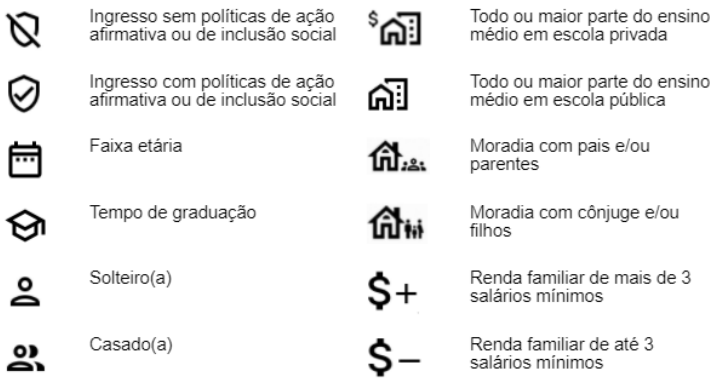


Figura 5: Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 1. Fonte: os autores (2022).



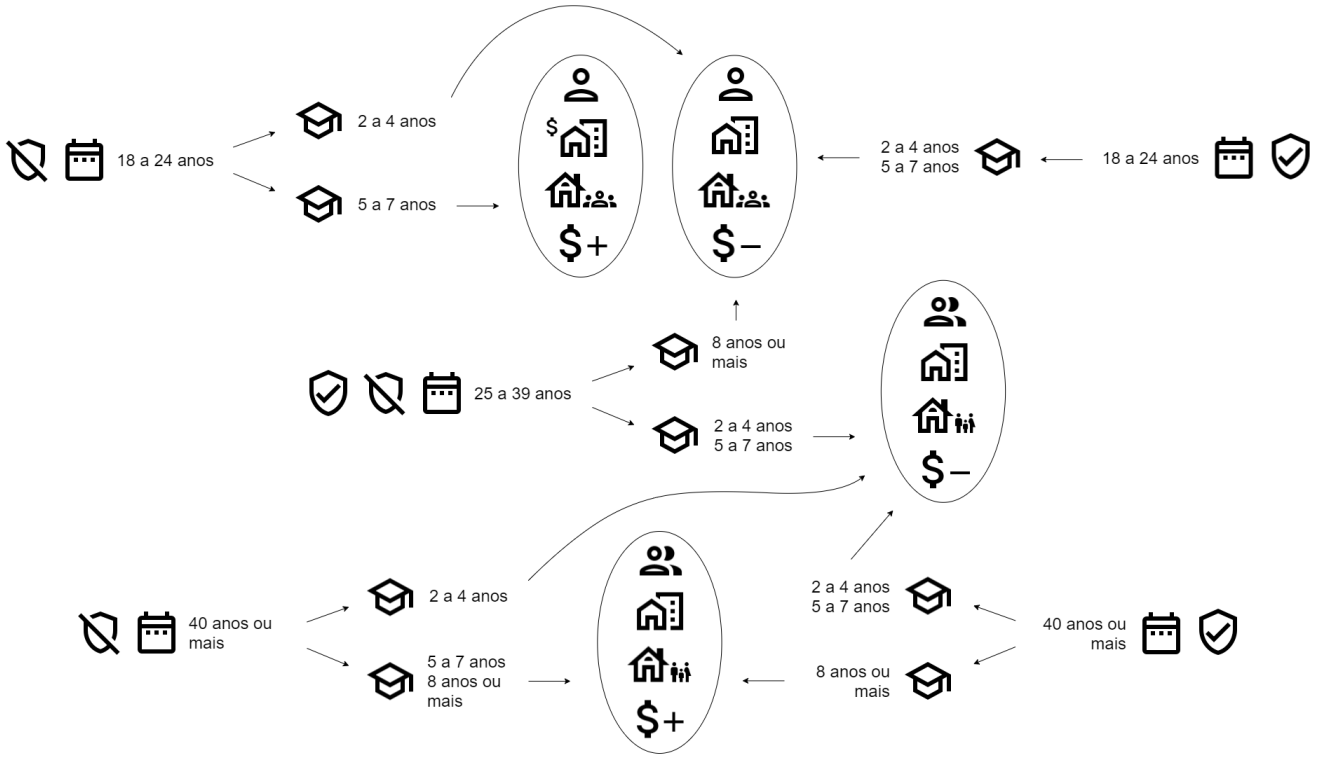


Figura 6: Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 2. Fonte: os autores (2022).

