UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Ebony Marques Rodrigues

Técnicas de Aprendizado de Máquina para Descoberta de Conhecimento sobre Dados Abertos do Ensino Superior Público Brasileiro

Recife

2021

Ebony Marques Rodrigues

Técnicas de Aprendizado de Máquina para Descoberta de Conhecimento sobre Dados Abertos do Ensino Superior Público Brasileiro

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Orientadora: Roberta Macêdo Marques Gouveia

Recife

2021

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

R696t Rodrigues, Ebony Marques
Técnicas de aprendizado de máquina para descoberta de conhecimento sobre dados abertos do ensino superior público brasileiro / Ebony Marques Rodrigues. - 2021.
60 f. : il.

Orientadora: Roberta Macedo Marques Gouveia.
Inclui referências e apêndice(s).

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

1. Enade. 2. Censo da educação superior. 3. Kdd. 4. Crisp-dm. 5. Aprendizado de máquina. I. Gouveia, Roberta Macedo Marques, orient. II. Título

CDD 004

Ebony Marques Rodrigues

Técnicas de Aprendizado de Máquina para Descoberta de Conhecimento sobre Dados Abertos do Ensino Superior Público Brasileiro

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em 10 de dezembro de 2021.

BANCA EXAMINADORA

Roberta Macêdo Marques Gouveia (orientadora)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Gabriel Alves de Albuquerque Júnior
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Rodrigo Lins Rodrigues
Departamento de Educação
Universidade Federal Rural de Pernambuco

Aos meus pais, à minha avó e ao meu irmão, que me apoiam desde sempre.

Agradecimentos

Em primeiro lugar, eu agradeço a Deus por ter o pai e a mãe que eu tenho. Eles são as pessoas mais importantes da minha vida. Neste momento, prestes a concluir a graduação, percebo que não houve um momento em que eles deixaram de me apoiar. Desde a escolha do curso até a defesa do TCC, todas as palavras que os meus pais me destinam são de incentivo, e eu sei que daqui para frente não será diferente. Não existem palavras capazes de expressar todo o amor que eu sinto por vocês, Ricardo e Vania, painho e mainha. Eu sou extremamente grato por tudo o que fizeram, fazem e continuarão fazendo por mim, se Deus quiser. Vocês são extremamente importantes para mim. Eu amo vocês.

É óbvio que minha avó, Josefa, e meu irmão menor, Kelvin, também devem ser mencionados neste texto. À minha avó, de 88 anos, que passou a viver comigo, com meus pais e meu irmão há muito tempo, eu agradeço por todo o amor que ela emana diariamente, não só para mim, mas para todos os que são felizes por tê-la por perto. Sou grato por todas as palavras de incentivo que ela me diz e que continuará dizendo, se Deus quiser. Ao meu irmão, que concluirá o ensino médio em alguns anos, eu digo que realmente vale a pena cursar uma graduação relacionada à computação, apesar do pouco de estresse que a área pode causar. Brincadeiras à parte, eu sei que o seu futuro será brilhante, se Deus quiser. Vocês também são extremamente importantes para mim. Eu também amo vocês.

Eu sou extremamente grato à professora Roberta, do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, que é minha orientadora desde a época da iniciação científica. São quase dois anos de muito trabalho em que ela me orienta, ensina e auxilia com muita atenção e dedicação. Neste percurso, milhares de mensagens de texto e áudio foram trocadas, bem como dezenas de horas de reunião foram compartilhadas, de domingo a domingo, a qualquer momento do dia ou da noite, do início da pesquisa, passando pela publicação de um artigo, até a conclusão deste trabalho. Sou muito feliz por ter sido seu aluno, professora. Muito obrigado por tudo, de verdade.

Também vale lembrar de Kevelyn, Larissa, Isabel, Gabriel, Pamella, Heitor, Nícolas, Nicollas, Matheus R., Matheus P., Lucas, Taciana, Anderson e Luciano, além de outras pessoas que, de distintas formas, participaram desta jornada. Muito obrigado!

”Os maiores cientistas também são artistas.”
(Albert Einstein)

Resumo

Este trabalho trata do uso de técnicas dos métodos de *Knowledge Discovery in Databases* — KDD — e *Cross Industry Standard Process for Data Mining* — CRISP-DM — sobre bases de dados educacionais disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira — INEP — visando à descoberta de conhecimento que permita a identificação, assim como a compreensão, do contexto de formação de discentes de Instituições de Ensino Superior — IES — públicas brasileiras. Três cenários de mineração de dados são observados, tendo em vista métodos do Aprendizado de Máquina Supervisionado e do Aprendizado de Máquina Não Supervisionado, abrangendo experimentos de classificação, agrupamento e associação de dados. O primeiro cenário, que contempla dados de concluintes de cursos de graduação de graus bacharelado e licenciatura, objetiva prever o tempo aproximado de conclusão da graduação, considerando informações socioeconômicas dos estudantes, por meio de 16 modelos de classificação construídos com o emprego de algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural Perceptron Multicamadas. Os modelos XGBoost tiveram os melhores resultados em todos os experimentos. Por sua vez, o segundo cenário utiliza o algoritmo K-Means para a execução de um agrupamento de IES públicas que, a partir da análise de quatro grupos obtidos com a consideração de informações sobre despesas, quantidades de docentes e técnicos, localização e categoria administrativa das IES, entre outras, possibilitou a identificação de similaridades e dissimilaridades entre as instituições. Os grupos em questão, além de dados utilizados no primeiro cenário, que incluem informações sobre os estudantes, como faixa etária, tempo de graduação e forma de ingresso na graduação, observando se esse ocorreu por meio de políticas de ação afirmativa ou de inclusão social, entre outras, são considerados nos experimentos do terceiro cenário, com o uso do algoritmo Apriori, para a geração de regras de associação que podem suportar a descoberta de conhecimento no âmbito do ensino superior público brasileiro.

Palavras-chave: ENADE, Censo da Educação Superior, KDD, CRISP-DM, Aprendizado de Máquina.

Abstract

This work deals with the use of techniques from the methods of Knowledge Discovery in Databases — KDD — and Cross Industry Standard Process for Data Mining — CRISP-DM — on educational databases made available by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (National Institute of Educational Studies and Research Anísio Teixeira) — INEP — aiming to the discovery of knowledge that allows the identification, as well as the understanding, of the context in the formation of students from public Brazilian higher education institutions. Three data mining scenarios are observed, considering Supervised Machine Learning and Unsupervised Machine Learning methods, covering data classification, grouping and association experiments. The first scenario, which includes data from graduates of bachelor's and licentiate's undergraduate courses, aims to predict the approximate length of stay at graduation, considering the students' socioeconomic information, through 16 classification models built using Decision Tree, Random Forest, XGBoost and Multilayer Perceptron Neural Network algorithms. XGBoost models had the best results in all experiments. In turn, the second scenario uses the K-Means algorithm to perform a grouping of public higher education institutions that, based on the analysis of four groups obtained by considering information on expenses, numbers of professors and technicians, location and administrative category, among others, made it possible to identify similarities and dissimilarities between the institutions. The groups, in addition to data used in the first scenario, which include information about the students, such as age group, length of stay at graduation and form of admission to graduation, noting whether this occurred through affirmative action or social inclusion policies, among others, they are considered in the experiments of the third scenario, using the Apriori algorithm, for the generation of association rules that can support the discovery of knowledge in the context of Brazilian public higher education.

Keywords: ENADE, Higher Education Census, KDD, CRISP-DM, Machine Learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
ENEM	Exame Nacional do Ensino Médio
ENADE	Exame Nacional de Desempenho dos Estudantes
IES	Instituição de Ensino Superior
IGC	Índice Geral de Cursos
KDD	*Knowledge Discovery in Databases*
CRISP-DM	*CRoss Industry Standard Process for Data Mining*
Cine Brasil	Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequencias de Formação Específica do Brasil
XGBoost	*eXtreme Gradient Boosting*
MLP	*Multilayer Perceptron Neural Network*
EDM	*Educational Data Mining*
AUC/ROC	*Area Under the Curve/Receiver Operating Characteristic*
IBGE	Instituto Brasileiro de Geografia e Estatística

Sumário

1 Introdução

De acordo com o Artigo 205 da Constituição da República Federativa do Brasil, a educação, direito de todos e dever do Estado, deve ser promovida visando ao pleno desenvolvimento do indivíduo e à sua qualificação profissional (BRASIL, 1988). Para o cumprimento de competências constitucionais relacionadas à educação, políticas públicas educacionais são planejadas e executadas pela Administração Pública nas esferas federal, estadual e municipal.

A Lei de Acesso à Informação, observando o princípio da publicidade da Administração Pública, explícito no caput do Artigo 37 da Constituição Federal, estabelece que é dever dos órgãos e entidades públicas promover, sem a necessidade de requerimentos, a divulgação, em local de fácil acesso, de informações, que podem ser de interesse coletivo ou geral, por eles produzidas ou custodiadas (BRASIL, 2011). Diante disso, órgãos e entidades públicas devem realizar a chamada abertura de dados, que permite a fiscalização da atuação do Estado pela população.

Cabem ao Ministério da Educação e às secretarias de educação estaduais e municipais, em âmbito nacional, estadual e municipal, respectivamente, as atividades de coleta, junção, tratamento e disponibilização de dados, processados ou não, que, de acordo com a Lei de Acesso à Informação, podem ser empregados para a produção de conhecimento, contidos em qualquer meio, suporte ou formato (BRASIL, 2011).

Em âmbito nacional, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira — INEP —, autarquia federal vinculada ao Ministério da Educação, possui diversas atribuições, incluindo o monitoramento e a avaliação do Sistema Nacional de Educação (INEP, 2021i). As atividades que compreendem a realização do Censo Escolar da Educação Básica e do Censo da Educação Superior são competências do INEP, assim como as que abrangem a execução do Exame Nacional do Ensino Médio — ENEM — e do Exame Nacional de Desempenho dos Estudantes — ENADE — (INEP, 2021c). Para promover a abertura de dados, o INEP elabora e publica, anualmente, bases de dados dos sistemas em questão. Os dados pessoais de identificação de discentes e docentes não constam nas bases disponibilizadas ao público.

O ENADE é um instrumento empregado para avaliar o rendimento dos concluintes de cursos de graduação brasileiros sobre conteúdos programáticos previstos nas diretrizes curriculares dos cursos, bem como para avaliar o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional dos estudantes. O exame é realizado com a observação de um ciclo que compreende três anos como período de avaliação. Em cada ano, cursos de graduação específicos,

de áreas do conhecimento determinadas, são avaliados (INEP, 2021d). O Apêndice A deste trabalho apresenta os cursos avaliados em cada ano do ciclo.

Além da prova propriamente dita, o ENADE conta com um questionário de caráter socioeconômico cujo objetivo é coletar informações que permitam compreender o perfil do estudante e o contexto de sua formação para subsidiar processos avaliativos gerais de cursos de graduação e de Instituições de Ensino Superior — IES — de todo o Brasil. O preenchimento do questionário pelo estudante é requisito para a participação no exame (INEP, 2021h).

O Censo da Educação Superior caracteriza uma ferramenta de pesquisa sobre IES brasileiras. Por meio dele, o INEP subsidia o Ministério da Educação na consideração de suas competências ao tornar possível a compreensão do sistema brasileiro de educação superior. O Censo coleta, por exemplo, dados acerca de condições de infraestrutura, bem como sobre cursos, vagas ofertadas e matrículas das IES, observando os variados tipos de organização acadêmica e categoria administrativa existentes. O Censo contribui para os trabalhos de gestores do governo e de entidades públicas e privadas, instituições de ensino e organismos internacionais, bem como de pesquisadores e especialistas brasileiros e estrangeiros (INEP, 2021a).

Este trabalho tem como objetos de estudo dados das edições de 2016, 2017 e 2018 do ENADE e de 2018 do Censo da Educação Superior — constantes nas bases de IES, Cursos de Graduação e Docentes —, além de dados da base do Índice Geral de Cursos — IGC — de 2019 (INEP, 2021g; INEP, 2021f; INEP, 2021e). As edições do ENADE citadas foram observadas pois caracterizam o ciclo de avaliação completo mais recente realizado até o início da pesquisa, assim como as edições do Censo da Educação Superior e do IGC foram selecionadas por serem as mais recentes disponíveis quando os dados foram coletados. Os dados em questão foram empregados para basear a descoberta de conhecimento por meio de técnicas de análise exploratória e 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. Neste trabalho, a mineração de dados abrange métodos do Aprendizado de Máquina Supervisionado e do Aprendizado de Máquina Não Supervisionado.

Tratando do cenário de mineração de dados que considera métodos do Aprendizado de Máquina Supervisionado, 16 modelos de classificação foram construídos com o uso de algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural Perceptron Multicamadas para prever o tempo de graduação, designado pelo período entre o ano de início da graduação e o ano de realização do ENADE, de estudantes de IES públicas. Quanto aos cenários do Aprendizado de Máquina Não Supervisionado, modelos foram construídos com o uso dos algoritmos K-Means e Apriori para agrupar IES públicas e associar dados de seus estudantes.

A Figura 1 exibe 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 foram executadas sobre os dados observados. 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 (INEP, 2021b). As áreas do conhecimento em questão podem ser vistas por meio do Apêndice B.

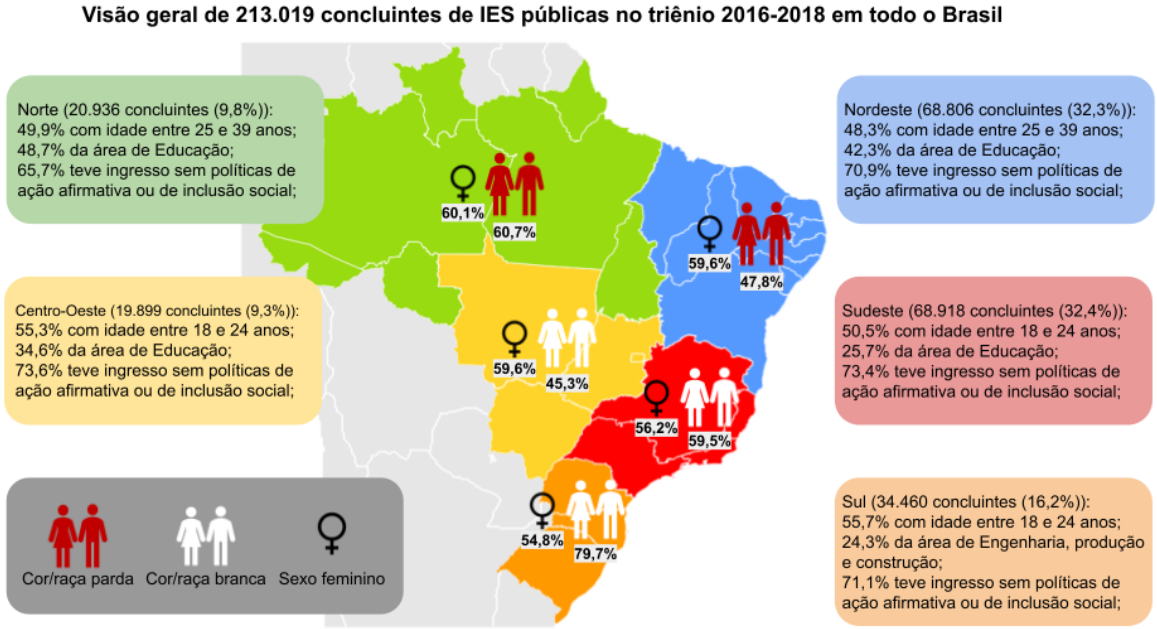


Figura 1 – Visão geral dos concluintes de IES públicas federais e estaduais de todo o Brasil que participaram do ENADE entre 2016 e 2018. Fonte: o autor (2021).