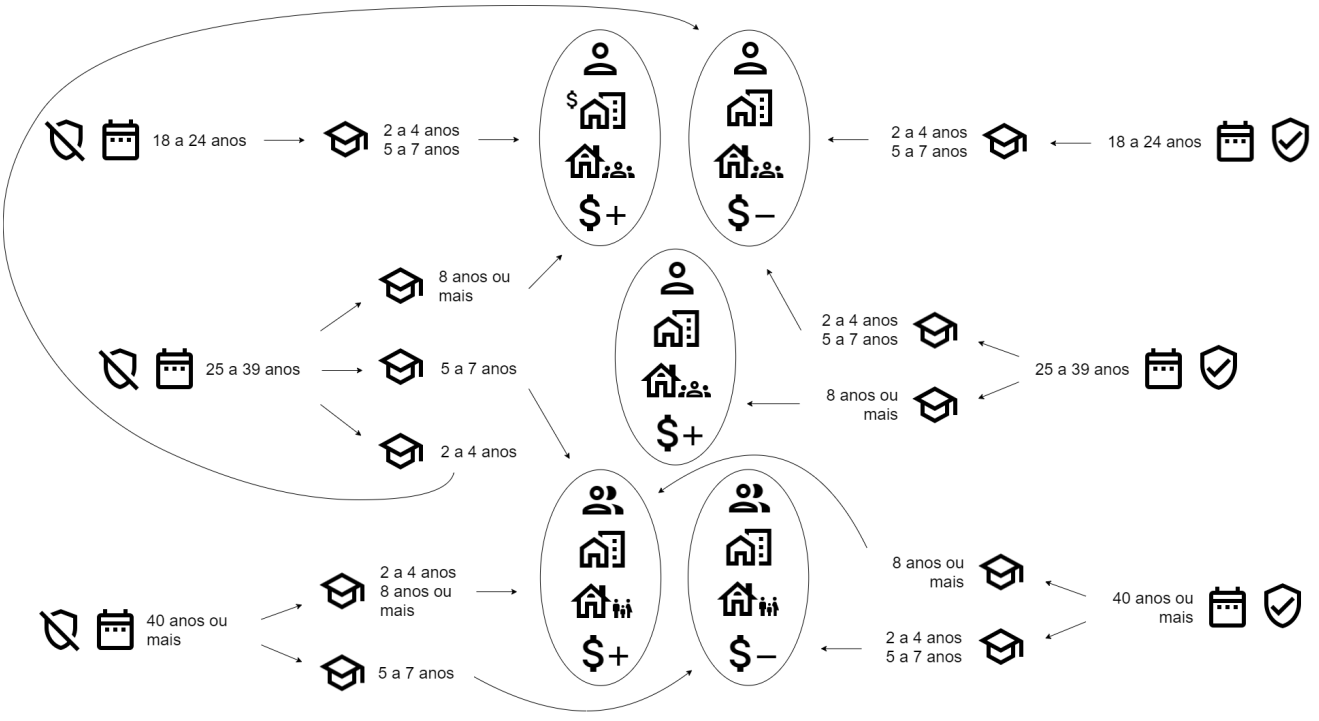


Figura 7: Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 3. Fonte: os autores (2022).

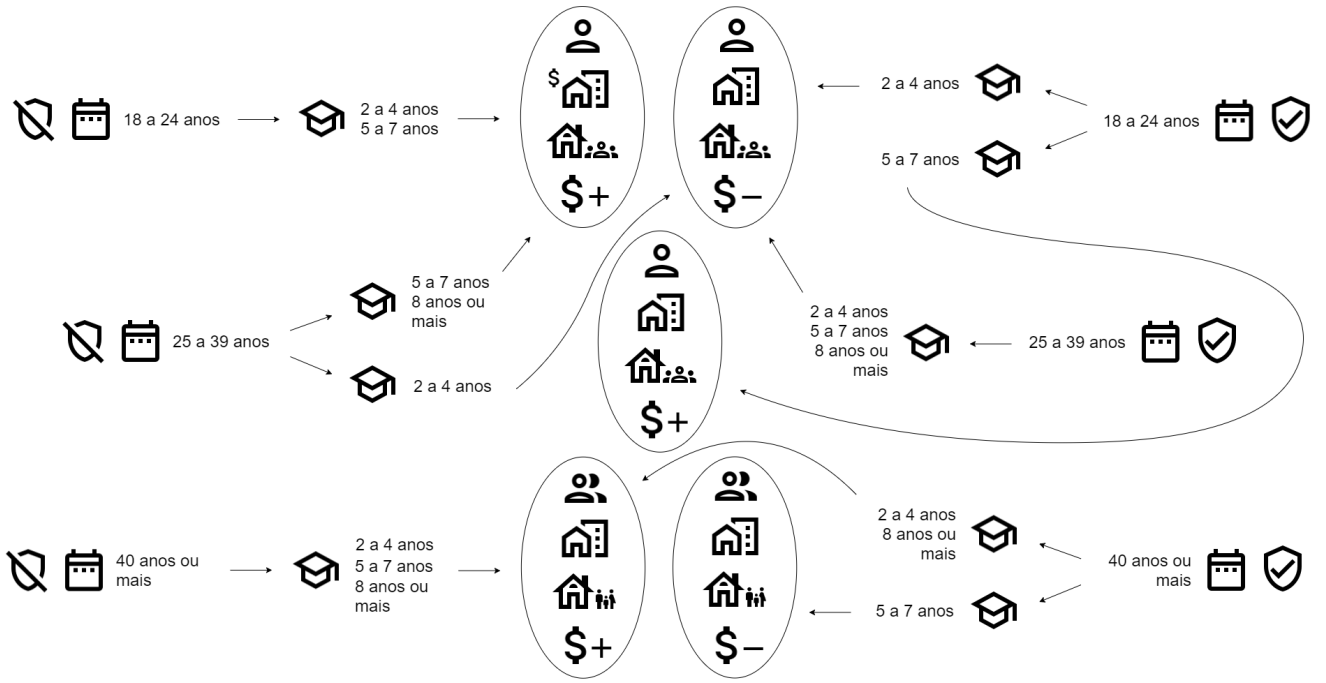


Figura 8: Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 4. Fonte: os autores (2022).

sem políticas de ação afirmativa ou de inclusão social, que estudaram em escola privada e têm renda familiar maior durante a conclusão do ensino superior.

Diante do exposto, é preciso considerar a importância e estimular o desenvolvimento de trabalhos com objetivos semelhantes aos deste, que tratem da compreensão do contexto de formação de discentes do ensino superior público no Brasil e que possam basear a elaboração de propostas de intervenção institucionais e de políticas públicas que, entre outros fins, busquem reduzir os índices de retenção e evasão em IES federais e estaduais brasileiras.

7. REFERÊNCIAS

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 487–499, 1994.

[2] C. Cechinel, R. Araujo, and D. Detoni. Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. *Revista Brasileira de Informática na Educação*, 23(03):1, 2015.

[3] M. F. Choji, C. D. N. Damasceno, I. I. Bittencourt, and S. Isotani. Mineração de dados do Enade de 2016 a 2018: uma análise sobre o município de Araçatuba/SP. *RENOTE*, 19(2):183–192, 2021.

[4] A. Dionisio, C. Rego, I. Ramos, M. Baltazar, and R. Lucas. Formas de organização e agrupamento das Instituições de Ensino Superior portuguesas. In *5ª Conferência FORGES – Autonomia e os Modelos de Governo e Gestão das Instituições de Ensino Superior*, 2015.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37, Mar. 1996.

[6] Feature-Engine. EqualFrequencyDiscretiser. Read the Docs. `https://feature-engine.readthedocs.io/en/latest/discretisation/EqualFrequencyDiscretiser.html`, 2021. Acesso em 10 de janeiro de 2021.