1.1 Motivação e Justificativa

A motivação deste trabalho está baseada no interesse em adquirir o respaldo científico necessário para evidenciar a possibilidade de descobrir conhecimento a partir dos dados, método e ferramentas considerados, o que pode permitir uma melhor compreensão sobre o contexto de formação de discentes em IES públicas federais e estaduais de todo o Brasil, além de uma reflexão acerca de similaridades e dissimilaridades entre as instituições, observando várias de suas características.

Acredita-se que os resultados percebidos em um estudo como este podem contribuir tanto para o aprimoramento de processos de ensino-aprendizagem quanto para a melhoria da gestão de recursos de IES públicas brasileiras.

1.2 Objetivos

Este estudo observa técnicas de Ciência de Dados (*Data Science*), Mineração de Dados (*Data Mining*), Inteligência Artificial (*Artificial Intelligence*) e Aprendizado de Máquina (*Machine Learning*) para investigar e diagnosticar padrões em dados abertos do ensino superior público no Brasil visando à descoberta de conhecimento acerca do tempo de permanência na graduação e perfis socioeconômicos de discentes de IES públicas federais e estaduais, tratando, inclusive, de características dessas instituições. Para isso, o trabalho tem como objetivos gerais:

- Compreender as bases de dados consideradas com o uso de técnicas de análise exploratória de dados;
- Selecionar conjuntos de dados de interesse a serem empregados visando à descoberta de conhecimento;
- Executar tarefas de pré-processamento e transformação sobre os dados selecionados para possibilitar a mineração de dados.

Quanto aos objetivos específicos deste estudo, busca-se:

- Prever o tempo aproximado de permanência na graduação de concluintes de cursos de graus bacharelado e licenciatura em IES públicas federais e estaduais brasileiras, considerando dados socioeconômicos, a partir de classificadores construídos com o uso de métodos do Aprendizado de Máquina Supervisionado;
- Identificar similaridades e dissimilaridades entre características de IES públicas federais e estaduais brasileiras, observando dados sobre infraestrutura e investimento, por meio de um agrupamento efetuado com a aplicação de métodos do Aprendizado de Máquina Não Supervisionado;
- Identificar os perfis socioeconômicos de concluintes de cursos de graduação de graus bacharelado e licenciatura em IES públicas federais e estaduais brasileiras, contemplando os resultados do agrupamento citado no objetivo anterior, por meio de regras de associação geradas com o emprego de métodos do Aprendizado de Máquina Não Supervisionado.

1.3 Estrutura do Trabalho

Este trabalho é constituído por seis capítulos, sendo este, de introdução, o primeiro, onde foram apresentados os objetos do estudo. A seguir estão dispostas breves descrições acerca do que os capítulos seguintes tratam:

- Capítulo 2 – Fundamentação Teórica: capítulo onde são expostos e definidos os instrumentos e conceitos utilizados no estudo;
- Capítulo 3 – Método e Ferramentas: capítulo que expõe o método observado no trabalho, bem como as ferramentas empregadas;
- Capítulo 4 – Cenários de Mineração de Dados: capítulo que apresenta os cenários de mineração de dados elaborados;
- Capítulo 5 – Resultados e Discussão: capítulo que exibe os resultados obtidos em cada cenário de mineração executado, tal como a sua discussão;
- Capítulo 6 – Considerações Finais: capítulo que contém as considerações finais sobre o trabalho.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos considerados para o desenvolvimento do estudo, abordando a definição de dados abertos, técnicas previstas pelos processos de KDD e CRISP-DM e operações de discretização, estratificação, binarização e *undersampling*, tratando, também, de análise exploratória e mineração de dados, com a observação de métodos de classificação, agrupamento e associação. Por fim, trabalhos relacionados a este são citados.

2.1 Dados Abertos

De acordo com a *Open Knowledge Foundation*, dados são abertos quando podem ser livremente usados, modificados e compartilhados por qualquer indivíduo para qualquer propósito (OKFN, 2021).

No âmbito da administração pública brasileira, o Tribunal de Contas da União indica cinco motivos para a abertura de dados, tratando de transparência na gestão pública, contribuição da sociedade com serviços inovadores ao cidadão, aprimoramento na qualidade dos dados governamentais, viabilização de novos negócios e obrigatoriedade por lei (BRASIL, 2015).

2.2 KDD e CRISP-DM

O termo *Knowledge Discovery in Databases*, que caracteriza o acrônimo KDD, foi apresentado em 1992 por Frawley, Piatetsky-Shapiro e Matheus (1992). Em *From Data Mining to Knowledge Discovery in Databases*, Fayyad, Piatetsky-Shapiro e Smyth (1996) destacam que o conhecimento é o produto final de uma descoberta baseada em dados e distinguem o processo de KDD de mineração de dados.

Fayyad, Piatetsky-Shapiro e Smyth (1996) dizem que o processo de KDD possui cinco etapas que, abrangendo tarefas de seleção, pré-processamento, transformação e mineração de dados e interpretação e avaliação de resultados, visam à descoberta de conhecimento útil a partir de dados. Os autores salientam que a mineração de dados é uma etapa do processo, que designa um método não trivial de identificação de padrões inéditos e válidos. Cada etapa depende da finalização da etapa anterior, que pode ser repetida quantas vezes for necessário até que os dados estejam adequados para uso posterior. O método de KDD é exposto na Figura 2.

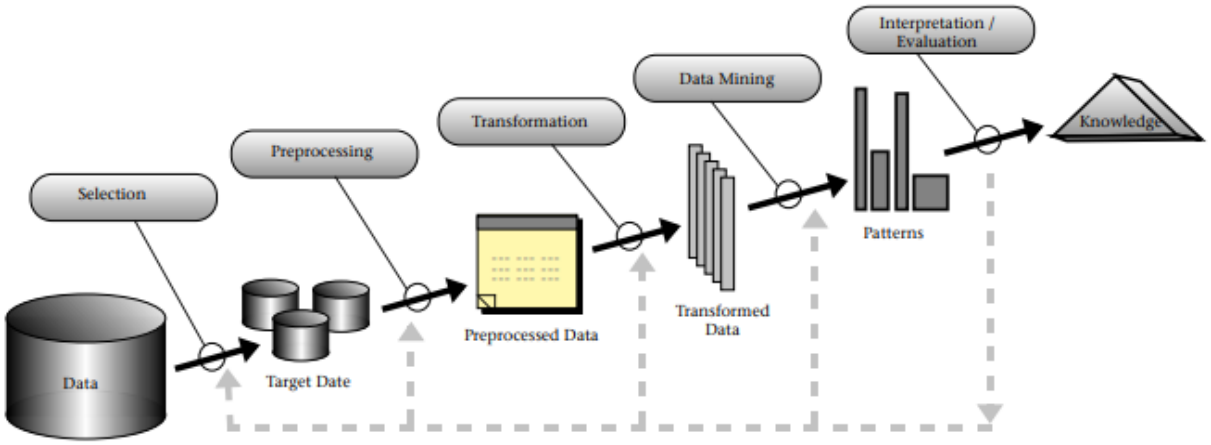


Figura 2 – Visão geral do KDD. Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996).

A primeira etapa do processo de KDD, denominada seleção de dados, é extremamente importante, uma vez que, para cada base de dados observada, há a necessidade de selecionar variáveis ou atributos e registros a serem utilizados na mineração de dados, designando a criação de subconjuntos de dados de interesse, também chamados de *target data* (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Após a seleção, é preciso analisar os dados dos conjuntos criados para identificar valores ausentes e outras inconsistências. A partir disso, para possibilitar a sequência do estudo, deve-se corrigir ou descartar os registros inconsistentes, de forma a não comprometer a qualidade dos modelos a serem construídos. Essas análises são compreendidas pela segunda etapa do método de KDD, denominada pré-processamento de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

As operações executadas durante a terceira etapa do KDD, denominada transformação de dados, designam tarefas de análise, formatação e reorganização de atributos e registros, de acordo com as necessidades percebidas para o atingimento dos objetivos do trabalho. Atributos podem ser criados ou alterados por meio de operações de discretização, estratificação, codificação e *undersampling*, entre outras (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A quarta etapa do processo de KDD, denominada mineração de dados, abrange o uso de métodos que possibilitam a descoberta de conhecimento. Os dados tratados nas etapas anteriores são utilizados para a identificação de padrões e correlações, objetivando a geração de informações úteis, a partir da observação de técnicas estatísticas, inteligência artificial e aprendizado de máquina (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Por fim, a quinta etapa do KDD abrange a execução de tarefas de interpretação e avaliação dos resultados obtidos pelos modelos construídos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Algumas métricas de avaliação, instrumentos empregados

nesta etapa, são abordadas em uma seção posterior.

De forma distinta do KDD, que visa à descoberta de conhecimento por meio dos dados propriamente ditos, o *Cross Industry Standard Process for Data Mining*, de acrônimo CRISP-DM, tem como objetos de estudo questões de negócio que possibilitam 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 (SHEARER, 2000). Esse método apresenta tarefas semelhantes às previstas pelo KDD em algumas de suas etapas. O CRISP-DM é exposto na Figura 3.

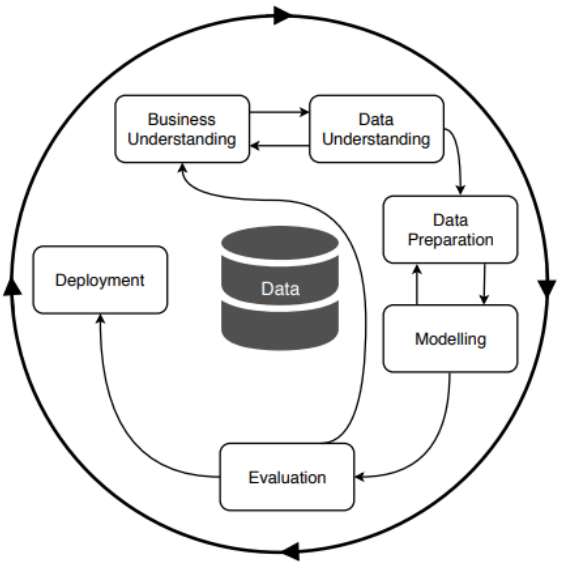


Figura 3 – Visão geral do CRISP-DM. Fonte: Martínez-Plumed et al. (2019).

A primeira etapa do CRISP-DM trata de entender o negócio considerado, abordando o seu objetivo. Nas etapas seguintes, de entendimento e preparação dos dados, o CRISP-DM prevê a compreensão das bases de dados observadas e, como o KDD, a execução de operações de processamento sobre os dados, de maneira a transformá-los, buscando tornar possível o seu uso para a mineração. A etapa seguinte, de modelagem, abrange, de fato, a execução de tarefas de mineração, visando ao atingimento do objetivo anteriormente definido, buscando a descoberta de conhecimento. Os resultados percebidos são avaliados na etapa de avaliação. Na última etapa do processo, de implantação, os resultados são apresentados e, a partir disso, podem ser utilizados em outras atividades (SHEARER, 2000).

Bastante empregado em trabalhos de mineração de dados que dependem do entendimento do negócio, o CRISP-DM costuma ser usado, também, no contexto educacional, tendo, inclusive, recebido uma proposta de adaptação, apresentada por Ramos et al. (2020), para basear a mineração na área.

2.3 Discretização, Estratificação, Binarização e *Undersampling* de Dados

Visando à mineração de dados, diversas operações podem ser executadas sobre os dados observados (HAN; KAMBER; PEI, 2012). Esta seção apresenta as tarefas de discretização, estratificação, binarização e *undersampling* de dados.

Atributos numéricos, cujas categorias são valores contínuos, podem ser discretizados de forma a equilibrar as frequências ou quantidades de observações e permitir o uso correto de algoritmos de aprendizado de máquina. A discretização consiste em apresentar os dados de maneira alternativa, por meio de intervalos ou faixas de valores. A partir dessa técnica de transformação, atributos originalmente contínuos tornam-se discretos, ou seja, as suas categorias passam a ter quantidades específicas e limitadas de valores possíveis (HAN; KAMBER; PEI, 2012).

A estratificação de dados consiste na separação de um conjunto de dados em subconjuntos de forma que cada subconjunto seja uma parte proporcional do conjunto original. A observação dessa operação é importante para que a mineração seja feita de forma correta, utilizando as devidas proporções de dados no treinamento e avaliação de modelos de aprendizado (HAN; KAMBER; PEI, 2012).

Vários algoritmos de aprendizado de máquina requerem o emprego de dados numéricos para a sua execução. Observando que a existência de dados não numéricos em uma base de dados é bastante comum, como aqueles resultantes de operações de discretização, é preciso representar os dados não numéricos em formato numérico, o que pode ser feito de diversas maneiras, por meio de codificação. Quando os dados têm o mesmo peso, é preciso usar um método de codificação que não altere tal característica. Variáveis *dummy* podem ser utilizadas, uma vez que possibilitam representar de forma binária dados de atributos que, em seu formato original, possuem duas ou mais categorias (SCIKIT-LEARN, 2021e).

Outra técnica de transformação de dados comumente utilizada corresponde à operação de *undersampling*, que designa uma solução para o problema de lidar com dados em quantidades desbalanceadas. Usar dados onde duas ou mais categorias de um atributo têm quantidades distintas de ocorrências em comparação com uma categoria específica pode ocasionar erros de aprendizado. A técnica de *undersampling* trata da remoção de determinada quantidade de observações de categorias predominantes, de forma que todas as categorias passem a ter a mesma frequência (HAN; KAMBER; PEI, 2012).

2.4 Análise e Mineração de Dados

A análise exploratória de dados caracteriza uma etapa fundamental em trabalhos com bases de dados. A análise trata da aplicação de métodos estatísticos sobre os dados, assim como a realização de tarefas de organização, resumo e visualização de informações, objetivando a compreensão dos objetos observados. Diante disso, a análise busca a apresentação das principais características dos dados considerados por meio de métodos visuais, visando, entre outros fins, à identificação de tendências (Tukey, 1977).

Segundo Han, Kamber e Pei (2012), a mineração de dados designa o processo de descoberta de padrões interessantes e conhecimento a partir de grandes quantidades de dados, cujas fontes podem ser variadas. O processo compreende operações de seleção, limpeza, integração e transformação de dados, além da mineração de dados propriamente dita, assim como tarefas de avaliação e apresentação do conhecimento obtido.

No âmbito educacional, a mineração de dados, denominada Mineração de Dados Educacionais (*Educational Data Mining* — EDM), visa ao desenvolvimento de métodos para explorar dados provenientes de ambientes educacionais em larga escala, de maneira a melhor compreender os estudantes, os seus contextos de aprendizado e assuntos relacionados (EDUCATIONALDATAMINING.ORG, 2021).

2.4.1 Aprendizado de Máquina Supervisionado e Não Supervisionado

O Aprendizado de Máquina trata da investigação de formas por meio das quais os computadores são capazes de aprender com base em dados. Uma das principais linhas de pesquisa da área abrange o aprendizado automático e o reconhecimento de padrões complexos visando à tomada de decisões inteligentes a partir de dados (HAN; KAMBER; PEI, 2012).

Han, Kamber e Pei (2012) afirmam que o Aprendizado de Máquina Supervisionado é um sinônimo para classificação de dados. A supervisão ocorre com o emprego de dados rotulados no treinamento dos modelos. Os autores também dizem que Aprendizado de Máquina Não Supervisionado é, em essência, um sinônimo de agrupamento, pois os dados de entrada não são rotulados e não há supervisão.

2.5 Classificação de Dados

De acordo com Witten, Frank e Hall (2011), a classificação caracteriza o método de análise onde um modelo classificador é construído para prever classes, que são representadas por valores discretos, por meio de dados. Para a criação de modelos de

classificação, o método de validação cruzada pode ser considerado, bem como algoritmos de Árvore de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*), XGBoost e Rede Neural Perceptron Multicamadas (*Multilayer Perceptron Neural Network* — MLP), entre outros, podem ser empregados. Esta seção trata desses conceitos.

2.5.1 Validação Cruzada

O método de reamostragem de dados denominado validação cruzada de *k* grupos (*k-fold cross-validation*) costuma ser utilizado com a observação de uma base de dados limitada para estimar como o modelo construído é capaz de prever ao considerar dados não usados durante o treinamento (WITTEN; FRANK; HALL, 2011).

O método funciona da maneira descrita a seguir. Primeiro, o conjunto de dados é embaralhado aleatoriamente; depois, o conjunto é dividido em *k* grupos; *k* vezes, um dos grupos deve ser utilizado como conjunto de dados de teste e os grupos restantes devem ser usados como conjuntos de treinamento; assim, para cada iteração, um modelo é treinado com os dados de treinamento e avaliado com os dados de teste, e suas pontuações são mantidas, enquanto o modelo é descartado; após todas as iterações, a habilidade geral de um modelo é resumida por meio das amostras de pontuação de *k* modelos. Em outras palavras, o método de validação cruzada de *k* grupos trata de treinar e avaliar modelos de aprendizado de máquina *k* vezes para resumir a habilidade de um modelo geral em prever. O tempo necessário para a construção do modelo com validação cruzada de *k* grupos é proporcional ao valor de *k* (WITTEN; FRANK; HALL, 2011).

2.5.2 Algoritmos de Classificação

O método de classificação de Árvore de Decisão designa um fluxograma que remete à uma árvore. Após a construção do modelo, é possível classificar um registro seguindo o fluxo da árvore, de cima para baixo, do nó raiz até uma folha. As principais vantagens da aplicação do método estão baseadas no fato de que as decisões tomadas pela árvore consideram as regras mais relevantes, que têm um bom grau de precisão (HAN; KAMBER; PEI, 2012)

A classificação com a observação do método de Floresta Aleatória ocorre por meio de um conjunto de árvores de decisão usado para aprimorar os resultados obtidos pelo método anterior. Classificar com uma árvore de decisão possui a desvantagem de que, quanto mais profunda a árvore for, maior será a chance de o modelo sofrer sobreajuste (*overfitting*). Porém, o método de Floresta Aleatória não costuma lidar com sobreajuste, pois constrói subconjuntos aleatórios com características das regras percebidas, gerando árvores menores. O principal hiperparâmetro do método é o número

de árvores consideradas. Quanto maior o número de árvores for, melhor será a performance do modelo (HAN; KAMBER; PEI, 2012).

De acordo com Chen e Guestrin (2016), *tree boosting* — impulsionamento de árvore, em tradução livre — é um método empregado sobre modelos de aprendizado de máquina altamente eficaz e amplamente utilizado. Os autores apresentam, no artigo *XGBoost: A Scalable Tree Boosting System*, o XGBoost, um sistema de aumento de árvore escalonável de ponta a ponta que promete, utilizando menos recursos, com a otimização de hardware e software, retornar resultados melhores do que os sistemas existentes.

O método de classificação por redes neurais é oriundo da psicologia e da neurobiologia e busca simular o comportamento de neurônios. Uma Rede Neural Perceptron Multicamadas designa um conjunto de unidades de entrada e saída conectadas por camadas intermediárias onde cada ligação tem um peso associado. Essa técnica costuma estar relacionada a ajustes de parâmetros bastante específicos e a longos períodos de treinamento. Uma das vantagens do uso de redes neurais trata do fato de que elas são capazes de identificar padrões para os quais jamais foram treinadas (WITTEN; FRANK; HALL, 2011).

2.5.3 Métricas de Avaliação

No âmbito da classificação de dados, tratando da avaliação de resultados dos modelos construídos, este trabalho observa as métricas de acurácia, precisão, *recall*, *f1-score* e *Area Under the Curve/Receiver Operating Characteristic* — AUC/ROC.

A acurácia caracteriza a média de predições identificadas de forma correta. Sensível a desbalanceamento de dados, essa métrica pode induzir a conclusões equivocadas sobre o desempenho de modelos de aprendizado (GOOGLE-DEVELOPERS, 2021a). A precisão associa valores de verdadeiros positivos com falsos positivos para retornar a proporção de identificações positivas corretas, enquanto a métrica de *recall* relaciona verdadeiros positivos e falsos negativos (GOOGLE-DEVELOPERS, 2021b). Ao contrário da acurácia, a *f1-score* usa verdadeiros e falsos positivos e negativos e é indicada na tentativa de não enviesar a avaliação das classes de maiores frequências quando há distribuição desigual de observações (SCIKIT-LEARN, 2021b). Por fim, a AUC/ROC considera as taxas de verdadeiros e falsos positivos para estimar o quão bem um modelo de classificação é capaz de prever (GOOGLE-DEVELOPERS, 2021c). Amplamente utilizada para a classificação binária, a AUC/ROC também pode ser empregada em problemas de classificação multiclasse, com o uso de estratégias como *One-vs-one* e *One-vs-the-rest* (SCIKIT-LEARN, 2021a).

2.6 Agrupamento de Dados

De acordo com Witten, Frank e Hall (2011), técnicas de *clusterização* ou agrupamento de dados são utilizadas quando não existem classes a serem previstas, mas há instâncias a serem separadas em grupos naturais. O Método do Cotovelo costuma ser usado nesse contexto, bem como o famoso algoritmo de agrupamento K-Means.

O Método do Cotovelo é um instrumento gráfico que auxilia na definição do número ideal de *clusters* ou grupos a serem gerados em um experimento de agrupamento a partir da observação da base de dados a ser empregada. Se o gráfico retornado pelo método assemelha-se a um braço, o ponto de inflexão na curva, que seria o cotovelo, é uma boa indicação de que o modelo tem melhores resultados nesse ponto (YELLOWBRICK, 2021).

O K-Means é um algoritmo de Aprendizado de Máquina Não Supervisionado bastante conhecido e empregado que trata de agrupamento de dados. Esse algoritmo, considerado um método simples e efetivo, a partir de um valor de entrada k, 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. Em seguida, o centroide ou a média das instâncias de cada grupo é calculado. Os centroides são tidos como os novos valores de centro dos respectivos grupos e todo o processo é repetido com os novos centros de grupos até que os mesmos pontos sejam atribuídos para cada grupo várias vezes consecutivas, o que sugere que os centros dos grupos estabilizaram-se e permanecerão os mesmos para sempre. A inicialização *k-means++* aprimora a velocidade e acurácia do K-Means com uma escolha cuidadosa dos centros de grupo iniciais, por meio do que conhece-se como sementes (*seeds*) (WITTEN; FRANK; HALL, 2011).

2.7 Associação de Dados

Regras de associação designam um método de mineração de dados que relaciona as ocorrências de itens de uma base de dados de forma 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 de Aprendizado de Máquina Não Supervisionado Apriori objetiva encontrar todos os conjuntos de itens possíveis em uma base de dados, recebendo valores mínimos para suporte e confiança como parâmetros (AGRAWAL; SRIKANT, 1994).

Sejam X e Y dois conjuntos de itens ou de categorias para um atributo de uma base. O suporte caracteriza a porcentagem de transações ou de 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 (AGRAWAL; SRIKANT, 1994). 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. Enfim, 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 onde o *lift* é muito maior do que 1 (IBM, 2021).

2.8 Trabalhos Relacionados

Esta seção apresenta trabalhos relacionados a este, considerando os cenários de mineração de dados elaborados, que compreendem o emprego de técnicas de análise exploratória de dados ou de métodos de classificação, agrupamento ou associação de dados, do Aprendizado de Máquina Supervisionado e Aprendizado de Máquina Não Supervisionado, sobre dados educacionais.

O trabalho de Nicolini, Andrade e Torres (2013) usou dados do ENADE de 2009 para verificar o desempenho de estudantes em universidades, centros universitários e faculdades públicas e privadas, de cursos de graduação de bacharelado em Administração, por meio de análise estatística, com o uso do software *Statistical Package for the Social Sciences*.

Carvalho, Cruz e Gouveia (2017) desenvolveram um trabalho que contemplou métodos do Aprendizado de Máquina Supervisionado para classificar, usando algoritmos de Árvore de Decisão e Naive Bayes, dados do Censo da Educação Superior de 2014 e 2015 com o objetivo de determinar a localização da IES do estudante a partir de informações socioeconômicas. A pesquisa de Souza (2020) tratou de construir modelos de classificação, com base em vários algoritmos, para abordar a evasão escolar, considerando dados de estudantes de uma IES privada de Minas Gerais. Os modelos buscaram prever se um estudante pode tornar-se evadido.

Oliveira e Brito (2019) empregaram dados de IES públicas do Censo da Educação Superior de 2017 para a execução de agrupamentos, com base nos algoritmos K-Means e K-Medoids e a observação do Coeficiente da Silhueta e do Índice de Calinski-Harabasz. Souza et al. (2009), visando identificar similaridades entre IES brasileiras que ofertam cursos da área de Ciências Contábeis, considerando a produção científica e o número de pesquisadores das instituições, realizaram um agrupamento, por meio da técnica de análise de *cluster* com o método Ward, a partir de dados sobre

artigos publicados em eventos. Dionisio et al. (2015) realizaram uma pesquisa com o objetivo de agrupar IES públicas portuguesas, observando as formas de organização e as tipologias das instituições, utilizando métodos de análise estatística multivariada.

Silva, Hoed e Saraiva (2019) utilizaram métodos do Aprendizado de Máquina Não Supervisionado para a geração de regras de associação sobre dados do ENADE de 2017, com o uso do algoritmo Apriori, buscando analisar fatores capazes de influenciar o desempenho de estudantes de cursos superiores de Computação. Silva et al. (2020) desenvolveram um estudo que observou técnicas de mineração de dados visando à identificação de desigualdades sociais a partir de dados do ENEM de 2019 por meio de experimentos de agrupamento e associação.

Por fim, Morais et al. (2021) usaram Modelagem por Equações Estruturais (*Structural Equation Modeling* — SEM) para examinar ”configurações universitárias”, características socioeconômicas, financiamento estudantil e políticas de cotas associadas a graduandos brasileiros no período entre 2013 e 2017.

Diante do exposto, percebe-se que diversos autores têm empregado dados do ENADE e do Censo da Educação Superior visando ao desenvolvimento de pesquisas sobre o ensino superior brasileiro, em especial, sobre o ensino superior público, o que indica que abordar o contexto de formação de estudantes de IES públicas, assim como tratar de questões de infraestrutura das instituições, é relevante. Este trabalho busca fornecer uma contribuição válida ao apresentar abordagens para a predição do tempo de permanência de discentes de cursos de graus bacharelado e licenciatura de IES públicas na graduação, tratar de perfis socioeconômicos desses estudantes e considerar as similaridades e dissimilaridades entre as características de suas instituições.

2.9 Considerações Finais

Este capítulo tratou, além de trabalhos que usaram dados, método e ferramentas relacionados aos que foram utilizados nesta pesquisa, dos principais instrumentos teóricos observados para o desenvolvimento do estudo. O emprego dos conceitos expostos é abordado em detalhes no capítulo seguinte.

3 Método e Ferramentas

O método empregado para a execução deste estudo é constituído pela junção de etapas dos métodos de *Knowledge Discovery in Databases* e *Cross Industry Standard Process for Data Mining*. As linguagens de programação Python, versão 3.8.3, e R, versão 4.1.0, além das bibliotecas de Python Numpy, Pandas, Pandas Profiling, Feature Engine, Scikit-Learn, Imbalanced-Learn, XGBoost, Yellowbrick, Seaborn e Matplotlib e a biblioteca de R Arules, nas versões mais recentes, e o RapidMiner Studio, versão 9.8, ativado com licença educacional, são as ferramentas utilizadas.

Este capítulo versa sobre as tarefas de seleção, pré-processamento, transformação e mineração de dados executadas neste trabalho. Essas tarefas compõem etapas homônimas do KDD, assim como integram as etapas de entendimento e preparação dos dados e modelagem do CRISP-DM.

3.1 Seleção de Dados

As bases de dados das edições de 2016, 2017 e 2018 do ENADE têm, respectivamente, 141, 150 e 137 atributos e, juntas, contêm 1.301.607 registros de estudantes inscritos (INEP, 2021g). Quanto aos dados da edição de 2018 do Censo da Educação Superior, a base de IES tem 48 atributos e 2.537 registros de IES públicas e privadas, enquanto a base de Docentes conta com 41 atributos e 397.893 registros e a base de Cursos de Graduação possui 112 atributos e 38.256 registros (INEP, 2021f).

Com o uso da ferramenta de automodelagem do RapidMiner Studio, considerando a tarefa de seleção de atributos mais relevantes de uma base de dados de entrada por meio de seus graus de correlação e estabilidade, após a definição do método de aprendizado de máquina a ser empregado, subconjuntos de atributos de interesse foram selecionados a partir dos dados observados (MIERSWA et al., 2006). 30 atributos do ENADE, que podem ser vistos no Apêndice C, foram selecionados para basear, após a execução das operações de pré-processamento e transformação de dados tratadas a seguir, o cenário de classificação deste trabalho. Para isso, no processo de automodelagem do RapidMiner Studio, houve a definição do atributo a ser predito no cenário em questão, apresentado no capítulo seguinte. Alguns dos 30 atributos selecionados também basearam o cenário de associação de dados deste estudo. O processo tratado também permitiu a seleção de 19 atributos da base de IES do Censo da Educação Superior, usados para basear o cenário de agrupamento do trabalho, que também podem ser visualizados no Apêndice C.

É válido salientar que o RapidMiner Studio foi utilizado somente para a identificação dos atributos mais relevantes das bases de dados consideradas. Após a identificação, a seleção de dados propriamente dita, ou seja, a criação de bases de dados derivadas, foi concretizada com o emprego da linguagem de programação Python e da biblioteca de Python Pandas, bem como outras tarefas de etapas posteriores.