[7] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques, 2012.

[8] IBM. Lift in an association rule, 2021. Acesso em 10 de novembro de 2021.

[9] INEP. Censo da Educação Superior. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. `https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior`, 2021. Acesso em 12 de janeiro de 2021.

[10] INEP. Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica (Cine Brasil). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. `https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/cine-brasil`, 2021. Acesso em 30 de outubro de 2021.

[11] INEP. Exame Nacional de Desempenho dos Estudantes (ENADE). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. `http://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade`, 2021. Acesso em 12 de janeiro de 2021.

[12] INEP. IGC. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. http://portal.mec.gov.br/igc, 2021. Acesso em 16 de fevereiro de 2022.

[13] INEP. Microdados de Indicadores de Qualidade da Educação Superior. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-qualidade-da-educacao-superior, 2021. Acesso em 13 de agosto de 2020.

[14] INEP. Microdados do Censo da Educação Superior. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior, 2021. Acesso em 13 de agosto de 2020.

[15] INEP. Microdados do ENADE. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade, 2021. Acesso em 13 de agosto de 2020.

[16] INEP. Questionário do Estudante do ENADE. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade/questionario-do-estudante, 2021. Acesso em 13 de janeiro de 2021.

[17] R. B. Mansilha, M. A. Sellitto, D. P. Lacerda, and R. Serrano. Formação de clusters na docência por interesse de pesquisa: método de auxílio à tomada de decisões em cursos de nível superior. *Transinformação*, 30(3):15–38, junho 2022.

[18] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 935–940, New York, NY, USA, 2006. Association for Computing Machinery.

[19] R. B. d. Oliveira and J. A. d. M. Brito. Análise de Agrupamento Aplicada ao Estudo de Instituições de Ensino Superior Públicas. In *Revista do Seminário Internacional de Estatística com R*, Niterói, RJ, Brasil, 2019.

[20] F. Orji and J. Vassileva. Using Machine Learning to Explore the Relation Between Student Engagement and Student Performance. In *2020 24th International Conference Information Visualisation (IV)*, pages 480–485, 2020.

[21] P. Rojanavasu. Educational Data Analytics using Association Rule Mining and Classification. In *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, pages 142–145, 2019.

[22] Scikit-Learn. Sklearn.preprocessing.OneHotEncoder. Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html, 2021. Acesso em 10 de fevereiro de 2021.

[23] C. Shearer. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 2000.

[24] A. Silva, R. Hoed, and P. Saraiva. Análise do Desempenho dos Alunos de Cursos Superiores em Computação no ENADE - Uma Abordagem usando Mineração de Dados. In *Atas da conferência Ibero-Americana*, pages 207–214, 12 2019.

[25] V. Silva, L. Moreno, L. Gonçalves, S. Soares, and R. S. Júnior. Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 72–81, Porto Alegre, RS, Brasil, 2020. SBC.

[26] F. Souza, S. Rover, A. Gallon, and S. Ensslin. Análise das IES da Área de Ciências Contábeis e de seus Pesquisadores por meio de sua Produção Científica. *Contabilidade Vista amp; Revista*, 19(3):15–38, maio 2009.

[27] G. Stoupas, A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Ranking universities via clustering. In *Proceedings of the 18th International Conference on Scientometrics & Informetrics (ISSI), Leuven, Belgium*, 2021.

[28] L. Teodoro and A. M. Kappel. Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. *Revista Brasileira de Informática na Educação*, 28(0):838–863, 2020.

[29] M. Torres-Samuel, C. L. Vásquez, M. L. Cardozo, N. Bucci, A. Viloria, and D. Cabrera. Clustering of Top 50 Latin American Universities in SIR, QS, ARWU, and Webometrics Rankings. *Procedia Computer Science*, 160:467–472, 2019. The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2019) / The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2019) / Affiliated Workshops.

[30] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3 edition, 2011.

[31] Yellowbrick. Elbow Method. Yellowbrick. https://www.scikit-yb.org/en/latest/api/cluster/elbow.html, 2021. Acesso em 7 de novembro de 2021.