Entre os 30 atributos selecionados das bases do ENADE, há informações sobre o ano de realização da prova, 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. —, o curso de graduação — como a área do conhecimento, modalidade, turno, região etc. —, a IES — a categoria administrativa —, e as respostas para o questionário socioeconômico — como o estado civil, cor/raça, renda familiar, motivo de escolha do curso de graduação, se o ingresso ocorreu por meio de políticas de ação afirmativa ou de inclusão social etc. Entre os 19 atributos selecionados na base de dados de IES do Censo da Educação Superior, constam informações como o código de identificação no e-MEC, nome, categoria administrativa, organização acadêmica, localização, quantidade de técnicos, despesas da instituição etc.

Neste trabalho, 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 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. Diante disso, foram selecionados 226.612 registros de participantes das três edições do ENADE citadas, além de 108 registros da base de dados de IES do Censo da Educação Superior, de instituições vinculadas aos estudantes em questão. As bases de Docentes e Cursos de Graduação do Censo da Educação Superior, bem como a base de dados do IGC das instituições, também foram empregadas.

3.2 Pré-processamento de Dados

Diversas verificações de inconsistências foram executadas sobre os dados observados por meio da linguagem de programação Python e das bibliotecas de Python Pandas e Pandas Profiling durante a etapa de pré-processamento de dados, bem como as devidas correções, quando necessárias. Várias operações de limpeza e padronização de dados também foram efetuadas, além da junção dos registros das três edições do ENADE em uma única base.

Especificamente, quanto à busca por inconsistências, as primeiras análises trataram de verificar se os registros de participantes do ENADE têm valores para o atributo

do 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. Se um registro possui valores iguais para o ano de realização do ENADE e o ano de início da graduação, ele deve ser removido do conjunto de dados, pois essa situação designa uma inscrição de um estudante ingressante em seu curso de graduação. Como este trabalho considera somente dados de estudantes concluintes, dados de ingressantes não são desejados. As análises restantes sobre os dados do ENADE verificam se o ano de conclusão do ensino médio e se ano de início da graduação do estudante são maiores do que 2018 e se a idade do participante é menor ou igual à diferença entre o ano de realização do exame e o ano de conclusão do ensino médio. Os registros que enquadraram-se em alguma das inconsistências citadas foram descartados.

Ainda tratando dos dados oriundos das bases do ENADE, quanto às operações de padronização de dados, 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. Algumas alterações gerais também foram realizadas sobre todos os dados, incluindo a tarefa de renomear os atributos para tornar os seus nomes intuitivos. Além disso, todos os registros do ENADE que têm valores ausentes, ou seja, registros que não contêm valores para determinados atributos, foram descartados.

Observando os dados do conjunto de IES do Censo da Educação Superior selecionado, as tarefas de pré-processamento efetuadas designam verificações de inconsistências, em especial, sobre os dados financeiros das instituições, além de alterações da estrutura dos atributos, abrangendo, por exemplo, o ato de renomeá-los.

Diante do exposto, uma ampla variedade de inconsistências foi verificada e tratada, tornando a base de dados unificada e processada do ENADE, com 213.019 registros, íntegra. Esta base tem dados de 94% dos concluintes de cursos de graduação de graus bacharelado e licenciatura de 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 2016 e 2018 em todo o Brasil. Assim, 6% (13.593) dos registros do conjunto inicialmente selecionado foram considerados inconsistentes e descartados. Os estudantes cujos dados foram observados para sequência do estudo estão vinculados às 108 instituições constantes na base de dados de IES.

3.3 Transformação de Dados

Considerando as atividades que constituem a etapa de transformação de dados, a partir de dados existentes, atributos foram criados ou transformados para assumir de-

terminadas estruturas, designando, por exemplo, alterações de categorias de valores possíveis. Além disso, tarefas de discretização e estratificação foram executadas sobre os dados, assim como métodos de binarização e *undersampling*. As tarefas citadas foram efetuadas 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.

Seis atributos foram criados para integrar a base de dados do ENADE, que fundamentou os experimentos de classificação e associação de estudantes na posterior etapa de mineração de dados. Alguns desses atributos foram criados a partir de dados oriundos de outras bases — por exemplo, o atributo que armazena o grau acadêmico do curso de graduação do estudante, criado por meio 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 tem a grande área do conhecimento do curso, criado a partir de sua área de formação geral e do manual para classificação de cursos Cine Brasil de 2018. Outro atributo criado, tendo em mente os cenários de classificação e associação de dados, possui a diferença entre o ano de realização do ENADE e o ano de início da graduação pelo estudante, definida como o tempo de graduação neste estudo. Quanto ao cenário de associação, é justo salientar a criação do atributo que trata da diferença entre o ano de conclusão do ensino médio e o ano de início da graduação. Todos os seis atributos criados no âmbito do ENADE podem ser visualizados em detalhes no Apêndice C.

Sobre o conjunto de dados de IES, que baseou o cenário de agrupamento, quatro atributos foram criados para integrá-lo. O primeiro possui a quantidade de docentes vinculados à IES, criado a partir de dados oriundos da base de Docentes do Censo da Educação Superior, enquanto outro possui os valores totais de despesas da IES, criado por meio da soma de todos os valores de despesas, originalmente separados em sete atributos. O terceiro atributo contém o IGC da instituição, criado a partir da base de dados de IGC. O último atributo, essencial para o cenário de associação de dados de estudantes, compreende o resultado do cenário de agrupamento de IES, designando o grupo ao qual a IES pertence. Os quatro atributos podem ser vistos em detalhes no Apêndice C.

Atributos do questionário socioeconômico do ENADE têm como categorias uma forma de expressar o quanto o participante do exame concorda com a afirmação contida na questão em consideração. Por exemplo, o atributo BIBLIOTECA_FISICA_IES, relacionado à questão nº 64 do questionário, contém a afirmação ”A biblioteca dispôs das referências bibliográficas que os estudantes necessitaram” e originalmente possui oito categorias: *“discordo totalmente”*, *“discordo”*, *“discordo parcialmente”*, *“concordo*

parcialmente”, *“concordo”*, *“concordo totalmente”*, *“não se aplica”* e *“não sei responder”*. Após a tarefa 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”*.

Observando os dados do ENADE, a discretização do atributo que armazena o período entre o ano de conclusão do ensino médio e o ano de início da graduação foi efetuada de forma automática por meio do método *EqualFrequencyDiscretiser* da biblioteca Feature Engine, de maneira a equilibrar as frequências de dados, bem como as operações de discretização dos atributos de IES que armazenam a despesa total e as quantidades de docentes e técnicos da instituição (FEATURE-ENGINE, 2021). Voltando a tratar do ENADE, as tarefas de discretização dos atributos que contêm a faixa etária e o tempo de graduação do estudante foram realizadas manualmente. No primeiro caso, houve a consideração das faixas etárias utilizadas pelo Instituto Brasileiro de Geografia e Estatística — IBGE — em seus estudos. No segundo, a discretização ocorreu após análises acerca do tempo de graduação em cursos de graus bacharelado e licenciatura, a partir de duas abordagens, como exposto na seção que apresenta o cenário de classificação de dados deste trabalho.

A estratificação de dados foi executada para cinco grupos, observando os experimentos de classificação, por meio do método *StratifiedKFold* da biblioteca Scikit-Learn (SCIKIT-LEARN, 2021c). Tratando da binarização, o método *OneHotEncoder* da biblioteca Scikit-Learn foi utilizado para a implementação de variáveis *dummy*, objetivando a codificação dos dados usados nos cenários de classificação e agrupamento deste estudo (SCIKIT-LEARN, 2021e). A operação de *undersampling*, realizada com o uso do método *RandomUnderSampler* da biblioteca Imbalanced-Learn, foi executada sobre o atributo que armazena o tempo de graduação, no contexto do cenário de classificação (IMBALANCED-LEARN, 2021).

Após as tarefas de pré-processamento e transformação, que compreenderam a criação de atributos a partir de atributos selecionados das bases de dados originais, assim como a alteração e a exclusão de atributos originais após o seu uso para a criação de novos, a base de dados do ENADE permaneceu com 30 atributos e a base de IES passou a ter 13. Os dados do ENADE foram duplicados e dispostos em duas bases, de maneira que cada base tivesse um conjunto de atributos específico, para possibilitar a execução dos cenários de classificação e associação de dados deste estudo. A depender do cenário e experimento que determinada base de dados fundamenta, seus atributos podem assumir categorias de valores distintas. Por meio dos apêndices deste trabalho, é possível observar os resultados das operações tratadas nesta seção. O conjunto de dados exibido no Apêndice D contém todos os 30 atributos do ENADE e foi utilizado nos experimentos de classificação. O conjunto de IES, que

já encontrava-se em uma base à parte, apresentado no Apêndice E, possui todos os 13 atributos de IES e fundamentou o cenário de agrupamento. O conjunto exposto no Apêndice F tem 22 atributos do ENADE e baseou os experimentos de associação.

3.4 Mineração de Dados

Diante dos objetivos deste trabalho, foram executados experimentos de classificação, no contexto do primeiro cenário de mineração de dados, do Aprendizado de Máquina Supervisionado, assim como um experimento de agrupamento, no segundo cenário, e experimentos de associação, no terceiro cenário, do Aprendizado de Máquina Não Supervisionado. A Tabela 1 tem uma visão geral dos cenários, tratando dos objetivos, método e ferramentas utilizados. Os referidos cenários e experimentos são apresentados em detalhes no capítulo seguinte.

Tabela 1 – Visão geral dos cenários de mineração de dados. Fonte: o autor (2021).

Cenário	**Base de dados**	**Objetivo e método**	**Ferramentas**
Classificação	30 atributos e 213.019 registros do ENADE	Construção de modelos para a previsão do tempo de graduação a partir de quatro abordagens de categorias e balanceamento da variável dependente	Python, Pandas, Scikit-Learn, XGBoost, Seaborn, MatplotLib e outras
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, além do atributo que identifica o grupo ao qual a IES do estudante está vinculada	Geração de regras de associação para a percepção de perfis socioeconômicos de discentes de IES públicas federais e estaduais por meio de 92 experimentos	R, Arules e outras

3.5 Considerações Finais

Ao apresentar o método e as ferramentas usadas para a execução do estudo, este capítulo expôs em detalhes as tarefas de seleção, pré-processamento e transformação de dados realizadas, previstas pelos processos de KDD e CRISP-DM. Entre as tarefas, há operações de seleção de atributos e registros e verificações de inconsistências e tratamento sobre registros, além da criação e alteração de atributos por meio de técnicas de discretização, estratificação, binarização e *undersampling* de dados.

4 Cenários de Mineração de Dados

Este capítulo trata dos três cenários de mineração de dados executados neste trabalho. É válido destacar que o primeiro cenário, de classificação, foi executado em âmbito local, enquanto os outros cenários, de agrupamento e associação, foram executados a partir da plataforma Google Colaboratory.

4.1 Cenário 1: Classificação de Dados de Estudantes

O primeiro cenário de mineração de dados deste trabalho tratou da construção e avaliação de modelos de classificação baseados em algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural Perceptron Multicamadas, por meio da linguagem de programação Python e das bibliotecas de Python Pandas, Scikit-Learn e XGBoost, para a previsão do tempo de permanência na graduação, ou simplesmente tempo de graduação, de estudantes. Buscou-se criar modelos de aprendizado de máquina supervisionado com boas pontuações de acurácia, precisão, *recall*, *f1-score* e AUC/ROC, o que caracteriza a possibilidade de descobrir conhecimento. Todos os modelos foram construídos com o uso do método de validação cruzada de cinco grupos.

Os modelos baseados no algoritmo de Árvore de Decisão foram construídos a partir do classificador *DecisionTreeClassifier* da biblioteca de Python Scikit-Learn, com os parâmetros padrões (SCIKIT-LEARN, 2021g). Por sua vez, os modelos de Floresta Aleatória foram implementados com base no classificador *RandomForestClassifier* da biblioteca de Python Scikit-Learn, e consideraram os parâmetros padrões, onde o número de árvores é 100 (SCIKIT-LEARN, 2021f). O classificador *XGBClassifier* da biblioteca de Python XGBoost foi utilizado para a criação dos modelos XGBoost, com os parâmetros padrões (XGBOOST, 2021). Por fim, para a construção de redes neurais, o classificador *MLPClassifier* da biblioteca de Python Scikit-Learn foi empregado, com parâmetros padrões (SCIKIT-LEARN, 2021d).

Como exposto anteriormente, este cenário foi baseado pelo conjunto de 30 atributos exibido no Apêndice D e teve a variável 'TEMPO_GRADUACAO', que armazena a diferença entre o ano de início da graduação e o ano de realização do ENADE, como dependente, ou seja, variável classe. O conjunto de dados tem todos os 213.019 registros de concluintes de cursos de graduação de graus bacharelado e licenciatura de IES federais e estaduais no Brasil entre 2016 e 2018 com tempo de graduação de 2 anos ou mais que foram percebidos após a execução de tarefas de pré-processamento e transformação de dados. Diante do exposto, o cenário tem como objetivo prever o tempo aproximado de permanência de estudantes de IES públicas brasileiras na gra-

duação. É válido destacar que a evasão acadêmica não é considerada nos dados do ENADE.

Quatro experimentos, que distinguem-se quanto às categorias da variável dependente e quanto ao balanceamento dos dados da variável, foram executados. Em cada experimento, quatro modelos de classificação foram construídos por meio de algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural Perceptron Multicamadas. Assim, 16 modelos foram implementados no âmbito do primeiro cenário de mineração de dados deste trabalho.

Tratando das categorias da variável dependente, os experimentos 1 e 2 observam duas classes — 'Entre 2 e 4 anos' e '5 anos ou mais' —, e os experimentos 3 e 4, três classes — 'Entre 2 e 4 anos', 'Entre 5 e 7 anos' e '8 anos ou mais' —, como pode ser visto na Tabela 2. Sobre os experimentos 1 e 2, cada um emprega um conjunto de dados específico. O primeiro usa um conjunto desbalanceado, com todos os dados possíveis, enquanto o segundo utiliza um conjunto balanceado, com os dados balanceados a partir da classe com a menor quantidade de observações. A situação quanto aos conjuntos de dados empregados é análoga nos experimentos 3 e 4. Dessa forma, os modelos construídos nos experimentos 1 e 3 foram treinados e avaliados com base em um conjunto de dados desbalanceado, enquanto os modelos dos experimentos 2 e 4 tiveram um conjunto de dados balanceado nessas tarefas.

Tabela 2 – Quantidades de observações dos conjuntos de dados empregados nos experimentos 1, 2, 3 e 4. Fonte: o autor (2021).

Classe	**Observações do conjunto desbalanceado**	**Observações do conjunto balanceado**
0: Entre 2 e 4 anos 1: 5 anos ou mais	119.947 93.072 (Experimento 1)	93.072 93.072 (Experimento 2)
0: Entre 2 e 4 anos 1: Entre 5 e 7 anos 2: 8 anos ou mais	119.947 83.882 9.190 (Experimento 3)	9.190 9.190 9.190 (Experimento 4)

A abordagem correta para a execução de experimentos de classificação compreende o balanceamento de classes da variável dependente na etapa de treino e o uso de todos os dados possíveis na etapa de teste. Entretanto, neste trabalho, decidiu-se executar os experimentos da maneira detalhada para a percepção das consequências da observação das abordagens de balanceamento de dados implementadas.

4.2 Cenário 2: Agrupamento de Dados de IES

No segundo cenário de mineração de dados deste estudo, houve a execução de um experimento de agrupamento de instituições de ensino superior públicas federais e estaduais brasileiras, a partir do algoritmo K-Means, com o emprego da inicialização *k-means++*, por meio da linguagem de programação Python e das bibliotecas de Python Pandas e Scikit-Learn, visando à obtenção de grupos de IES para a identificação de similaridades e dissimilaridades entre as suas características.

O experimento observou 108 registros de IES públicas brasileiras e 13 atributos que podem ser vistos no Apêndice E. Entre os atributos empregados, há informações sobre despesas totais, quantidades de docentes e técnicos, localização, categoria administrativa e Índice Geral de Cursos — IGC — da IES. O Método do Cotovelo (*Elbow Method*), implementado a partir do *KElbowVisualizer* da biblioteca de Python Yellowbrick, foi empregado para a definição da quantidade de grupos a serem gerados, com base na métrica da distorção (YELLOWBRICK, 2021).

Com a análise de características das IES, os grupos resultantes do experimento foram definidos. Dessa forma, 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, do último cenário de mineração de dados deste estudo.

4.3 Cenário 3: Associação de Dados de Estudantes e IES

Por fim, o terceiro cenário de mineração de dados deste trabalho tratou da execução de vários experimentos de associação sobre dados de estudantes de IES públicas federais e estaduais de todo o Brasil, observando os resultados obtidos no agrupamento do cenário anterior, com a utilização do algoritmo Apriori, por meio da linguagem de programação R e da biblioteca de R Arules. Os recursos disponibilizados por essa biblioteca justificam o seu uso.

Os experimentos de associação efetuados foram baseados em 24 subconjuntos da base de dados de 213.019 registros de participantes do ENADE entre 2016 e 2018, de cursos de graduação de graus bacharelado e licenciatura de IES públicas federais e estaduais de todo o Brasil, cujo tempo de graduação, tido neste estudo como o período caracterizado entre o ano de início da graduação e o ano de realização do ENADE, foi de 2 anos ou mais. 22 atributos, que podem ser vistos no Apêndice F, foram observados para a realização dos experimentos.

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 4, 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.

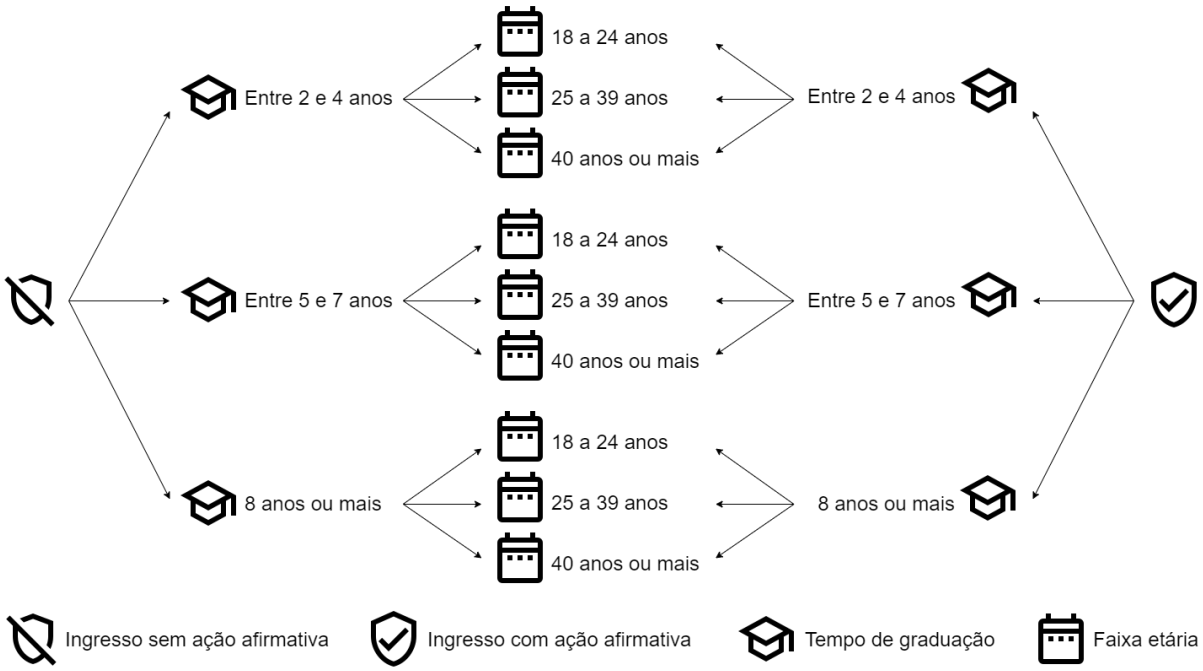


Figura 4 – Visão geral da estrutura dos experimentos de associação. Fonte: o autor (2021).

Os experimentos objetivaram a obtenção de regras a partir de duas abordagens. A primeira 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, concretizados 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 tempos 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 única de 213.019 registros de participantes do ENADE, observando a execução de perspectivas de associação específicas, com quantidades de registros de estudantes do contexto muito baixas, o suporte a ser definido deveria ser diretamente proporcional, e isso ainda não garantiria a percepção de regras com os atributos desejados. Diante do exposto, 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 e 5 a 7 anos, que tinham de 18 a 24 anos, 25 a 39 anos ou 40 anos ou mais, assim como de estudantes 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. Como destacado quando o primeiro cenário foi apresentado, é justo perceber que os dados empregados neste estudo tratam de concluintes de cursos de graduação em IES públicas que tiveram presenças válidas no ENADE e, dessa forma, a evasão estudantil não é considerada nessa abordagem.

4.4 Considerações Finais

Este capítulo tratou em detalhes dos três cenários de mineração de dados executados na busca pelo atingimento dos objetivos do estudo. Os cenários consideram métodos de classificação, agrupamento e associação, do Aprendizado de Máquina Supervisionado e do Aprendizado de Máquina Não Supervisionado. No Cenário 1, há a construção de 16 modelos de classificação a partir de algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural Perceptron Multicamadas e de dados soci-

oeconômicos de discentes de IES públicas brasileiras. O Cenário 2 trata de agrupar as IES desses discentes para a identificação de similaridades e dissimilaridades entre as características das instituições, com o uso do algoritmo K-Means. Por fim, o Cenário 3 busca a compreensão de perfis socioeconômicos dos discentes, observando o agrupamento do Cenário 2, por meio de regras de associação geradas com o algoritmo Apriori. Os resultados são apresentados no próximo capítulo.

5 Resultados e Discussão

Este capítulo expõe os resultados obtidos por meio da execução de experimentos nos três cenários de mineração de dados do trabalho, bem como a sua discussão.

5.1 Cenário 1: Classificação de Dados de Estudantes

A Figura 5 apresenta os resultados de classificação obtidos pelos modelos construídos nos quatro experimentos executados no contexto do primeiro cenário de mineração de dados. A Figura 6 mostra detalhes do aprendizado dos modelos XGBoost dos experimentos 1, 2, 3 e 4, enquanto as figuras 7, 8, 9 e 10 expõem as suas matrizes de confusão. As figuras 11, 12, 13 e 14 têm os gráficos de radar de todos os experimentos. As figuras 15, 16, 17 e 18 contêm as curvas AUC/ROC de todos os modelos por experimento. A Figura 19 exibe os tempos de treinamento e avaliação dos modelos. Detalhes do aprendizado, matrizes de confusão e curvas AUC/ROC dos modelos de Árvore de Decisão, Floresta Aleatória e Rede Neural Perceptron Multicamadas de todos os experimentos podem ser vistos por meio do Apêndice G.

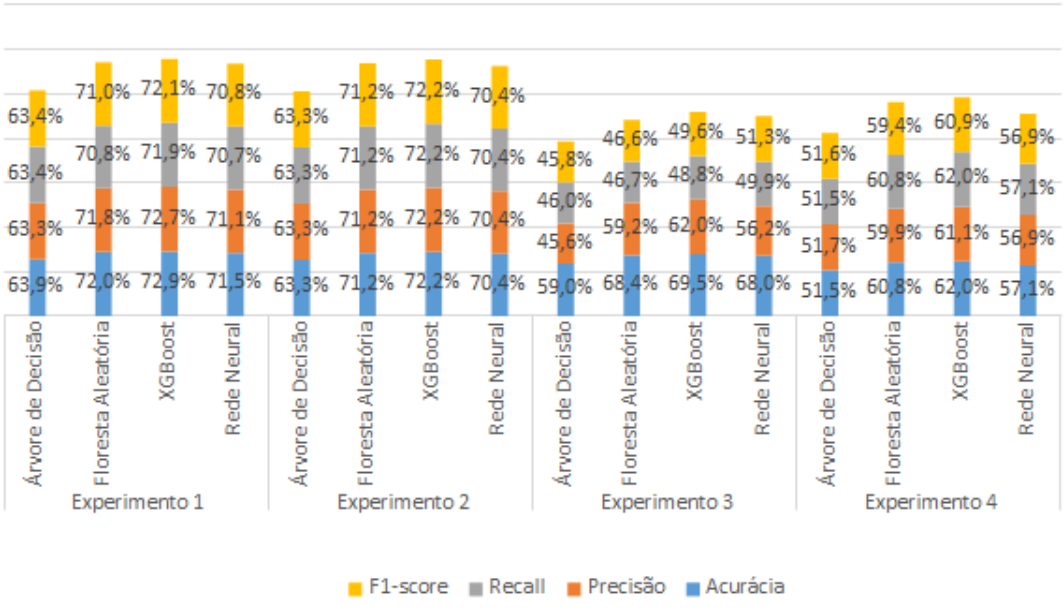


Figura 5 – Resultados de classificação. Fonte: o autor (2021).

Apesar do uso de dados balanceados e desbalanceados por categoria da variável dependente, os resultados, nos experimentos 1 e 2, que consideraram duas classes para a variável em questão, são bastante próximos, bem como nos experimentos 3 e 4, que observaram três classes. Os modelos construídos com o algoritmo XGBoost apre-

sentaram os melhores resultados em todos os experimentos. Os modelos de Floresta Aleatória e Rede Neural Perceptron Multicamadas tiveram resultados muito próximos.

Nos experimentos 1 e 2, os modelos XGBoost compreenderam as características de estudantes que levaram cerca de 2 a 4 anos, assim como os que levaram 5 anos ou mais, para realizar o ENADE após o início da graduação. Quanto aos experimentos 3 e 4, os modelos XGBoost compreenderam as características de estudantes que levaram de 2 a 4 anos entre o ano de início da graduação e o ano de realização do ENADE. O modelo XGBoost do Experimento 4 também conseguiu compreender o perfil de estudantes que levaram 8 anos ou mais entre os momentos considerados.

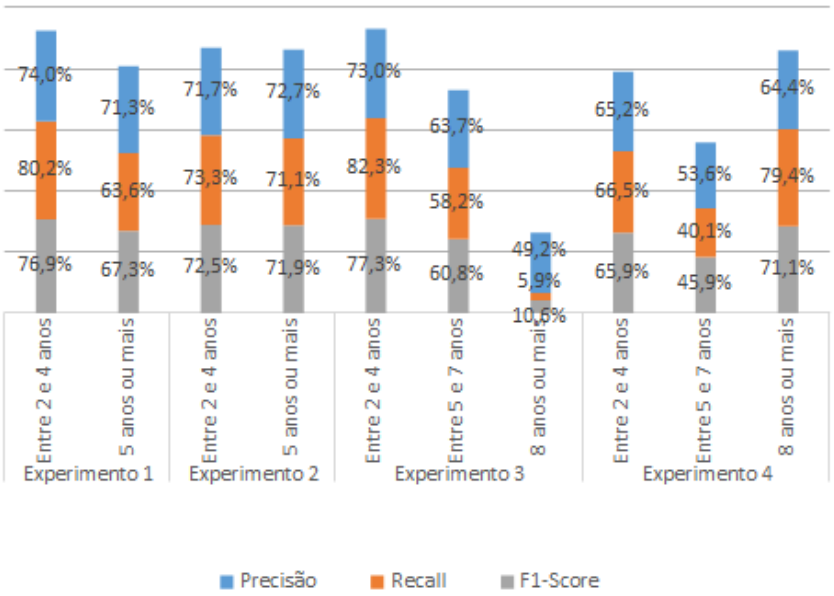


Figura 6 – Detalhes de aprendizado dos modelos XGBoost nos experimentos 1, 2, 3 e 4. Fonte: o autor (2021).

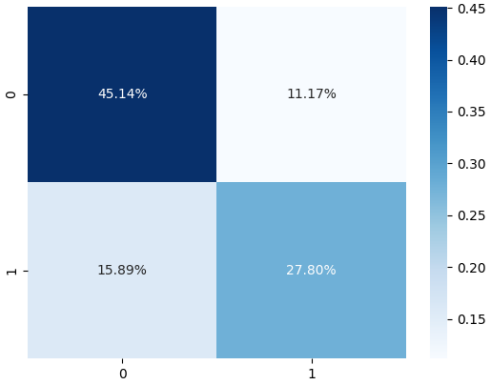


Figura 7 – Matriz de confusão do XGBoost do Exp. 1. Legenda: '0' = Entre 2 e 4 anos - '1' = 5 anos ou mais. Fonte: o autor (2021).

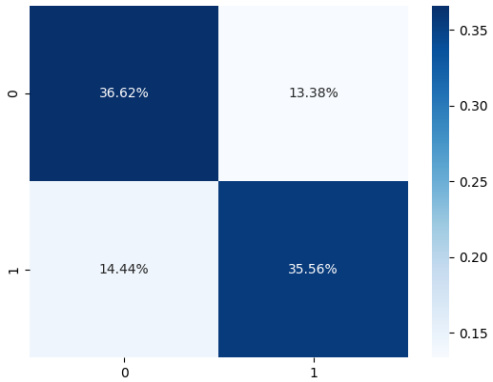


Figura 8 – Matriz de confusão do XGBoost do Exp. 2. Legenda: '0' = Entre 2 e 4 anos - '1' = 5 anos ou mais. Fonte: o autor (2021).

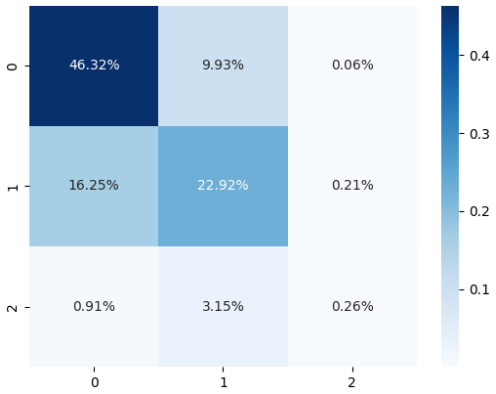


Figura 9 – Matriz de confusão do XGBoost do Exp. 3. Legenda: '0' = Entre 2 e 4 anos - '1' = Entre 5 e 7 anos - '2' = 8 anos ou mais. Fonte: o autor (2021).

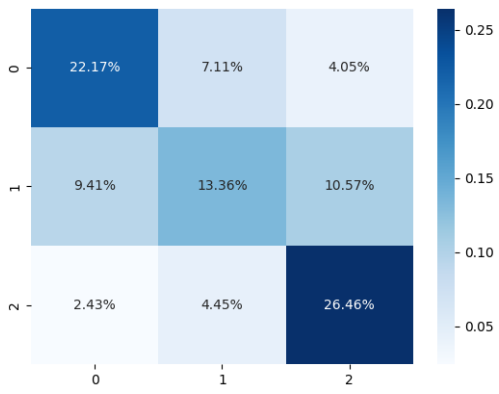


Figura 10 – Matriz de confusão do XGBoost do Exp. 4. Legenda: '0' = Entre 2 e 4 anos - '1' = Entre 5 e 7 anos - '2' = 8 anos ou mais. Fonte: o autor (2021).

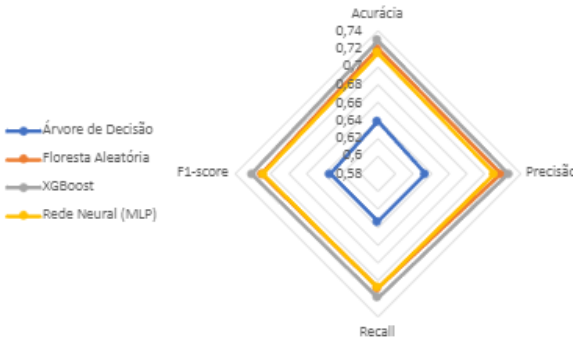


Figura 11 – Gráfico de radar dos modelos do Exp. 1. Fonte: o autor (2021).

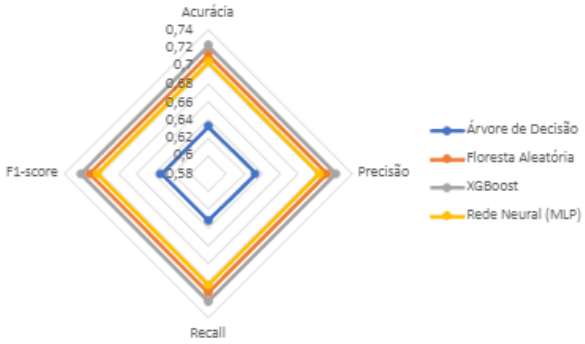


Figura 12 – Gráfico de radar dos modelos do Exp. 2. Fonte: o autor (2021).

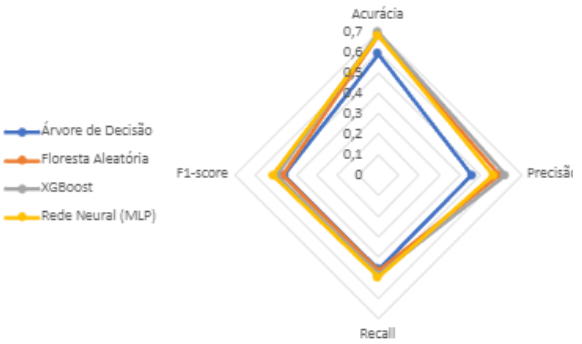


Figura 13 – Gráfico de radar dos modelos do Exp. 3. Fonte: o autor (2021).

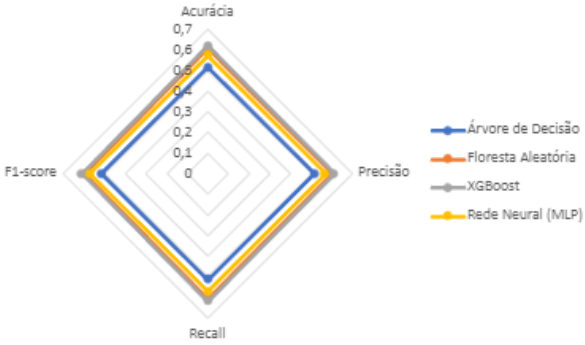


Figura 14 – Gráfico de radar dos modelos do Exp. 4. Fonte: o autor (2021).

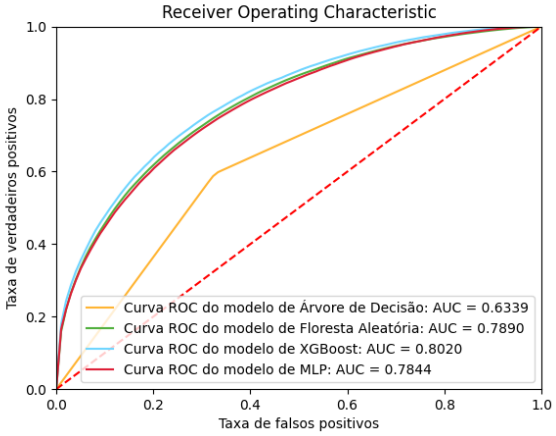


Figura 15 – Curvas AUC/ROC dos modelos do Exp. 1. Fonte: o autor (2021).

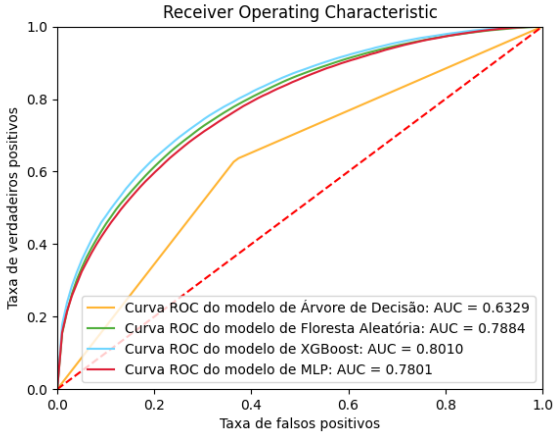


Figura 16 – Curvas AUC/ROC dos modelos do Exp. 2. Fonte: o autor (2021).

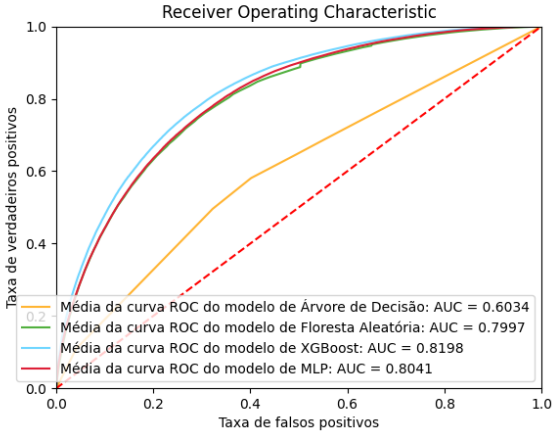


Figura 17 – Médias das curvas AUC/ROC dos modelos do Exp. 3. Fonte: o autor (2021).

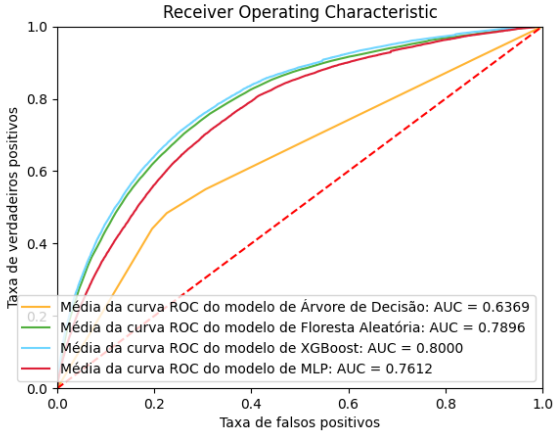


Figura 18 – Médias das curvas AUC/ROC dos modelos do Exp. 4. Fonte: o autor (2021).

Considerando os tempos de treinamento e teste, nota-se que, nos experimentos 1 e 2, com o uso de duas classes para a variável dependente, os modelos XGBoost, além de retornarem os melhores resultados, foram treinados e avaliados mais rapidamente do que os modelos de Floreta Aleatória e Rede Neural Perceptron Multicamadas. Nos experimentos 3 e 4, com o emprego de três classes para a variável dependente, os modelos de Floresta Aleatória foram treinados e avaliados mais rapidamente do que os modelos XGBoost, tendo resultados um pouco inferiores. Em todos os experimentos, os modelos de Rede Neural Perceptron Multicamadas foram os que apresentaram o maior custo temporal.

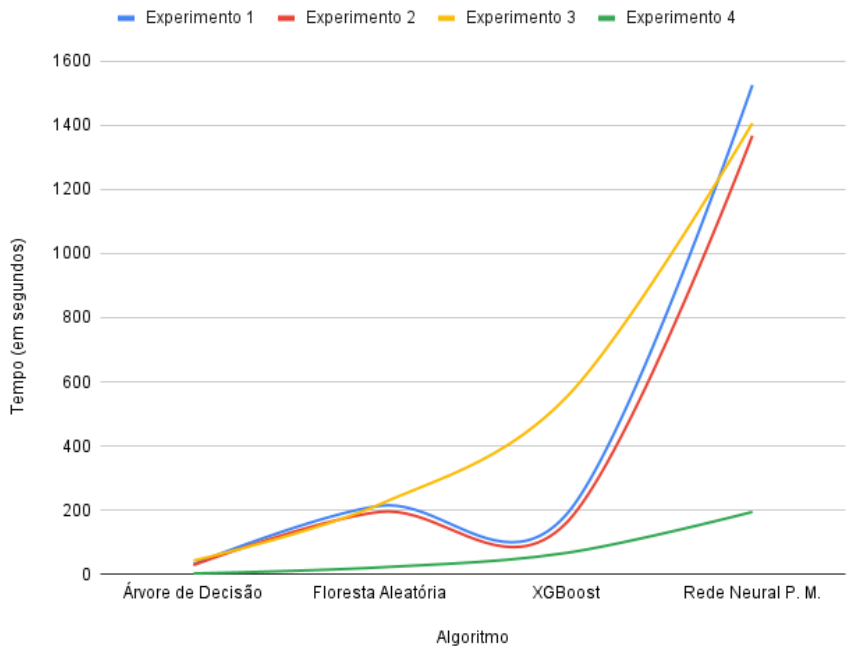


Figura 19 – Tempos de treinamento e teste dos modelos por experimento. Fonte: o autor (2021).

5.2 Cenário 2: Agrupamento de Dados de IES

Quanto aos resultados do agrupamento do segundo cenário, que observou 108 instituições de ensino superior públicas federais e estaduais brasileiras, o Método do Cotovelo retornou o valor k = 4 como quantidade ideal de grupos a serem gerados, como exibe a Figura 20. As figuras 21, 22, 23 e 24 apresentam, respectivamente, uma visão geral dos grupos 1, 2, 3 e 4, definidos com a observação de características da maior parte de suas IES, a partir de informações sobre despesas totais, quantidades de docentes e de técnicos, localização e categoria administrativa. O Apêndice H expõe uma visão mais detalhada sobre o agrupamento, bem como apresenta as IES que integram cada grupo.

Observando as figuras 21, 22, 23 e 24, 20,3% das IES consideradas no experimento de 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%. Por fim, o Grupo 4 tem 26,9% das IES analisadas. Diante disso, é justo afirmar que os grupos obtidos são balanceados.

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

com despesas entre R$794.566.153,68 e R$4.016.243.944,85.

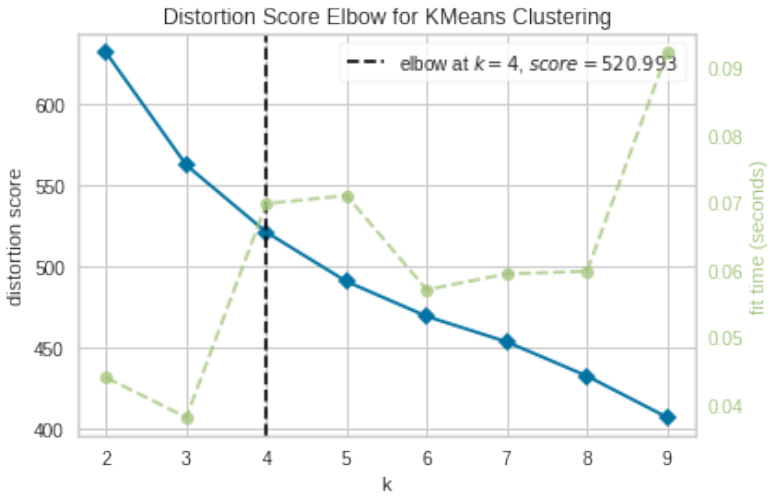


Figura 20 – Método do Cotovelo para o K-Means com a inicialização *k-means++* e a métrica da distorção. Fonte: o autor (2021).

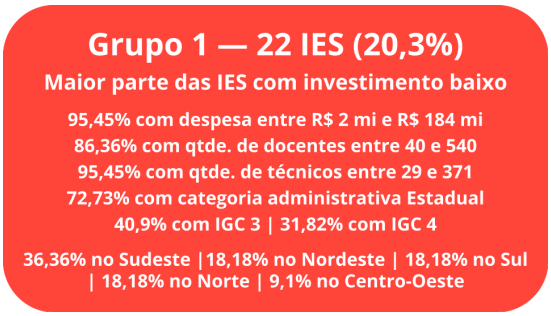


Figura 21 – Visão geral do Grupo 1. Fonte: o autor (2021).

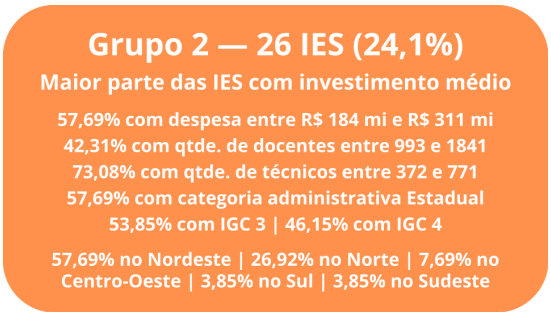


Figura 22 – Visão geral do Grupo 2. Fonte: o autor (2021).

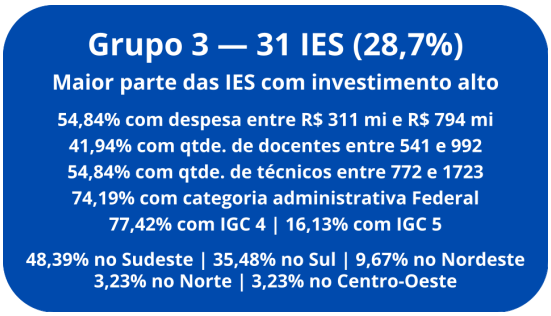


Figura 23 – Visão geral do Grupo 3. Fonte: o autor (2021).

Figura 24 – Visão geral do Grupo 4. Fonte: o autor (2021).

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 1841 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 1723 técnicos. Por fim, considerando as IES do Grupo 4, a maior parte delas tem entre 1842 e 4197 docentes e entre 1724 e 14581 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 sua vez, 74,19% das IES do Grupo 3 possuem a categoria de IES federal, bem como 86,21% 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. A maior parte (36,36%) das IES do Grupo 1 localiza-se na região Sudeste, assim como a maior parte (48,39%) das IES do Grupo 3. Na região Nordeste, está localizada a maior parte (57,69%) das IES do Grupo 2, assim como a maior parte (34,48%) das IES do Grupo 4.

Citando exemplos da 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, há 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 tem 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 exibidos, percebe-se que o Grupo 1, em sua maioria, é constituído de IES de categoria administrativa estadual, da região Sudeste, de investimento baixo, com baixas quantidades de docentes e técnicos, quando comparado com os demais grupos. No Grupo 2, a maior parte das IES também tem 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, onde a maior parte das IES tem investimento alto, nota-se que as instituições são federais e têm altas quantidades de docentes e técnicos, com localização nas regiões Sudeste e Sul. A maior parte das IES do Grupo 4 possui 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.

Nicolini, Andrade e Torres (2013), que usaram dados do ENADE de 2009 para

tratar do desempenho de estudantes de cursos de graduação de bacharelado em Administração, concluíram que universidades têm melhores resultados quando comparadas com centros universitários e faculdades. Os autores dizem que ”em parte, o senso comum prevalece”, uma vez que ”os melhores profissionais, em geral, continuam saindo das universidades públicas”. Os resultados obtidos no presente trabalho, com a interpretação das características das IES após o agrupamento, mostram que IES públicas, em especial, federais, têm investimento elevado, o que pode explicar o maior desempenho de seus estudantes no ENADE daquele ano. Os dados empregados neste estudo tratam de uma realidade mais recente, porém, considerando o histórico de investimento de universidades públicas, levantado por Moreno (2018), percebe-se que essas instituições tiveram orçamento aproximado nos períodos observados.

5.3 Cenário 3: Associação de Dados de Estudantes e IES

As regras obtidas com a execução dos 92 experimentos de associação do terceiro cenário que tiveram os maiores valores para suporte, confiança e *lift* foram selecionadas, interpretadas e empregadas na construção dos resumos visuais expostos a seguir. O Apêndice I apresenta essas regras em detalhes, com os seus atributos e valores de suporte, confiança e *lift*. Observando os dados de estudantes considerados, bem como o agrupamento realizado no cenário anterior, a Figura 25 exibe uma visão geral da alocação de estudantes por grupo de IES.

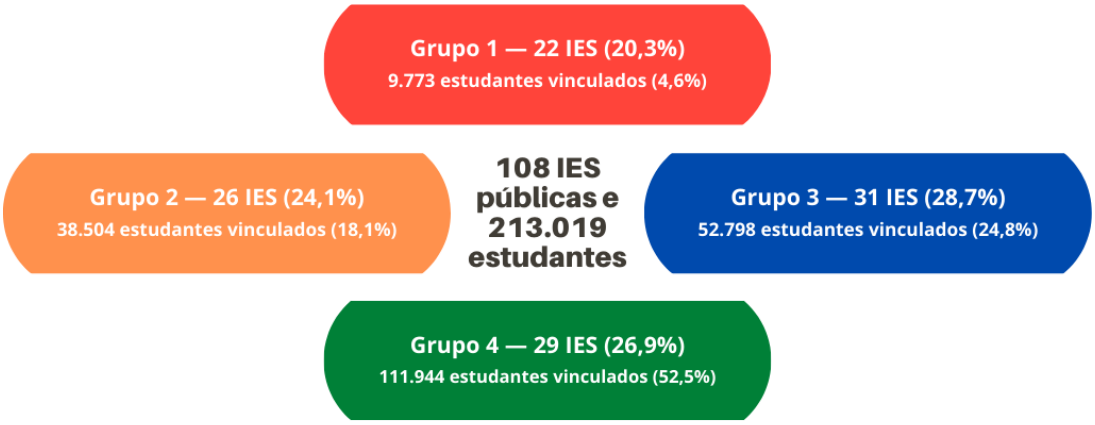


Figura 25 – Quantidade de estudantes por grupo de IES. Fonte: o autor (2021).

A Figura 26 exibe a visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES que integram o Grupo 1, grupo em que a maior parte das IES tem investimento baixo. A Figura 27 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 28 expõe regras de estudantes relacionados à IES do Grupo 3, cuja maior parte das IES tem investimento alto. Por fim, a Figura 29 resume

as regras sobre estudantes associados à IES do Grupo 4, grupo em que a maior parte das IES tem investimento muito alto.

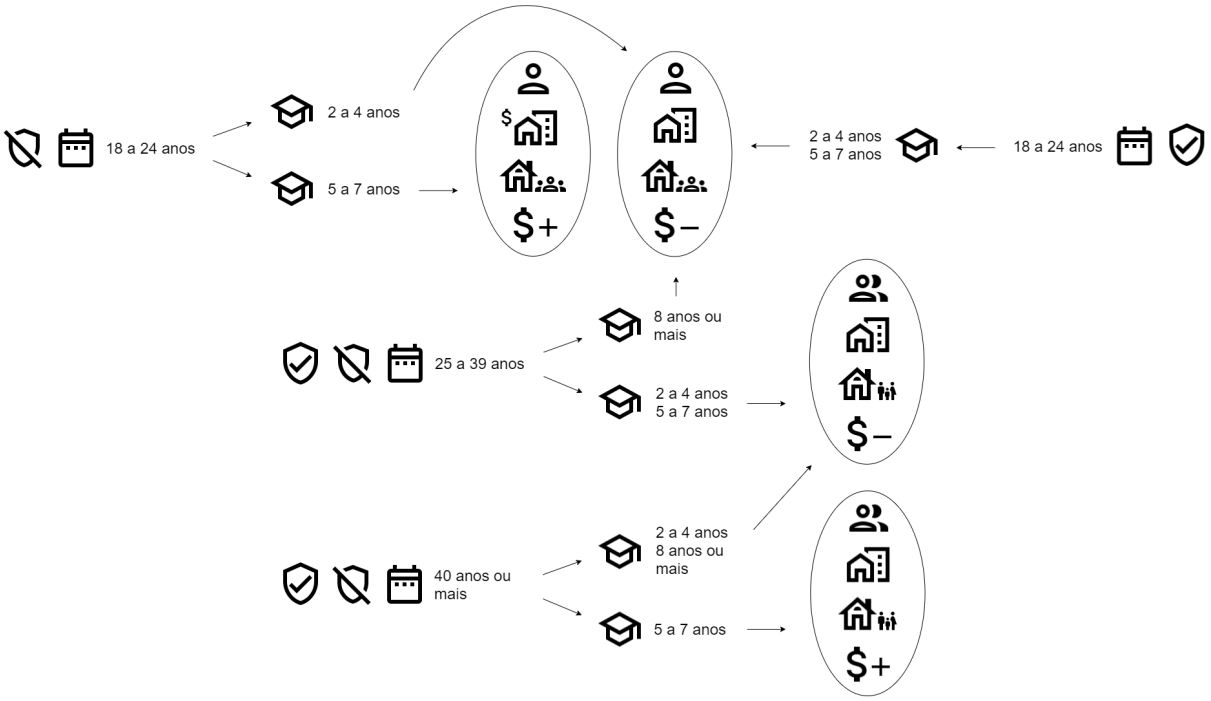
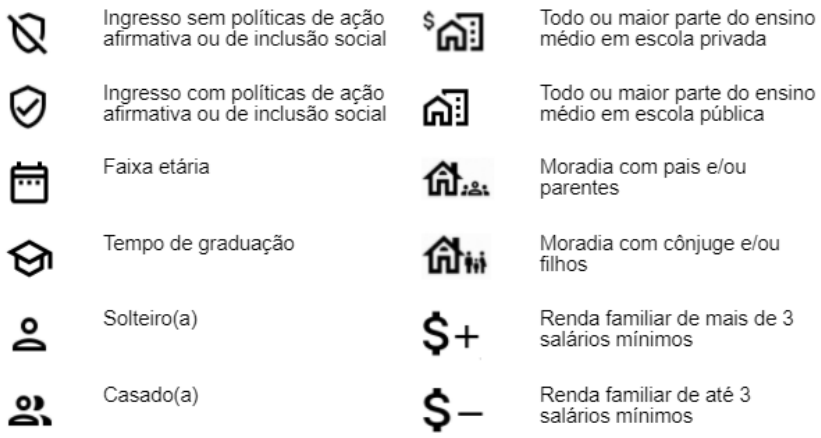


Figura 26 – Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 1. Fonte: o autor (2021).



Legenda das figuras 26, 27, 28 e 29.

A partir das regras geradas, 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, em geral, têm perfis semelhantes independentemente do grupo ao qual as suas IES estão vinculadas: são solteiros(as), 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. A única exceção trata de estudantes de

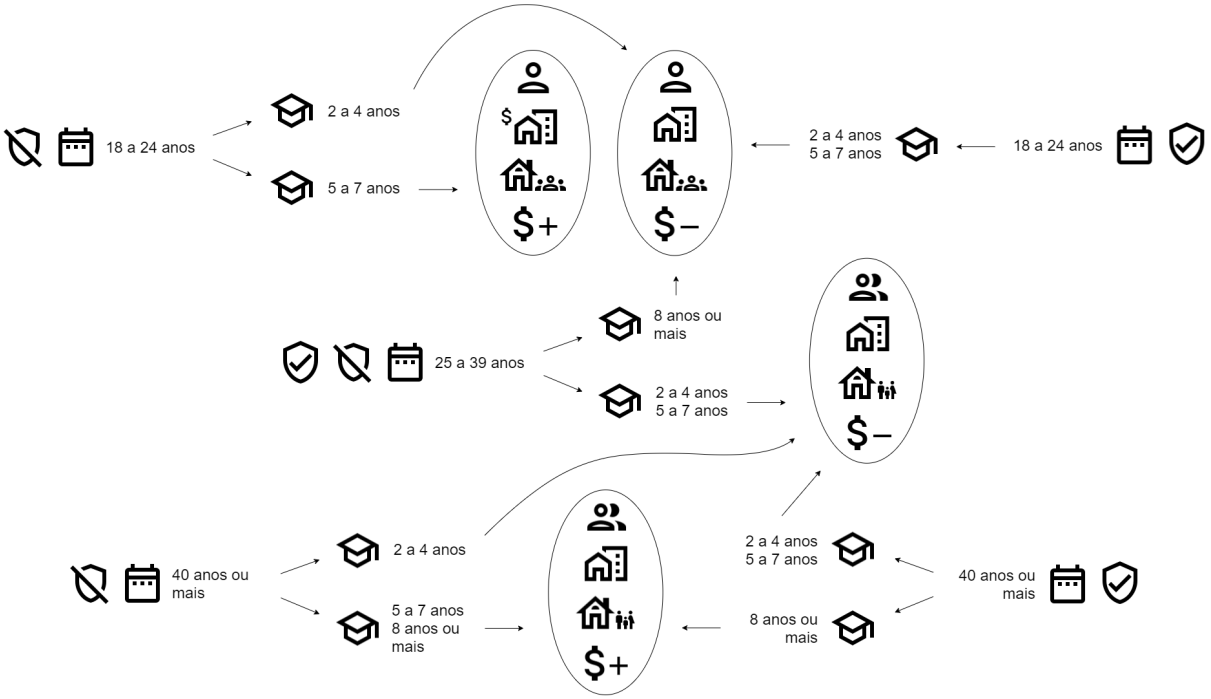


Figura 27 – Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 2. Fonte: o autor (2021).

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, em geral, são solteiros(as), 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. As exceções estão relacionadas a estudantes de IES dos grupos 1 e 2, que, a depender do tempo de graduação, podem ter estudado em escola pública e possuir 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, em geral, solteiros(as), 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. A exceção é designada por estudantes de IES do Grupo 3, que, dependendo do tempo de graduação, podem ter renda familiar acima de 3 salários. Considerando os 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

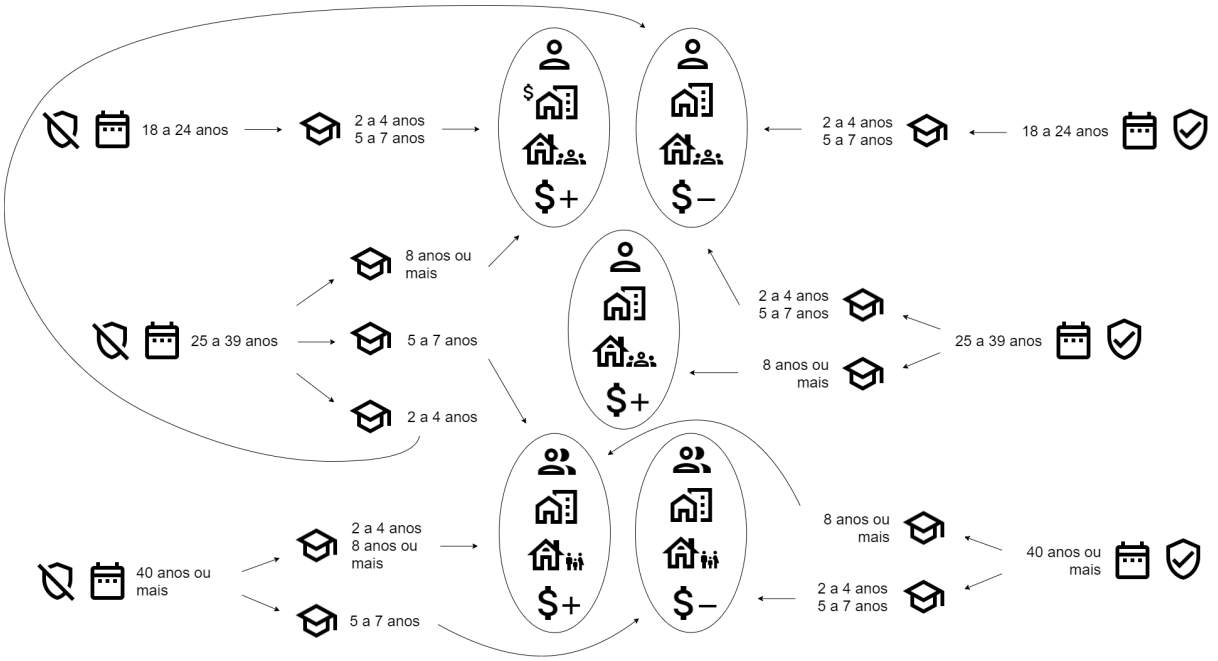


Figura 28 – Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 3. Fonte: o autor (2021).

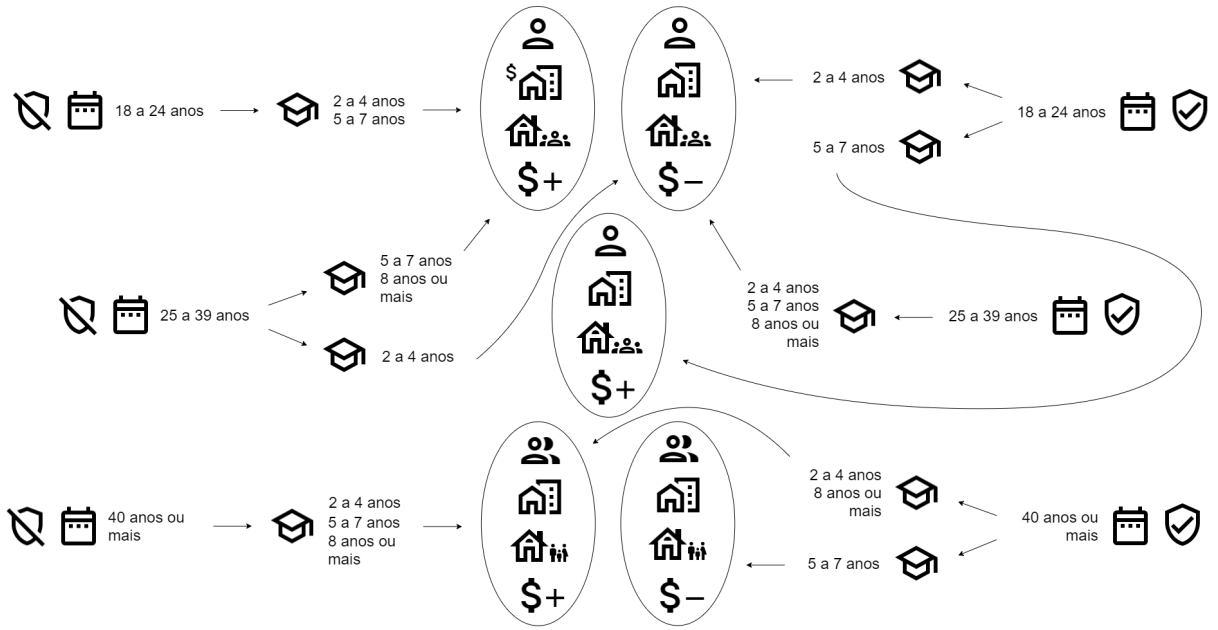


Figura 29 – Visão geral das regras de associação geradas sobre dados de estudantes vinculados à IES do Grupo 4. Fonte: o autor (2021).

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(as) 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(as), 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.

5.4 Considerações Finais

Este capítulo tratou dos resultados percebidos nos três cenários de mineração de dados executados. No Cenário 1, que tratou de prever o tempo aproximado de permanência de discentes de IES públicas federais e estaduais brasileiras na graduação, os modelos construídos com o uso do algoritmo XGBoost tiveram o melhor desempenho. Esses modelos foram treinados e avaliados mais rapidamente nos experimentos onde a variável dependente teve duas classes. No Cenário 2, quatro grupos de IES vinculadas aos estudantes com dados empregados no Cenário 1 foram gerados e definidos a partir das características da maior parte de suas instituições. Os grupos têm IES de investimento baixo, médio, alto e muito alto. Contemplando os resultados do Cenário 2, o Cenário 3 abordou a geração de regras de associação que possibilitaram a identificação de perfis socioeconômicos de estudantes cujos dados também foram considerados no Cenário 1, observando a sua forma de ingresso no ensino superior, entre outras informações.

6 Considerações Finais

Considerando que os objetivos deste estudo, que buscaram evidenciar a possibilidade de descobrir conhecimento por meio dos dados, método e ferramentas empregados, foram atingidos com êxito, é justo afirmar que o trabalho em questão designa uma contribuição válida para o contexto tratado, observando a importância da execução de pesquisas no âmbito do ensino superior público brasileiro.

No primeiro cenário de mineração de dados, 16 modelos de classificação foram construídos a partir de quatro experimentos, que tiveram quantidades de categorias e abordagens de balanceamento distintas para a variável dependente, por meio de algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural Perceptron Multicamadas. Os modelos XGBoost apresentaram o melhor desempenho nos quatro experimentos. Quanto ao balanceamento de dados da variável dependente, percebe-se que utilizar dados balanceados retornou melhores resultados, como era esperado. Essa parte do estudo pode estimular o desenvolvimento de trabalhos relacionados com a consideração de que a execução de pesquisas que tratem do tempo de graduação de discentes no âmbito educacional público é importante, uma vez que possibilita a compreensão sobre a permanência dos estudantes nas IES, o que contribui para a melhoria da gestão de recursos e o aprimoramento de processos de ensino-aprendizagem das instituições.

Sobre o agrupamento de IES públicas federais e estaduais brasileiras, 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, tal como nota-se que as instituições de maior investimento estão localizadas, principalmente, nas regiões Nordeste, Sudeste e Sul do Brasil. Tratando da alocação de IES pernambucanas nos grupos gerados, a Universidade Federal do Vale do São Francisco — UNIVASF — e a Universidade de Pernambuco — UPE — estão no Grupo 2, enquanto a Universidade Federal Rural de Pernambuco — UFRPE — e a Universidade Federal de Pernambuco — UFPE — encontram-se no Grupo 4. Quanto ao IGC, em 2019, a UPE teve conceito 3, enquanto UNIVASF, UFRPE e UFPE tiveram conceito 4.

As regras geradas no âmbito do cenário de associação permitiram uma percepção geral acerca dos perfis socioeconômicos dos discentes de IES públicas brasileiras, considerando os grupos aos quais as suas IES estão vinculadas, resultantes do cenário

de agrupamento. Os resultados do terceiro cenário indicam que, em geral, 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 poder aquisitivo menor no momento de conclusão da graduação quando comparados com estudantes de mesma idade que ingressaram no ensino superior sem políticas de ação afirmativa ou de inclusão social, que estudaram em escola privada e possuem poder aquisitivo 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, reduzam os índices de retenção e evasão em IES federais e estaduais brasileiras.

Referências

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. [S.l.: s.n.], 1994. (VLDB '94), p. 487–499. ISBN 1558601538. Citado 2 vezes nas páginas 26 e 27.

BRASIL. *Constituição da República Federativa do Brasil de 1988*. Brasília, DF: Senado Federal: Centro Gráfico, 1988. Citado na página 14.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. *Diário Oficial da União*, Brasília, DF, 2011. ISSN 1676-2339. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Citado na página 14.

BRASIL. *5 Motivos para a Abertura de Dados na Administração Pública*. 2015. Tribunal de Contas da União. <https://portal.tcu.gov.br/5-motivos-para-a-abertura-de-dados-na-administracao-publica.htm>. Acesso em 11 de dezembro de 2021. Citado na página 19.

CARVALHO, J.; CRUZ, L.; GOUVEIA, R. Descoberta de Conhecimento com Aprendizado de Máquina Supervisionado em Dados Abertos dos Censos da Educação Básica e Superior. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, v. 6, n. 1, p. 674, 2017. ISSN 2316-8889. Citado na página 27.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. *CoRR*, abs/1603.02754, 2016. Disponível em: <http://arxiv.org/abs/1603.02754>. Citado na página 25.

DIONISIO, A. et al. 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*. [S.l.: s.n.], 2015. Citado na página 28.

EDUCATIONALDATAMINING.ORG. *Educational Data Mining*. 2021. Educational Data Mining. <https://educationaldatamining.org/>. Acesso em 10 de novembro de 2021. Citado na página 23.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>. Citado 3 vezes nas páginas 8, 19 e 20.

FEATURE-ENGINE. *EqualFrequencyDiscretiser*. 2021. Read the Docs. <https://feature-engine.readthedocs.io/en/latest/discretisation/EqualFrequencyDiscretiser.html>. Acesso em 10 de janeiro de 2021. Citado na página 33.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge Discovery in Databases: An Overview. *AI Magazine*, v. 13, n. 3, p. 57, Sep. 1992. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1011>. Citado na página 19.

GOOGLE-DEVELOPERS. *Classification: Accuracy*. 2021. Machine Learning Crash Course, Google Developers. <https://developers.google.com/machine-learning/crash-course/classification/accuracy>. Acesso em 15 de fevereiro de 2021. Citado na página 25.

GOOGLE-DEVELOPERS. *Classification: Precision and Recall*. 2021. Machine Learning Crash Course, Google Developers. <https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>. Acesso em 15 de fevereiro de 2021. Citado na página 25.

GOOGLE-DEVELOPERS. *Classification: ROC Curve and AUC*. 2021. Machine Learning Crash Course, Google Developers. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>. Acesso em 15 de fevereiro de 2021. Citado na página 25.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. Disponível em: <http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1>. Citado 4 vezes nas páginas 22, 23, 24 e 25.

IBM. *Lift in an association rule*. 2021. Acesso em 10 de novembro de 2021. Disponível em: <https://www.ibm.com/docs/en/db2/9.7?topic=associations-lift-in-association-rule>. Citado na página 27.

IMBALANCED-LEARN. *RandomUnderSampler*. 2021. Scikit-Learn. <https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html>. Acesso em 6 de novembro de 2021. Citado na página 33.

INEP. *Censo da Educação Superior*. 2021. 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>. Acesso em 12 de janeiro de 2021. Citado na página 15.

INEP. *Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica (Cine Brasil)*. 2021. 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>. Acesso em 30 de outobro de 2021. Citado na página 16.

INEP. *Competências do INEP*. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/competencias>. Acesso em 11 de janeiro de 2021. Citado na página 14.

INEP. *Exame Nacional de Desempenho dos Estudantes (ENADE)*. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <http://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade>. Acesso em 12 de janeiro de 2021. Citado na página 15.

INEP. *Microdados de Indicadores de Qualidade da Educação Superior*. 2021. 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>. Acesso em 13 de agosto de 2020. Citado na página 15.

INEP. *Microdados do Censo da Educação Superior*. 2021. 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>. Acesso em 13 de agosto de 2020. Citado 2 vezes nas páginas 15 e 29.

INEP. *Microdados do ENADE*. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade>. Acesso em 13 de agosto de 2020. Citado 2 vezes nas páginas 15 e 29.

INEP. *Questionário do Estudante do ENADE*. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <http://portal.inep.gov.br/questionario-do-estudante>. Acesso em 13 de janeiro de 2021. Citado na página 15.

INEP. *Sobre o INEP*. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <http://inep.gov.br/web/guest/sobre-o-inep>. Acesso em 11 de janeiro de 2021. Citado na página 14.

MARTÍNEZ-PLUMED, F. et al. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, Institute of Electrical and Electronics Engineers (IEEE), dez. 2019. ISSN 1041-4347. Citado 2 vezes nas páginas 8 e 21.

MIERSWA, I. et al. YALE: Rapid Prototyping for Complex Data Mining Tasks. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2006. (KDD '06), p. 935–940. ISBN 1595933395. Disponível em: <https://doi.org/10.1145/1150402.1150531>. Citado na página 29.

MORAIS, A. M. de P. et al. Universities, socioeconomic standards and inclusion policies: Assessing the effects on the performance of Brazilian undergraduates. *Studies in Educational Evaluation*, v. 70, p. 100996, 2021. ISSN 0191-491X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0191491X21000225>. Citado na página 28.

MORENO, A. C. *90% das universidades federais tiveram perda real no orçamento em cinco anos; verba nacional encolheu 28%*. 2018. G1. <https://glo.bo/3GVFFow>. Acesso em 11 de dezembro de 2021. Citado na página 48.

NICOLINI, A.; ANDRADE, R.; TORRES, A. A. G. Comparando os resultados do ENADE 2009 por número de instituições e número de estudantes: como anda o desempenho acadêmico dos cursos de administração? *Administração: Ensino e Pesquisa*, v. 14, p. 161, 03 2013. Citado 2 vezes nas páginas 27 e 47.

OKFN. *What is open?* 2021. Open Knowledge Foundation. <https://okfn.org/opendata/>. Acesso em 11 de dezembro de 2021. Citado na página 19.

OLIVEIRA, R. B. de; BRITO, J. A. de M. Análise de Agrupamento Aplicada ao Estudo de Instituições de Ensino Superior Públicas. In: *Revista do Seminário*

Internacional de Estatística com R. Niterói, RJ, Brasil: [s.n.], 2019. Disponível em: <https://periodicos.uff.br/anaisdoser/article/view/29316/17030>. Citado na página 27.

RAMOS, J. et al. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2020. p. 1092–1101. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12865>. Citado na página 21.

SCIKIT-LEARN. *Multiclass classification*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/multiclass.html#multiclass-classification>. Acesso em 7 de novembro de 2021. Citado na página 25.

SCIKIT-LEARN. *sklearn.metrics.f1_score*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>. Acesso em 7 de novembro de 2021. Citado na página 25.

SCIKIT-LEARN. *Sklearn.model_selection.StratifiedKFold*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.htm>. Acesso em 12 de fevereiro de 2021. Citado na página 33.

SCIKIT-LEARN. *sklearn.neural_network.MLPClassifier*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html>. Acesso em 6 de novembro de 2021. Citado na página 35.

SCIKIT-LEARN. *Sklearn.preprocessing.OneHotEncoder*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>. Acesso em 10 de fevereiro de 2021. Citado 2 vezes nas páginas 22 e 33.

SCIKIT-LEARN. *sklearnsklearn.ensemble.RandomForestClassifier*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. Acesso em 6 de novembro de 2021. Citado na página 35.

SCIKIT-LEARN. *sklearn.tree.DecisionTreeClassifier*. 2021. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>. Acesso em 6 de novembro de 2021. Citado na página 35.

SHEARER, C. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, v. 5, n. 4, 2000. Citado na página 21.

SILVA, A.; HOED, R.; SARAIVA, P. Análise do Desempenho dos Alunos de Cursos Superiores em Computação no ENADE - Uma Abordagem usando Mineração de Dados. In: *Atas da conferência Ibero-Americana*. [S.l.: s.n.], 2019. p. 207–214. Citado na página 28.

SILVA, V. et al. 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*. Porto Alegre, RS, Brasil: SBC, 2020. p. 72–81. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12763>. Citado na página 28.

SOUZA, A. M. de. *Machine Learning e a evasão escolar: análise preditiva no suporte à tomada de decisão*. 134 p. Dissertação (Mestrado) — Fundação Mineira de Educação e Cultura, Belo Horizonte, 2020. Citado na página 27.

SOUZA, F. C. de et al. Análise das IES da Área de Ciências Contábeis e de seus Pesquisadores por meio de sua Produção Científica. *Contabilidade Vista amp; Revista*, v. 19, n. 3, p. 15–38, maio 2009. Disponível em: <https://revistas.face.ufmg.br/index.php/contabilidadevistaerevista/article/view/359>. Citado na página 27.

Tukey, J. W. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977. (Behavioral Science: Quantitative Methods). Citado na página 23.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. Amsterdam: Morgan Kaufmann, 2011. (Morgan Kaufmann Series in Data Management Systems). ISBN 978-0-12-374856-0. Disponível em: <http://www.sciencedirect.com/science/book/9780123748560>. Citado 4 vezes nas páginas 23, 24, 25 e 26.

XGBOOST. *XGBoost Python API Reference*. 2021. XGBoost Read the Docs. <https://xgboost.readthedocs.io/en/latest/python/python_api.html>. Acesso em 6 de novembro de 2021. Citado na página 35.

YELLOWBRICK. *Elbow Method*. 2021. Yellowbrick. <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>. Acesso em 7 de novembro de 2021. Citado 2 vezes nas páginas 26 e 37.

Apêndices

Os apêndices deste trabalho podem ser visualizados em <https://osf.io/cvu3h/?view_only=1bad2d79da5d4a96b66bd0945fa7b607> e em <https://bit.ly/3DX5e7D>.